FINAL REPORT

Multimodal Sensor Fusion for UXO Classification and Remediation

SERDP Project MR18-1440

FEBRUARY 2020

Aaron Marburg
University of Washington Applied Physics Laboratory

SERDP
DOD ▪ EPA ▪ DOE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01/Feb/2020	SERDP Final Report	12/29/2017 - 12/28/2019

4. TITLE AND SUBTITLE

Multimodal Sensor Fusion for UXO Classification and Remediation

5a. CONTRACT NUMBER
18-C-0005

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

Aaron Marburg

5d. PROJECT NUMBER
MR18-1440

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Washington Applied Physics Laboratory
1013 NE 40th St
Seattle, WA 98105

8. PERFORMING ORGANIZATION REPORT NUMBER
MR18-1440

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Strategic Environmental Research and Development Program (SERDP)
4800 Mark Center Drive, Suite 16F16
Alexandria, VA 22350-3605

10. SPONSOR/MONITOR'S ACRONYM(S)
SERDP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
MR18-1440

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Despite the ubiquity of robotic systems in deep-ocean intervention, such approaches have limited impact in shallow-water UXO remediation, due in large part to the relatively crude and non-dextrous nature of the current state of the art in tele-operated manipulation. Computer-assisted or -controlled approaches over great promise for addressing the fundamental issues in subsea tele-operation, allowing safe and effective execution of UXO remediation tasks; however, such computer assistance requires accurate digital models of the UXO in place on the seabed. While terrestrial research can rely on a variety of structured light and LiDAR-based sensors to generate such models in near realtime, no such turnkey solutions exist for subsea application, particularly for operation in the shallow, turbid waters where UXO remediation is of highest priority. This program examines the use of visible light stereo cameras and a high frequency forward-looking sonar, combined with platform motion, to construct and update 3D reconstructions of UXO on the sea floor.

15. SUBJECT TERMS

Multimodal Sensor Fusion, UXO, UXO Classification, Remediation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Aaron Marburg
UNCLASS	UNCLASS	UNCLASS	UNCLASS	66	**19b. TELEPHONE NUMBER** *(Include area code)* 206-685-8461

Abstract

Introduction and Objectives

Despite the ubiquity of robotic systems in deep-ocean intervention, such approaches have limited impact in shallow-water UXO remediation, due in large part to the relatively crude and non-dextrous nature of the current state of the art in tele-operated manipulation. Computer-assisted or -controlled approaches offer great promise for addressing the fundamental issues in subsea tele-operation, allowing safe and effective execution of UXO remediation tasks; however, such computer assistance requires accurate digital models of the UXO in place on the seabed. While terrestrial research can rely on a variety of structured light and LiDAR-based sensors to generate such models in near realtime, no such turnkey solutions exist for subsea application, particularly for operation in the shallow, turbid waters where UXO remediation is of highest priority. This program examines the use of visible light stereo cameras and a high frequency forward-looking sonar, combined with platform motion, to construct and update 3D reconstructions of UXO on the seafloor.

Technical Approach

This work encompasses four major tasks: (1) The construction of sensor platform containing stereo 4K cameras and a 2.1 MHz imaging sonar, along with software allowing time-synchronized recording of data from all sensors. This system is mounted on a camera gantry which allows repeated, constrained motion approximating the close inspection of a UXO preceding and during manipulation. (2) Collection of data sets with the sensor capture system; including relevant metadata and the estimation of a ground truth world structure and camera trajectory. (3) Development of reprojection models between all sensors, particularly a procedure for using observed data to estimate the mechanical offset between the camera center and the origin of the sonar data. And finally, (4) the extension of LSD-SLAM, a monocular Simultaneous Localization and Mapping (SLAM) algorithm, to meet the particulars of the described application, including making use of stereo for direct scale measurements, improving model convergence given relatively small camera motion, and inclusion of sonar data.

Results

The improved LSD-SLAM algorithm is shown to produce a converged 3D model in realtime of a test scene using stereo video, including an estimate of camera trajectory. Outside of an undiagnosed scale error, this trajectory has a high degree of agreement with the independently measured ground truth trajectory for both camera position and attitude. An effective camera-to-sonar calibration procedure is also demonstrated, including preliminary results in projecting sonar data into the visual frame.

Benefits

This program developed hardware and software tools for gathering synchronized stereo video and imaging sonar data of objects, including estimation of ground truth scene structure and camera trajectory. It also showed the effectiveness of stereo visual approaches for 3D reconstruction in low-turbidity conditions, allowing continued progress towards the original application, assistive ROV manipulation, when those conditions are present. There remains significant further work to be done both in ensuring the visual reconstruction is robust and in making use of acoustic data either in supplement to, or in lieu of optical data.

Executive Summary for SERDP MR18-1440: Multimodal Sensor Fusion for UXO Classification and Remediation

PI: Aaron Marburg (amarburg@uw.edu); University of Washington Applied Physics Laboratory

Introduction and Objectives

Due to the crushing pressures found at depth, work in the deep ocean relies heavily on remotely operated vehicles (ROVs), tele-operated robotic platforms which carry optical and acoustic sensors, manipulators, and other tools for observation of and interaction with the benthic environment. Despite their critical importance to subsea operations, ROVs remain technologically very simple, passing realtime sensor data over a high-bandwidth umbilical to operators on a surface support ship, who then issue low-level motion commands back to the ROV's thrusters and manipulators. While interaction with the environment is a key ROV capability, limitations in sensor fidelity, the quality of feedback from the manipulators, and limited control ergonomics mean the viability of delicate, dextrous tasks depends largely on an individual operator's skill. These limitations, as well as the capital and operating costs of such systems, reduce their competitive advantage relative to divers for many shallow water tasks, except in scenarios with high operational risk.

Given the inherent risk to divers in UXO remediation, there is a clear opportunity for adapting these marine robotic technologies to UXO response. However, given the importance of delicate handling during remediation, the current state of the art in tele-operated manipulation is insufficient. Preliminary results have shown the benefits of computer-guidance or assistance for ROV manipulation (Rydén et al., 2013; Choi et al., 2015), building on a broad base of research into the use of intermediating or augmentative computer control to improve the precision and performance of remote manipulation in applications ranging from telesurgery (Taylor et al., 2016) to exoplanetary exploration. Such assistive control can range from near-transparent systems where the operator retains full control while operating within an envelope of computer-defined "safety stops" which prevent known dangerous actions, up to fully autonomous control where the computer directs the manipulator and the operator provides only high-level, supervisory inputs.

Computer-assisted or robotic manipulation in an unstructured environment requires that the system sense the geometry of the objects to be manipulated as well as the surrounding environment as a basis for understanding current state, planning future actions, and detecting unexpected or fault behaviors. For seafloor munitions response this translates to the generation of digital models of UXO in place on the seafloor of sufficient resolution and accuracy to allow algorithmic identification of the object and execution of a remediation plan by a robotic manipulator.

The collection of such models underwater is non-trivial. The physics of both light and sound in water, as well as the engineering and economic challenges inherent in the development of marine sensors, lead to a paucity of turn-key underwater 3D sensing technologies. Further, given the shallow depths and frequently soft bottom types in areas of greatest concern for UXO remediation, water clarity is often severely limited, handicapping optical techniques. Finally, any perceptual products must be available in near-realtime, as this determines the delay between observing the scene and taking action; and it is highly desirable that the model be dynamic and readily updated. An ideal sensor would provide provide a dense 3D model of the scene with a refresh rate comparable to the physical dynamics of the world (e.g., video rate); short of that, it should be feasible to update the model on demand with minimal delay.

Technical Approach

The project has four major phases:

- **Development of an optical-acoustic data collection system** which emulates the sensor front end of an ROV. A core requirement is the ability to move in a manner representative of an ROV surveying an object on the seafloor. However, this trajectory must be independently measurable to allow evaluation of the accuracy of the trajectory calculated from sensor data. A free-floating sensor package (or a small ROV) would be allow more realistic motion, but would also be expensive to construct, more complex, and would not offer a straightforward method for accurate pose measurement. Instead, a motion gantry system was developed which allows an orbiting motion around objects of interest. While constrained, this trajectory is more repeatable, and more easily instrumented for ground truth.

  A second critical requirement is ensuring the data from the two sensor modalities are correctly time-synchronized and available in realtime for processing by the reconstruction algorithm.

- **Collection of realistic test data** of objects of interest including calibration shapes, generic objects and UXO stand-ins. The importance of the collection of thorough, well-documented sample data sets was underlined throughout the program, and a critical outcome from this project was a focus on collecting and curating reference optical-acoustic data sets to support future research.

- **Development of inter- and intra-sensor reprojection models.** The reprojection model of each camera on its own and as a stereo pair can be calculated using computer vision best practices. Similarly, we assume the imaging sonar is metric given an accurate sound velocity, such that the bearing and range to any given sonar bin is correct. The question of measuring and estimating distortions in the sonar model, while interesting, was deemed out of scope. The most challenging component is the estimation of the physical offset between the origins of the camera(s) and of the sonar, which allows the reprojection of world points between sensor frames.

- **Development of an opti-acoustic reconstruction algorithm.** The selected approach is based on an LSD-SLAM, an existing open source realtime visual reconstruction algorithm (Engel, Schöps, et al., 2014), which is then adapted for the particulars of program, including the addition of depth information from stereo cameras, and inclusion of data from the imaging sonar. As executed, this phase is divided into three sub-tasks:

  - Adoption and refactoring of the existing open source 3D reconstruction algorithm.
  - Extension of the algorithm to incorporate stereo video.
  - Extension of the algorithm to incorporate sonar data.

Results

The resulting data collection system was used to capture time-synchronized data sets of the sensors observing UXO and non-UXO objects.

Camera-sonar calibration

An algorithm for the calculation of the mechanical offset between the stereo camera pair and the imaging sonar from shared views of a fiducial object was developed. This offset, paired with sensor models for both

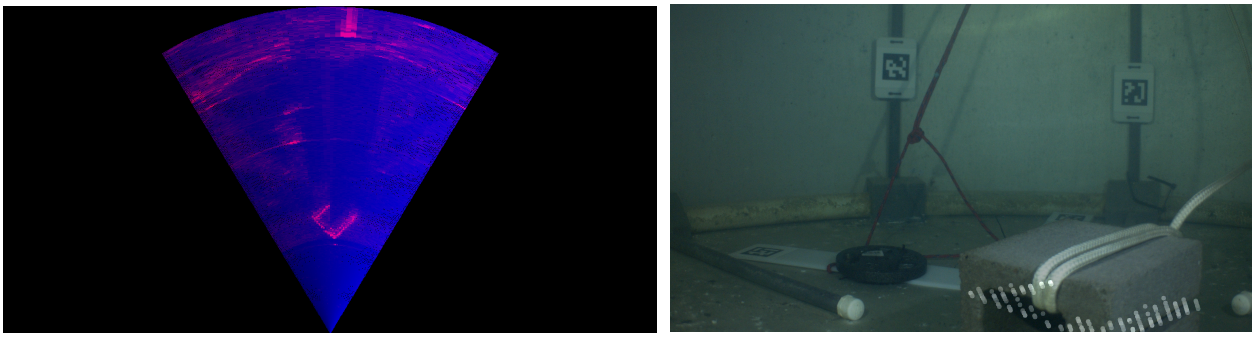


Figure 1: The calculated sonar-to-camera extrinsic calibration allows reprojection of high-intensity points from sonar data (left) into the perspective of the camera (right), noting that the projected points are placed on the vertical centerline of the sonar's beam pattern; the true vertical location of the observed targets within the sonar's vertical aperture cannot be measured from the data.

types of sensors allows reprojection of world points between sensor images, an essential task for data fusion. An exemplar result can be seen in Figure 1, with a rendering of the imaging sonar data on the left, and an image from the left camera on the right. The high intensity sonar returns corresponding to the cinderblock are reprojected into the camera image – the longer range returns from the back wall of tank are not shown for clarity. Note that the imaging sonar exhibits a ambiguity due to its vertical beamwidth. For illustrative purposes, this projection places the reprojected points on the centerline of the sonar beam pattern.

3D Reconstruction from stereo optical imagery

In this program, ground truth scene structure and camera trajectory are calculated using Metashape, a post-processed photogrammetric reconstruction package, as this proved to be more accurate than the low-cost IMU installed on the sensor head. Sample output from this tool is shown in Figure 2. Use of this commercial software also provides a reference implementation of visual reconstruction, although for reasons of performance it is not suitable as a realtime reconstruction input for robotic manipulation.

A reconstruction from the developed SLAM algorithm can be seen in Figure 3. Unlike the post-processed reconstruction shown in 2, this reconstruction is built iteratively and in realtime as video data is captured. Compared to the reference reconstruction it is more sparse (e.g., missing portions of the tank floor and walls) because the underlying algorithm preferentially focuses on regions of high image gradient, as those tend to correspond to physical edges in the world. Further the reconstruction is "fuzzy" in some locations due to the temporally converging nature of the algorithm, which uses new imagery to incrementally improve the model but does not retain that information to achieve global optimality as is done in Metashape.

An initial examination of the post-processed and realtime trajectories revealed a scaling error in the LSD-SLAM results, despite the use of stereo imagery which should provide results to scale. The results which follow include a compensatory scaling factor calculated to minimize 3D RMS error between the two trajectories and as such should be considered indicative. We assume the scaling error is due to an undiagnosed bug in our SLAM implementation.

With that caveat, the calculated post-processed and realtime trajectories are shown in Figures 4, 5 and 7; the positioning errors are shown in 6.

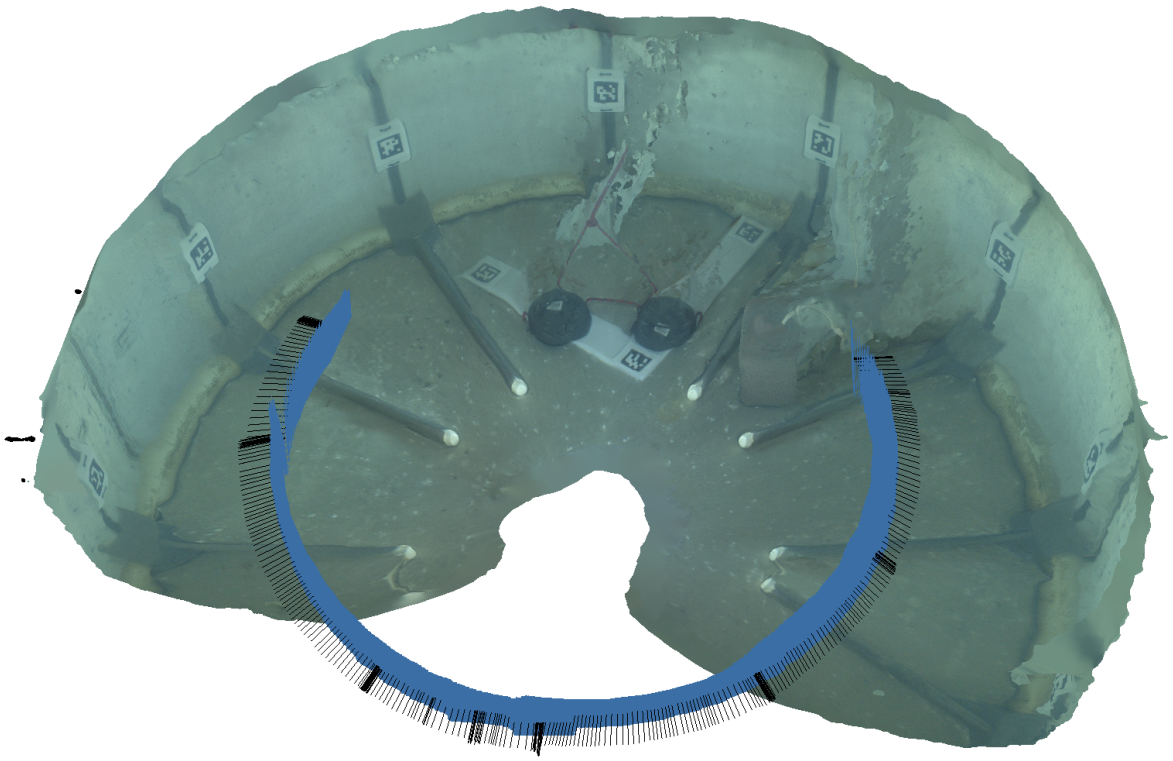


Figure 2: 3D reconstruction of camera images from Metashape showing the cinderblock and L-shaped fiducial. Note the failure of reconstruction near the taglines on both the fiducial and cinderblock which may be due to the relatively low texture on the tank wall near the taglines, or may be due to motion in the lines during data collection.

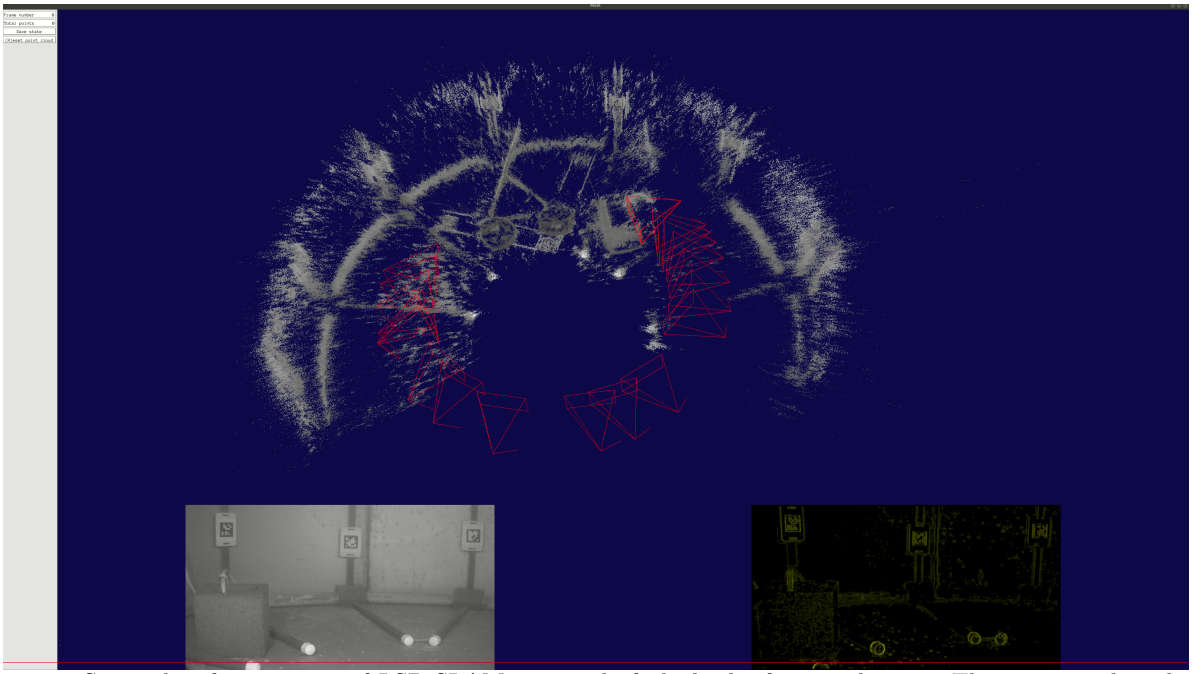


Figure 3: Screenshot from output of LSD-SLAM near end of playback of a test dataset. The main window shows the resulting point cloud (white) and the position of each of the keyframes (red). The lower left window shows the current frame of video input, while the lower right shows the associated gradient image.

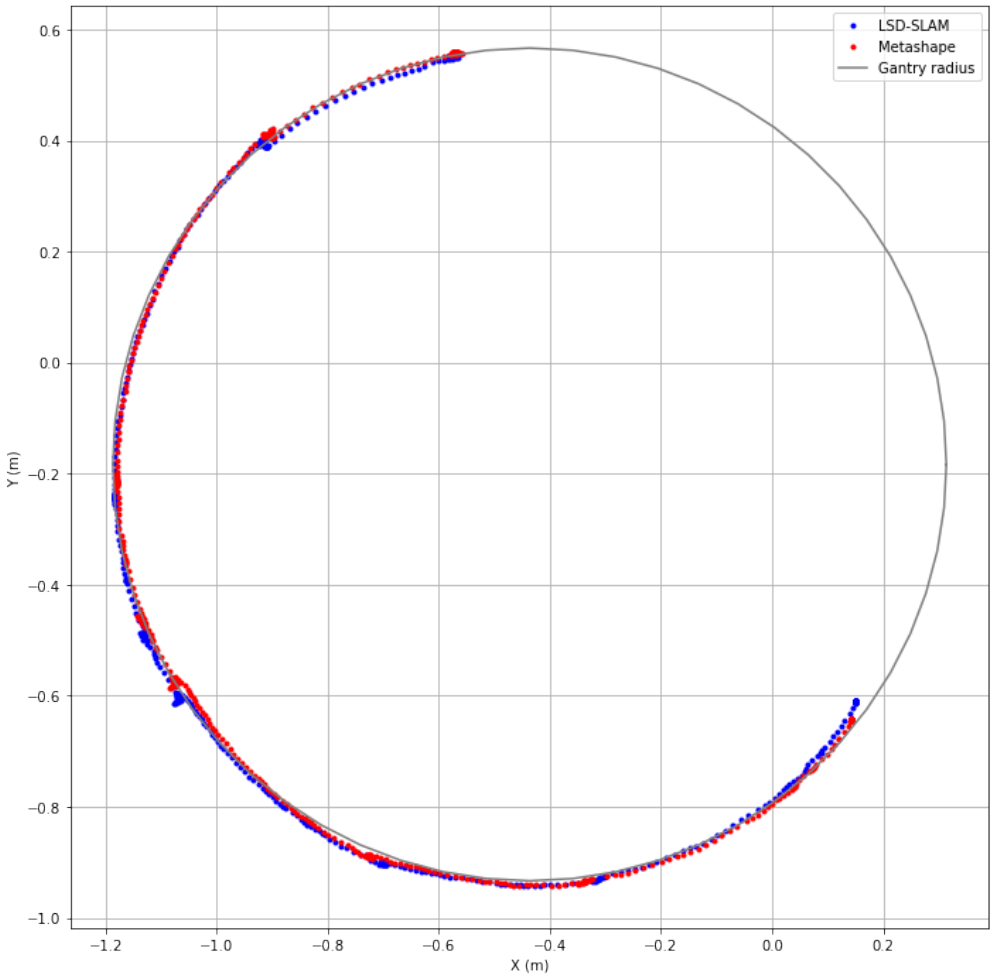


Figure 4: X-Y projection of Metashape and scale-corrected LSD-SLAM trajectories after aligning the first LSD-SLAM camera position with the corresponding camera in the Metashape trajectory and applying the calculated scale correction. The known radius of the motion gantry is also shown. The absolute position of the Metashape trajectory is derived from multiple observations of the L-shaped fiducial in the original image data.

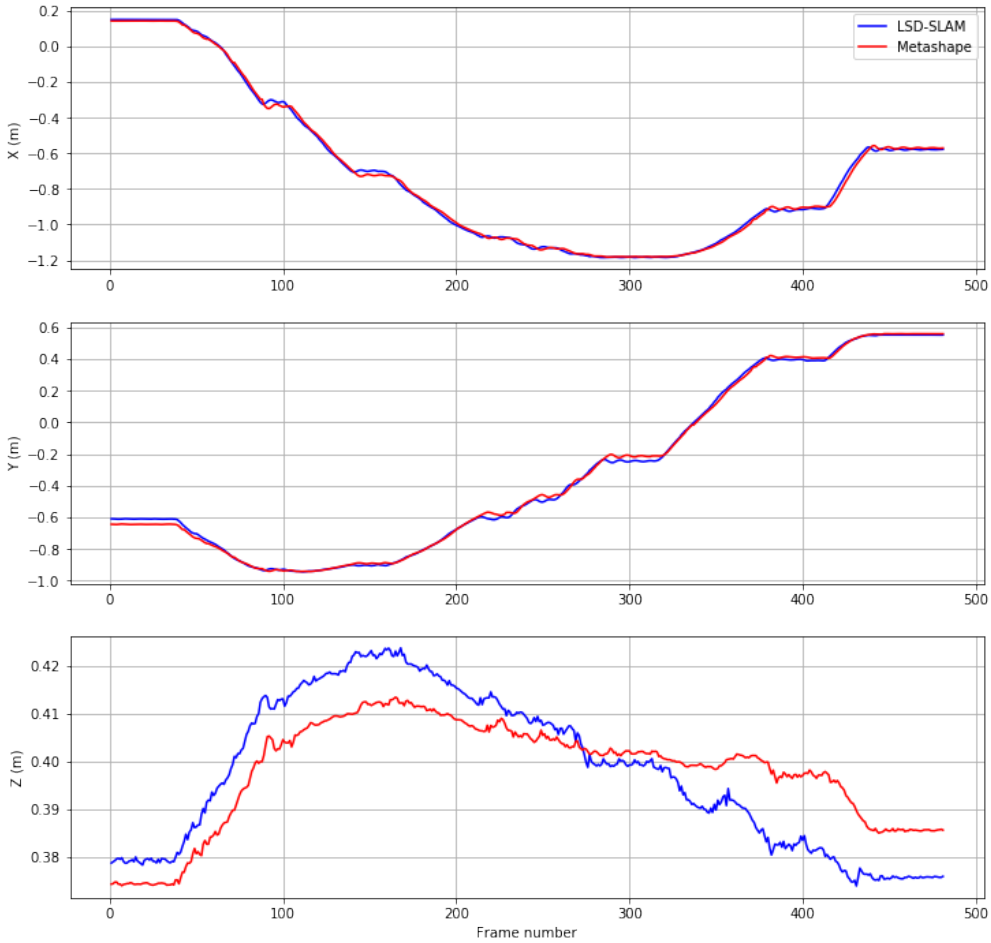


Figure 5: X, Y and Z components for Metashape and scale-corrected LSD-SLAM trajectories.

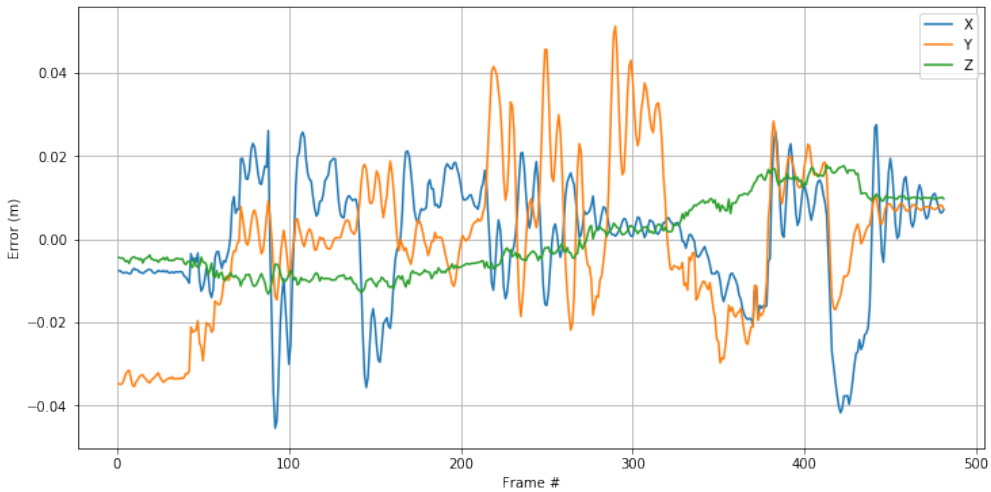


Figure 6: X, Y and Z errors between Metashape and scale-corrected LSD-SLAM trajectories.

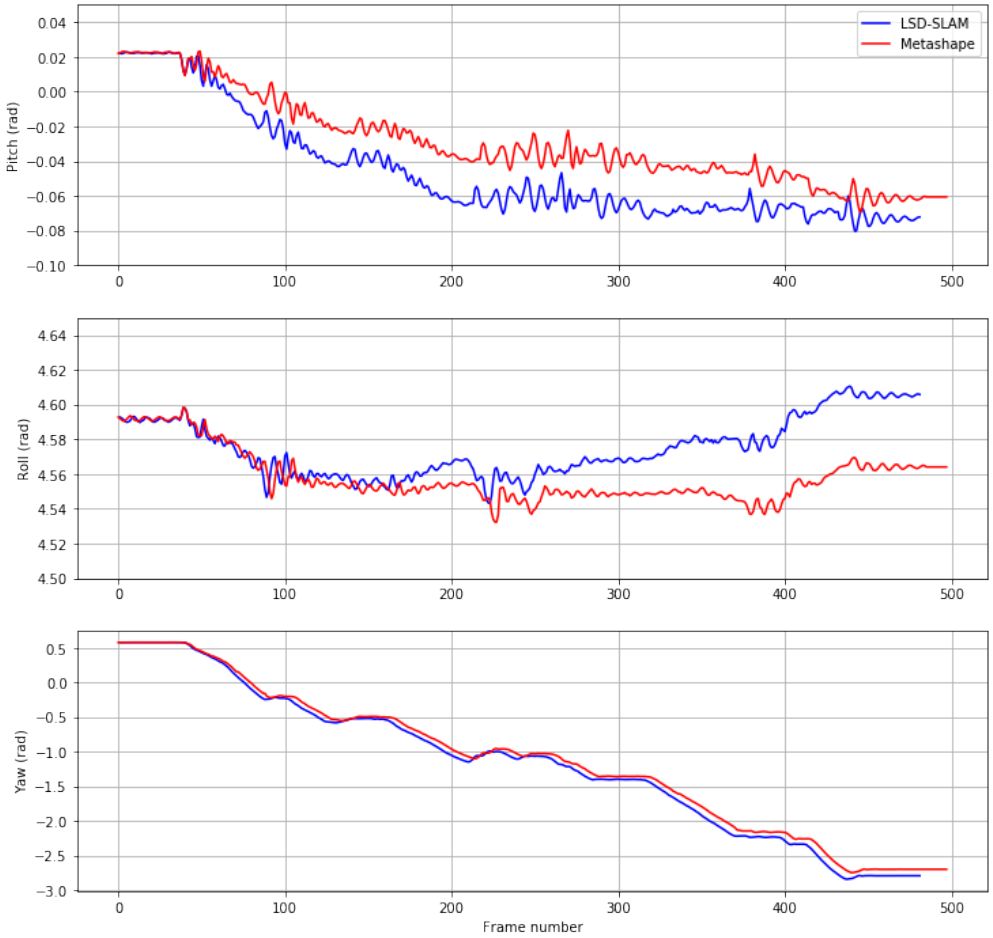


Figure 7: Roll, pitch, and yaw of Metashape and scale-corrected LSD-SLAM trajectories.

Implications for Future Research and Benefits

Overall, this project made significant progress towards the collection and analysis of simultaneous optical and acoustic data of objects on the seafloor, including construction of a novel testbed for collecting consistent test datasets and independently estimating ground truth scene structure and sensor head motion. As shown, stereo visual construction was performed using data collected by the system, although this required far greater development than initially anticipated due to the need to explicitly re-introduce stereo processing to the selected SLAM algorithm, and to improve SLAM system performance in low-motion underwater images as found in this test scenario, as well as low quality in the original software implementation. Furthermore, we demonstrated a successful approach to camera-sonar calibration allowing mapping of sonar data into the video frame. Ultimately, there was insufficient project scope to achieve the final step of leveraging acoustic data within the visual reconstruction framework.

Despite this, there is significant value in the ability to capture time-synchronized data from both optical and acoustic sensors in a controlled setting which will directly benefit a range of problems related to fine-scale sensing and reconstruction in underwater environments. At a trivial level, the data sets collected within the project, and similar data sets can be immediately applied to address monocular- and stereo visual, sonar-only, and fused vision-sonar reconstruction. Moreover, while we did not structure our scenario to require realtime position information during reconstruction, the ability to develop an accurate sensor track in post-processing allows the synthesis of positioning information for playback. This in turn enables testing of algorithms (in a playback mode) which rely on external positioning information for reconstruction (as in e.g., Guerneve et al., 2018). This is particularly powerful as positioning on ROVs typically relies on a combination of an IMU and acoustic sensors which are both complex and difficult to emulate in a constrained tank environment.

Similarly, the capacity to efficiently generate testing data is useful given the rise of machine-learning based data processing approaches. Such supervised algorithms require large sets of labelled data for training. When the cost of collecting sample data is high (e.g., with underwater video), the paucity of exemplar data handicaps algorithm development. As an example, data captured during this project has been used to test a proof of concept in pre-training object recognition algorithms before operating in novel environments. Given a target object, data captured in the test tank can be used to train a network to detect that object under the imaging conditions *present in the tank*. How well does that network function when searching for the object in other locations, potentially with different levels of particulates, ambient lighting, etc? Can the process be improved either through artificial augmentation of the training data (simulating different underwater imaging conditions) or through in situ recalibration of the object detector algorithm itself, recognizing and adapting to the differences between the training circumstances (test tank) and the current conditions (open water) without affecting the underlying object detection system?

As such, there are multiple steps forward which leverage the existing investment in testing infrastructure and expertise. Continued refinement of the optical and opti-acoustic reconstruction techniques investigated in this project is the clearest avenue for continued work. In addition to the core problem of integrating sonar data into the optical SLAM framework, a closely related problem is quantifying the degradation of the optical reconstruction as water clarity declines. Not only is this important for defining the boundaries of usability for optical reconstruction underwater, it also addresses the problem of determining when optical reconstruction is likely to fail (based on input data), or has failed (based on output data), which in turn is critical to any algorithm which opportunistically switches between the modalities based on the relative strengths or qualities of the data at any point in time.

An alternative is to redouble efforts to support autonomous manipulation solely with acoustics. While

clearly more challenging, any such approach would naturally be applicable to a broader range of turbidities. A natural extension would continue to investigate approaches to reconstruction from imaging sonars. While there have been a number of successes in this space in recent years, strong caveats remain about the use of many such algorithms, including a requirement for precise positioning information, or for travelling in particular trajectories or sensor motions. As previously outlined, the existing infrastructure allows straightforward collection of large quantities of imaging sonar data which can be correlated to sensor trajectory using either the current strategy of photogrammetric reconstruction or through the installation of improved motion capture infrastructure on the motion gantry.

Contents

List of Figures

List of Tables

Acknowledgements

The author would like to thank Mitchell Scott who contributed significantly to the development and analysis of results presented. He would also like to thank the students who assisted with the hardware development including Aleah Deschmidt, Echo Wood, Stefan Layanto, Tanner Poling, and Karl Skeel.

Keywords

Unexploded ordnance, simultaneous localization and mapping, stereo vision, imaging sonar, forward looking sonar.

Acronyms

SDI	Serial Data Interface
SLAM	Simultaneous Localization and Mapping
SfM	Structure from motion
UXO	Unexploded ordnance
UW-APL	University of Washington Applied Physics Laboratory

Chapter 1

Introduction

Due to the crushing pressures found at depth, work in the deep ocean relies heavily on remotely operated vehicles (ROVs), tele-operated robotic platforms which carry optical and acoustic sensors, manipulators, and other tools for observation of and interaction with the benthic environment. Despite their critical importance to subsea operations, ROVs remain technologically very simple, passing realtime sensor data over a high-bandwidth umbilical to operators on a surface support ship, who then issue low-level motion commands back to the ROV's thrusters and manipulators. While interaction with the environment is a key ROV capability, limitations in sensor fidelity, the quality of feedback from the manipulators, and limited control ergonomics mean the viability of delicate, dextrous tasks depends largely on an individual operator's skill. These limitations, as well as the capital and operating costs of such systems, reduce their competitive advantage relative to divers for many shallow water tasks, except in scenarios with high operational risk.

Given the inherent risk to divers in UXO remediation, there is a clear opportunity for adapting these marine robotic technologies to UXO response. However, given the importance of delicate handling in remediation, the current state of the art in tele-operated manipulation is insufficient. Preliminary results have shown the benefits of computer guidance or assistance for ROV manipulation (Rydén et al., 2013; Choi et al., 2015), building on a broad base of research into the use of intermediating or augmentative computer control to improve the precision and performance of remote manipulation in applications ranging from telesurgery (Taylor et al., 2016) to exoplanetary exploration. Such assistive control can range from near-transparent systems where the operator retains full control while operating within an envelope of computer-defined "safety stops" which prevent known dangerous actions, up to fully autonomous control where the computer directs the manipulator and the operator provides only high-level, supervisory inputs.

Computer-assisted or robotic manipulation in an unstructured environment requires that the system sense the geometry of the objects to be manipulated as well as the surrounding environment as a basis for understanding current state, planning future actions, and detecting unexpected or fault behaviors. For seafloor munitions response this translates to the generation of digital models of UXO in place on the seafloor of sufficient resolution and accuracy to allow algorithmic identification of the object and execution of a remediation plan by a robotic manipulator.

The collection of such models underwater is non-trivial. The physics of both light and sound in water, as well as the engineering and economic challenges inherent in the development of marine sensors, lead to a paucity of turn-key underwater 3D sensing technologies. Further, given the shallow depths and frequently soft bottom types in areas of greatest concern for UXO remediation, water clarity is often severely limited,

handicapping optical techniques. Finally, any perceptual products must be available in near-realtime, as this determines the delay between observing the scene and taking action; and it is highly desirable that the model be dynamic and readily updated. An ideal sensor would provide provide a dense 3D model of the scene with a refresh rate comparable to the physical dynamics of the world (e.g., video rate); short of that, it should be feasible to update the model on demand with minimal delay.

Chapter 2

Objectives

This project explores a specific approach to this problem, using a combination of stereo cameras and a 2D imaging sonar to construct 3D models of objects on the seafloor without an explicit requirement for a precise external position estimate (e.g. from an navigation-grade IMU), driven by an end goal of providing perceptual input for robotic manipulation of UXO on the seafloor by an inspection-class ROV. This in turn drives a specific usage scenario, from which a set of requirements can be derived, as detailed in Table 2.1. As discussed in later chapters, these missions-specific requirements are often at odds with the assumptions made in more general 3D reconstruction / SLAM research, and a greater understanding of this divergence in expectations is a specific outcome of this project.

Despite the caveats given previously around water clarity, turbidity may not always completely occlude cameras. When optical data *is* available, it is likely to be the preferred modality as it contains greater information density about the object and environment. Moreover, the mathematical and algorithmic bases for vision-based 3D reconstruction are far more mature than the associated development for acoustic sensors. Starting with an optical approach addresses a subset of conditions where vision-based solutions are appropriate and allows immediate contribution to the larger robotics problem. Acoustic data can then be introduced incrementally, building on the algorithmic base provided by the optical solution. An ideal system would use both optical and acoustic data at all times, balancing the data from both modalities to achieve the best model.

Building from this goal, this project has four major phases:

- **Development of an optical-acoustic data collection system** which emulates the sensor front end of an ROV (Chapter 3). A core requirement is the ability to move in a manner representative of an ROV surveying an object on the seafloor. However, this trajectory must be independently measurable to allow evaluation of the accuracy of the trajectory calculated from sensor data. A free-floating sensor package (or a small ROV) would allow more realistic high-degree-of-freedom motion, but would also be expensive to construct, more complex, and would not offer a straightforward method for accurate direct pose measurement. Instead, a motion gantry system was developed which allows an orbiting motion around objects of interest. While constrained, this trajectory is more repeatable and more easily instrumented for ground truth.

  A second critical requirement is ensuring the data from the two sensor modalities are correctly time-synchronized and available in realtime for processing by the reconstruction algorithm. This necessitated

Scenario Phase	**Derived Requirement**
1. The remediation platform (ROV) locates the target UXO and approaches to within manipulation range (approx. 1 m).	*This project does not focus on the task of long- to medium-range identification and localization of UXO, although the forward-looking sonar used in this project would be suitable for this task, and the data collected in this project would be suitable for supporting this research.*
2. The platform executes a series of maneuvers (e.g., orbiting, examining from multiple angles) to construct a 3D model of the UXO sufficient to segment the UXO from the background, identify grip points and plan manipulation. This period of observation may be subject to spatial and temporal variations in the water clarity.	The system will construct a 3D model of a relatively constrained volume (approx 1m cube) given relatively few distinct views from a constrained set of viewing locations. The model will be generated using all available data, subject to the water clarity. The model will be generated in realtime or near-realtime to allow immediate action.
3. If the platform is free-floating, it will require a continuous estimate of its position relative to the object.	The model should provide a low-latency, high update rate estimate of platform-to-object position through both the mapping and manipulation phases.
4. The system will be compatible with lower-cost ROVs, including those with low-grade or no attitude and velocity measurement capability.	The model will make no assumptions about the accuracy or availability of an external pose or velocity input.
5. The platform will then execute a manipulation task (either through assistive tele-operation or autonomously).	The system should continue to update the model at a rate sufficient to monitor progress of the manipulation or to detect faults.

Table 2.1: Remediation mission concept of operations and derived requirements for sensing system.

the development of a suite of custom data recording tools rather than use of the standard clients provided by the sensor manufacturers.

- **Collection of realistic test data** of objects of interest including calibration shapes, generic objects and UXO stand-ins (Chapter 6). The importance of the collection of thorough, well-documented sample data sets was underlined throughout the program, and a critical outcome from this project was a focus on collecting and curating reference optical-acoustic data sets to support future research.

- **Development of inter- and intra-sensor reprojection models (Section 4).** The reprojection model of each camera on its own and as a stereo pair can be calculated using computer vision best practices. Similarly, we assume the imaging sonar is metric given an accurate sound velocity, such that the bearing and range to any given sonar bin is correct. The question of measuring and estimating distortions in the sonar model, while interesting, was deemed out of scope. A challenging component is the estimation of the physical offset between the origins of the camera(s) and of the sonar, which allows the reprojection of world points between sensor frames.

- **Development of an opti-acoustic reconstruction algorithm (Section 5).** The selected approach is based on an existing open source realtime visual reconstruction algorithm, adapting it for the particulars of the problem, including the addition of depth information from stereo cameras, and inclusion of data from the imaging sonar. As executed, this phase is divided into three sub-tasks:

  - Adoption and refactoring of LSD-SLAM, an open source 3D reconstruction algorithm.
  - Extension of LSD-SLAM to incorporate stereo video.
  - Extension of LSD-SLAM to incorporate sonar data.

2.1 Existing Sensor Technologies

To place our solution in context, it's necessary to briefly review the current state of the art in manipulation-range subsea sensors. Here, subsea sensing technologies are classified into three major categories, defined not by by their fundamental medium of operation (e.g. optical versus acoustic) but by their conceptual mode of operation. Further details on specific implementations, as well as differing opinions on the appropriate taxonomy for such sensors, can be found in recent survey papers (e.g., Massot-Campos and Oliver-Codina, 2015).

2.1.1 Direct structure sensors

Direct structure sensors are capable of efficiently measuring full 3D scene structure, either by capturing the entire scene directly with each sampling of the sensor or by rastering a sensor at high speed over the scene. Terrestrially, this category includes a variety of accessible, and increasingly ubiquitous technologies including structured light and time-of-flight sensor, as well as steered LiDAR. Although there are a range of caveats and limitations on the use of each, the sheer usefulness of these types of sensors for a range of mechatronic applications (e.g., robotics, autonomous vehicles, AR/VR) will continue to exert strong market pressure towards more capable, less expensive implementations.

Underwater, the strongest disincentive for the direct adaptation of terrestrial implementations for marine use is the high absorbance of the near-infrared wavelengths commonly used in both structured light and

LiDAR sensors — NIR being used because it is not visible by the human eye, and thus not distracting to human observers, while still being readily observed with standard CMOS camera sensors. Working around this limitation, both structured light (Bruno et al., 2011; Inglis et al., 2012, e.g.,) and LiDAR (Tetlow and Spours, 1999; Hildebrandt et al., 2008, e.g.,) have been demonstrated subsea using visible spectrum light. This approach introduces the significant disadvantage of broadcasting a visible, often strobed pattern onto the scene, a severe handicap for human operators trying to simultaneously interpret the video, but potentially less of a constraint for robotic systems. Despite the potential of such sensors, none have achieved commercial success.

This category also includes unassisted (without an explicit structured light pattern) stereo vision, although the strong interdependence between image quality and depth map quality, as well as the relatively high computational requirements, constrain its adoption. However, as stereo requires only ubiquitous, low cost sensors, there is strong pressure to develop and refine robust vision-based stereo algorithms.

There is little direct impediment to deploying stereo depth estimation underwater, as subsea cameras are relatively ubiquitous. However, the nature of much subsea imaging is directly at odds with the needs of dense stereo estimation, as underwater imagery is frequently ill-lit with low contrast or color differentiation (McGlamery, 1975), and large areas of little or no texture. Despite this, stereo reconstruction has been applied successfully in a number of cases, frequently in deeper water where water is more reliably clear (Negahdaripour and Firoozfam, 2006; Hogue et al., 2007; Johnson-Roberson et al., 2010; Oleari et al., 2015).

Moreover, all three of these technologies are inherently optical and will suffer degraded performance in turbid waters to some degree, severely reducing their utility in the shallow water scenarios highlighted in the MR Statement of need. Unfortunately, the nature and rate of degradation with turbidity for each modality has been studied, at best, qualitatively. Of the three, structured light is likely to degrade most quickly as the projected pattern will scatter off suspended matter in the water column. Similarly, camera-based methods like stereo will suffer from both backscatter effects and the decreased object definition from the occluding particulate matter, although the former can be mediated somewhat by increased separation between the lights and cameras.

The most viable non-optical alternative to these approaches is the used of multibeam sonar capable of beamforming a 2D receive array, a COTS example of which is the Echoscope line from Coda Subsea. In its current instantiation, the Echoscope does achieve the stated goals of developing a direct depth map for each beam over a 2D field of view, at greater-than-1 Hz rates. However, the Echoscope features a low resolution of 128x128 beams over a $50\,\text{deg} \times 50\,\text{deg}$ field of view. Initial, promising attempts to generate 3D structure underwater with the Echoscope can be found in Davis and Lugsdin (2005) and Lagudi et al. (2016). The Echoscope is also by far the most expensive single sensor discussed here.

2.1.2 Profiling sensors

In contrast to direct structure sensors, profiling sensors measure scene structure directly, but only over a 1D field of view (along a line or swath). Optical approaches include 1-D rastered LiDAR which actively measure the time of flight of the laser pules, as well as similar laser line disparity sensors which optically measure the parallax-induced offset of a laser line projected onto the scene.

While technologically distinct from the optical sensors, high-frequency profiling multibeam sonars such as the Blueview BV5000 are comparable in net effect. These sonars employ a linear array of receive elements to generate a fan-shaped receive pattern with a broad field of view (45-120deg) in one dimension, which is beamformed into a number of distinct beams; and a narrow (typ. 1-5 deg) aperture in the orthogonal

direction. By measuring the range to the strongest return, the sonar is able to estimate the distance to a surface along each beam in the profile line. Like 2D imaging sonars (discussed below), the system is unable to resolve the angle of a return in the orthogonal axes, but due to the narrow aperture, each beam's range measurement relates to a relatively small physical patch on the reflecting surface. While sonar have lower spatial and depth resolutions compared to laser-based solutions, they are unaffected by turbidity, although they are impacted by the ambient acoustic conditions, multipath and the acoustic response of the object being scanned.

The fundamental disadvantage of both types of profiling sensors is the need to scan the sensor, typically mechanically, across the scene to produce a full 3D map, with across-track resolution determined by the ratio between measurement repetition and scanning speed. For the laser line and Blueview sensors mentioned above, the practical scan speeds are limited to $1 - 2\,\mathrm{deg/sec}$, requiring e.g., 45-90 seconds to complete a quadrant scan. Moreover, joining the discrete scans into a full 3D model requires knowledge of the sensor position for each ensonification as it moves across the scene. In the simplest cases, the sensor is mounted to a stationary seafloor tripod during the scan, although point cloud assembly can also be accomplished when the motion and pose of the platform is known with great precision (e.g., with a high-grade inertial navigation system).

In some cases, a lengthy, stationary sample acquisition can be accommodated in the mission concept of operations; and at present this is the most practical COTS approach to 3D structure estimation. However, this approach is far from dynamic and may only be appropriate for generating "ground truth" data sets. These sensors are generally unable to provide active updating of scene structure or vehicle position *during* during dynamic maneuvers, for example while a manipulator is approaching, grasping, and moving a UXO.

2.1.3 Decimating sensors

The final category includes sensors which have an inherent ambiguity in their capture of scene structure, a category which includes both single cameras and so-called 2D imaging sonars. Cameras inherently decimate the 3D world to a 2D image by discarding information about the depth (distance from sensor to world) for each pixel. Both stereo imaging and structure-from-motion (SfM) techniques attempt to recover this missing depth information by exploiting parallax between multiple camera images from multiple cameras (in the case of stereo) or multiple viewing angle (in the case of SfM). While both techniques can be successful in ideal circumstances, such processing depends highly on the quality of the captured images and properties of the reconstructed objects.

Like the profiling multibeam sonars described previously, 2D imaging sonars also feature a 1D receive array, again with a wide field-of-view in one axis which is divided into a number of beams. In contrast, however, the orthogonal direction has a broad aperture (typ. $10 - 20^\circ$ vs. $1 - 2^\circ$ for profiling sonars). A full receive strength-versus-time delay sequence is collected for each beam, which can be treated as a receive strength-versus-range measurement given the speed of sound. By pointing the sensor obliquely at a surface (e.g. the seafloor), this can be interpreted as a polar signal-versus-range "image." However, as the signal strength at a given range is integrated over the full cross-array aperture, the sonar is unable to localize the angle of return along the orthogonal direction (this ambiguity is described further in Section 4.1). Resolving this ambiguity is the essential problem in extracting 3D structure from such imaging sonars (Guerneve et al., 2018). Relative to other sensor modalities (e.g. vision), processing of imaging sonar is also challenging due to the complexities of the underlying physics e.g., aspect-dependent acoustic response of different targets, multipath, and sidelobe effects.

2.1.4 The Role of Platform Motion and Motion Measurement

A common factor underlying much of the discussion of sensor options is the role of sensor/platform motion in expanding or extending sensor field of view. For some sensors (e.g. profiling laser and sonar systems) this motion is essential for generating a usable 3D model, while for others (e.g. structured light sensors) motion provides greater coverage and potentially redundant views to improve scene quality but is not essential for generation of a 3D model.

Systems must be further differentiated between cases where vehicle motion is measured precisely (of a similar order of magnitude as the resolution of the exteroceptive sensors), imprecisely, or not at all. The first case vastly simplifies the association of multiple sensor readings into a synoptic view, as it allows straightforward reprojection of multiple readings into a shared coordinate frame. In some cases, for example, the use of profiling sonar from fixed tripods, measurement of this motion is inherent to the sensor concept of operations. In other cases, the requirement for precise motion estimation can be passed along to the sensor platform, leading to, for example, the need for survey-grade inertial navigation systems on many platforms used for synthetic aperture sonar measurement.

The latter two cases (imprecise or no motion measurement) are in fact closely related, as they place the burden of association on the data itself, with imprecise motion estimates offering a guiding prior on vehicle motion, but not of sufficient quality to obviate the need for data-based association. Despite the challenges inherent in performing data-based association, it remains an appealing prospect due to the weight, complexity and expense of precision inertial navigation systems. Similarly, making no assumptions about the availability of pose information minimizes the requirements for successful operation of an algorithm, as it can be run on any platform of opportunity with the correct sensors. Within this project, we have taken this latter approach, adopting an approach which derives sensor motion from the sensor data itself. In practice, a platform performing subsea UXO remediation would likely have a lower-grade MEMS IMU and a doppler velocity log, giving some estimate of vehicle motion, which could be integrated into the SLAM algorithm.

2.2 Discussion of Sensor Choice

At present, there is no practical, low-cost COTS sensor for the measurement of 3D scene structure underwater, particularly in mixed or turbid conditions, with every option carrying a significant compromise. For example, the Coda Echoscope solves the problem directly, albeit at limited resolution, but incurs significant cost and complexity. Profiling sensors, particularly the Blueview-style profiling multibeam sonars can also be used to generate 3D models at a moderate cost, but the operational challenge of needing to mechanically scan the sensor, and potentially a requirement for precise vehicle positioning information, severely constrains it utility. Finally, cameras and imaging sonars offer the lowest cost solution, with acceptable data resolution and update rate, but require significant, complex post-processing (even for the well-established optical case) to arrive at a 3D reconstruction.

Of all possible options, structured light is the most likely to make a significant commercial impact, however its application will be limited to deep water, low-turbidity applications. Given this, as well as the challenges with using a visible projected pattern while also relying on video for human control, the first practical application of subsea structured light sensors will likely be on robotic or autonomous platforms which do not have require human-usable video. However, for the munitions response application, we believe structured light is of little use given its potential for degradation in turbid waters.

Of the remaining options, the decimating sensors are the lowest cost, most readily available option.

They are also the only option which require an algorithmic advance, rather than an advance in instrument design or overcoming a physical constraint, to become practical. Given the strength of existing optical 3D reconstruction techniques, and the assertion that both optical and acoustic data will be available in many operational scenarios, a hybridized blending of the two was deemed the most appropriate path for development.

Chapter 3

Data Collection System Design

A core element of this project is the design of an apparatus for collecting time-synchronized stereo video and imaging sonar data. This apparatus is coarsely split into two sub-assemblies, a "sensor head" which contains sensors to approximate the front end of an inspection class ROV, and a "gantry" which allows the sensor head to be placed or moved relative to a test object in a test tank. Use of a true ROV was considered, as this would allow true high-degree of freedom motion and provide a platform for subsequent development in open water, however this approach was abandoned due to the significant increase in cost and complexity, as well as the difficulties in accurately measuring the pose and trajectory of the ROV, even in controlled or constrained spaces.

3.1 Hardware Specification and Design

The system is subject to a number of requirements:

- The sensor head shall contain a pair of video cameras operating at HD (or better) resolution and video frame rates ($\geq$ 25 Hz).
- Due to the negative effects of block compression artifacts on computer vision algorithms, the video data shall be collected in a lossless or near-lossless format. The cameras will be frame-synchronized to within 1/120th of a second.
- The sensor head shall contain a imaging sonar suitable for imaging objects at 20cm–2m range.
- The cameras and sonar shall be rigidly mounted relative to each other, such that the inter-sensor geometry is fixed.
- The sensor head shall contain lights sufficient for imaging objects 2m in front of the cameras.
- The sensor head and all components shall be designed for operation in salt water — given this program focuses on tank testing, a maximum operating depth was not set although having the ability to reuse components in later open water testing was taken into consideration.
- The gantry will allow the sensor head to travel through a path which allows inspection of the object of interest from multiple perspectives.
- The gantry will constrain that motion insofar as it makes estimates of camera trajectory more tractable.

Component	Manufacturer	Model
Imaging Sonar	Blueprint Subsea	Oculus M1200d
Cameras	BlackMagic Design	Micro Studio Camera 4k (qty 2)
Lenses	Olympus	M.Zuiko Digital 17mm f/1.8 Lens
Lights (Rev 1)	APL	Internally developed
" (Rev 2)	Big Blue Dive Lights	CB10000P (Qty 2)

Table 3.1: Key components for sensor head.

Mode	1200 MHz	2100 MHz
Range (max)	30m	10m
Range (min)	0.1m	
Range resolution	2.5mm	
Update rate (max)	40Hz	
Horizontal aperture	130°	60°
Vertical aperture	20°	12°
Angular resolution	0.6°	0.4°
Beam separation	0.25°	0.16°
Number of beams	512	

Table 3.2: Key specifications for Oculus M1200d imaging sonar.

3.1.1 Sensor Head

Sonar Selection

The sensor head is designed around the constituent sensors. The imaging sonar is a Blueprint Subsea Oculus M1200d, which offers both 1.2 and 2.1 MHz modes. Core specification for the sonar are provided in table 3.2.

The sonar has a single connector which provides power and a 100Mbit ethernet interface. The manufacturer provides a simple API which documents the network packets required to initiate pinging, and the format of the returned ping data. As detailed below, this protocol information was sufficient to develop custom sonar data recording tools.

In this project, the sonar was only used in the high frequency (HF) 2.1MHz mode. As shown in Table 3.1, in HF mode, the sonar has a horizontal field of view of 60 degrees consisting of 512 beams on 0.16° centers, each with an approx. 0.4° by 12° aperture. Although the sonar lists an update rate of 40Hz, in our testing, we observed data at approximately 4Hz – as this was sufficient for initial development we did not investigate the reasons for this mismatch nor attempt to drive the sonar at a faster rate.

Sonar data is returned in a tabular form consisting of receive strengths as a function of beam number (enumerated left-to-right) and range bin (near-to-far). Additional meta-information provided in each ping data packet allows calculation of the center bearing for each beam relative to directly ahead from the sensor, and the central range (in m) for each range bin. Either the speed is sound or the water salinity is provided to the sonar within the network packet which requests the sonar to start (or continue) pinging. In the latter case the sonar uses a built-in temperature sensor to calculate the speed of sound.

Camera Selection

The selection of cameras was driven in large part by the suitability of passing high-bandwidth video data through submersible cabling. Commercially, cameras of HD resolution (or higher) are available with a number

of different data buses, including gigabit ethernet, HDMI, USB3, and Serial Data Interface (SDI). Of those, HDMI and USB require multiple high speed differential signal pairs and there is little commercial support for passing such signals through the ubiquitous bulkhead connectors used in the subsea industry. Further, neither HDMI nor USB is designed for long distance transmission, leaving concerns about maintaining signal integrity between the sensor head and topside recording equipment located at a suitable standoff distance from the test pool. This could have been ameliorated by introducing an intermediating data converter to the sensor head itself, e.g. a single board computer to capture video from USB3 cameras and re-export it over ethernet; although the benefit would not balance the increase in complexity unless there was a projected need for processing data *in situ* within the sensor head.

The relative popularity of ethernet as a fieldbus (including for the Oculus sonar) has lead to a range of subsea ethernet connector and cabling options. Unfortunately, the data rate of uncompressed HD video exceeds the capacity of gigabit ethernet, with "raw" 1080p30, or similar from an ethernet machine vision camera, requiring 1.5Gbit/sec. In practice, true raw pixel values would not be transmitted over ethernet, instead relying on a chroma subspace sampling, or a lossless video codec; however, standard Gigabit ethernet leaves little margin for greater-than-HD resolutions or greater-than-video frame rates, which may be desirable depending on camera selection.

Serial Data Interface (SDI) is a physical layer standard within the video industry for the transmission of broadcast video over 75Ω coaxial cable. Standard High-Definition SDI (HD-SDI) has a bandwidth of 1.485 GBit/sec, with newer standards defining higher-speed multiples of 2.970 GBit/sec (3G-SDI), 6 Gbit/sec (6G-SDI), and 12 GBit/sec (12G-SDI), with the latter two sufficient for streaming "4k" (3840x2160) video at 30 and 60 frames per second respectively. Moreover, these exists a mature industry in SDI cameras, repeaters, recorders, etc. designed for professional video production. Further, given appropriate cabling, SDI is robust over distances of $\approx$ 100m.

From the perspective of suitability to underwater use, there are a selection of 75Ω coaxial wet connectors, although they are sufficiently specialized to command a price premium. Current COTS connectors do not have sufficient bandwidth for the higher SDI rates, with nominal bandwidth of 1.5Ghz, however, the potential for developing a higher-bandwidth coaxial connector seems feasible given the availability of lower-bandwidth coax connectors. There are also standards for the conversion of SDI to fiber, which is another option for transmission through wet cabling, although outside the budget of this project.

On this basis, the sensor head was designed with SDI cameras. The SDI signal is passed through Belden 4505R 12G-SDI-rated cabling which has been overjacketed with polyurethane. For cost and simplicity, the SDI cables were not connectorized in this design and instead are potted into fixed bulkheads on the camera housings. While inconvenient for handling and repair, this defers dealing with the unavailability of appropriate connectors with the understanding that a design path exists for employing future high-bandwidth coaxial wet connectors, though not within the limited engineering budget of this project.

This project utilizes a pair of Blackmagic Design Micro Studio 4k cameras. These cameras are compact, designed for remote use (e.g., it does not have a viewfinder or screen), offer native 4k resolution at 30fps, and use commercially standard Micro Four-Thirds photographic camera lenses. Video is output at up to 6G-SDI depending on resolution and framerate, and the cameras offer a unique SDI-loop feature where synchronization and configuration information can be sent *to* the cameras over a second "input" SDI link.

The SDI input and output for each camera is handled by an interface card in a data acquisition computer. Due to the transmission characteristics of SDI, this computer can be located at a distance from the camera head without loss of signal integrity. The role of this computer in capturing video, and control-

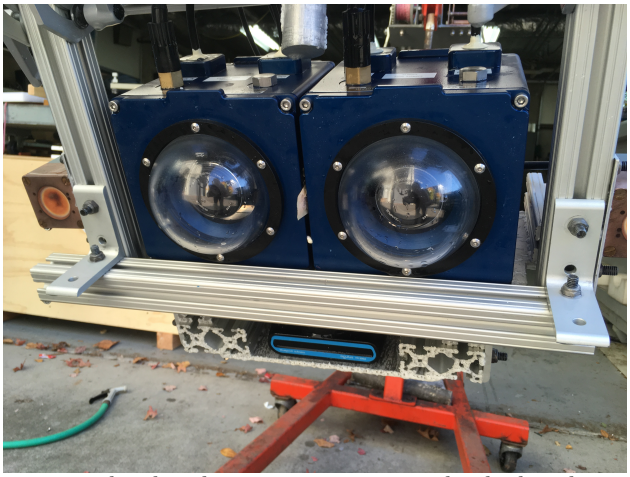


Figure 3.1: Complete sensor head with stereo cameras in individual enclosures and sonar below.

ling and synchronizing the cameras is discussed in section 3.2. The SDI data is passed through an input switcher/compositor and preview monitor which allows live preview of the video independent of the computer (see Figure 3.2).

The cameras are individually housed in aluminum enclosures designed and manufactured by The Sexton Co., utilizing their 4.5" acrylic dome port (figure 3.1) with a nominal depth rating of 100m. In addition to the two SDI *in* and *out* penetrations, a 4-pin Subconn rubber molded connector is used to power the camera. An internal rail system allows adjustment of the fore-aft positioning of the camera and lens to accommodate different lens lengths.

3.2 Software

The camera and sonar sensor head is supported by a suite of software interface libraries developed for the program, all of which are released under the Open Source MIT License. The particulars of each package, along with a link to its repository are included in Appendix A.

The overall goals of the software components are to:

- Provide an interface for viewing camera and sonar data in realtime and to make configuration changes to those instruments as needed e.g. adjust focus, resolution, max range, ping rate, etc.
- Accurately record the timing of each frame of data from each sensor.
- Record the data in a manner and file format which preserves the timing and sequencing of each data packet (video / sonar frames).
- Play both realtime and recorded video data through reconstruction tools.

The resulting software utilizes a unified data file which contains both the timestamped video and sonar data. In this way, the relationship between the data streams is preserved up to the point of playback, however doing this necessitated development of a specialized stereo-video-and-sonar data recording test suite. As described previously, the sonar communicates via TCP/IP packets over the sonar's ethernet interface. Vendor-supplied code was readily adapted to build a sonar parsing and processing toolkit.

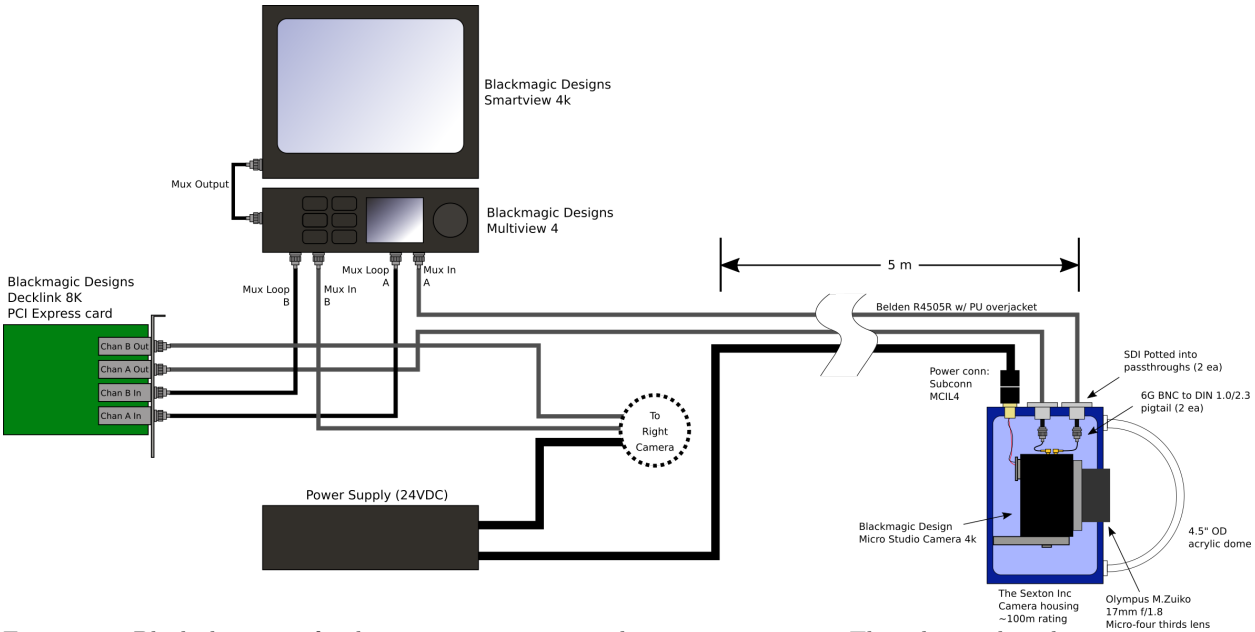


Figure 3.2: Block diagram of video acquisition system, showing one camera. The ethernet-based sonar acquisition system is not shown.

The cameras require a suite of software packages. As noted previously, the cameras output raw video in the SDI data format, which is captured by an SDI interface card installed in the data acquisition computer. Existing sample code was expanded to capture stereo imagery from the two cameras. The cameras are also remotely controlled and synchronized by SDI *output* from the interface card and *input* by the cameras via the the Blackmagic Design SDI Camera Control protocol, a proprietary protocol which inserts command data into the normally invisible inter-frame periods in the data. This protocol allows remote control of the camera and lens, including shutter and aperture, focus, and camera resolution. Further, the cameras are configured to synchronize their internal frame acquisition to this input SDI signal, providing the required frame synchronization between the two cameras.

A front-end application provides the main realtime features, including preview of the sonar and video data, remote configuration of the sensors, and the ability to record data. This version of the software records the data in a multi-track Quicktime container. The left and right video data is recorded in parallel Prores-encoded video tracks within the file, allowing that data to be viewed and extracted using standard video editing tools. The sonar data is encoded in a timestamped data track which is interleaved with the video data, creating a single file which efficiently stores all three data streams.

Later in the project, a second workflow was developed which adapts the sensor interface tools to the Robot Operating System (ROS). To accommodate ROS, the existing hardware interface packages were modified to output sensor data onto the ROS data bus, with the left and right video data published using the standard ROS image topics, and the sonar data published on a custom imaging sonar message type. By publishing this data and exposing the relevant control parameters, the sensor data can be previewed using the standard ROS data viewing tools (e.g., `rqt`), with a sonar-to-image node developed for sonar preview; and the sensor parameters can be controlled through the ROS *dynamic reconfiguration* mechanism. While using the ROS package, data can be recorded, and later played back, using the integral `rosbag` tool.

3.3 Gantry Design

A key component of the system design is the ability to move the sensor head in relation to the target object in a manner which approximates an ROV inspecting an object on the seafloor. At the same time, the desire was to keep any support structure simple and low cost. On the resulting gantry, the camera/sonar head is mounted in a manner which allows adjustment in both the pan and tilt direction. The sonar head is then mounted on the end of a $\approx$ 1 75cm radial arm which can pivot around a central pivot (an iteration of the design can be seen in Figure 6.1). Counterweights opposite the camera head minimize torquing moments on the central pivot. This pivot, in turn, attaches to a linear slide which runs along a rail. In total, this gantry provides four degrees of freedom, three of which can be changed dynamically during data collection:

- Linear position of central pivot along rail (can be locked out)
- Radial arm yaw around central pivot
- Sensor head yaw (relative to arm)

And one which is adjustable but typically fixed in place during operation:

- Sensor head pitch

In addition, the radius of the camera along the arm is adjustable, though not trivially. For simplicity, these degrees of freedom are purely manual with no powered actuation. In practice, the linear position was difficult to adjust in situ and was not used in later data collection, with the radial motion becoming critical to capturing full views of objects.

Chapter 4

System Calibration

Fusion of the data from the three sensors (two cameras, sonar) requires an understanding of the reprojection model for each sensor type as well as the physical relationship between each sensor. This information allows the reprojection of world points into each sensor frame and the transfer of points between sensors.

4.1 Sensor Models

The cameras are modelled using the standard pinhole camera model combined with the non-linear radial-tangential distortion model used in OpenCV / ROS. The camera-to-camera stereo alignment is modelled as a rigid body transformation between camera centers. Both of these quantities are standardized within computer vision and can be estimated using fiducial objects (e.g. checkerboards) using well-documented procedures (Zhang, 1999, e.g.,). As the two cameras are rigidly mounted to the sensor head, cameras calibrations were calculated in water once and used consistently for further analyses.

For the sonar we adopt the imaging model of Guerneve et al. (2018). The sonar operates from an origin point S, a *sonar center* comparable to the camera center used in the pinhole model. This sonar center has an attached polar coordinate system (r, θ, ϕ) where r is a range/radius, θ is the bearing of each formed beam, and ϕ is the aperture across the beam pattern. We adopt the convention of the Oculus sonar such that zero bearing is directly ahead of the sonar, with positive bearings to starboard.

The formed image is a projection onto the (r, θ) plane which consists of pixel intensities $I_s(r, \theta)$ over some field of view $\theta \in [\theta_1, \theta_2]$ and max range $r \in [0, r_{max}]$ (although there is a practical minimum range for the sonar, data is typically presented to zero range). The intensity of each pixel $I_s(r, \theta)$ is approximated by the sum of reflected intensities over the cross-beam aperture $[\phi_1, \phi_2]$

$$I_s(r, \theta) = \int_{\phi_1}^{\phi_2} \beta(\theta, \phi) V_s(r, \theta, \phi) \frac{\vec{v} \cdot \vec{n}_{r\theta\phi}}{\|\vec{v}\| \|\vec{n}_{r\theta\phi}\|} d\phi$$

where $\beta(\theta, \phi)$ is the angular distribution of energy known as the *beam pattern*, $V_s(r, \theta, \phi)$ is the albedo of the reflector, and the final term is the cosine between the direction of propagation of the beam $\vec{v}$ and the local surface normal of the object $\vec{n}$. This is a simplified model which neglects e.g., object reverberation, multipath, etc., but captures the decimation performed by the sensor in the ϕ axis.

In practice, the user-accessible sonar data is a further mapping of those reflected intensities $\hat{I}(r, \theta) = F(I(r, \theta))$, where the function $F(\cdot)$ includes signal conditioning or smoothing, as well as constant or range-

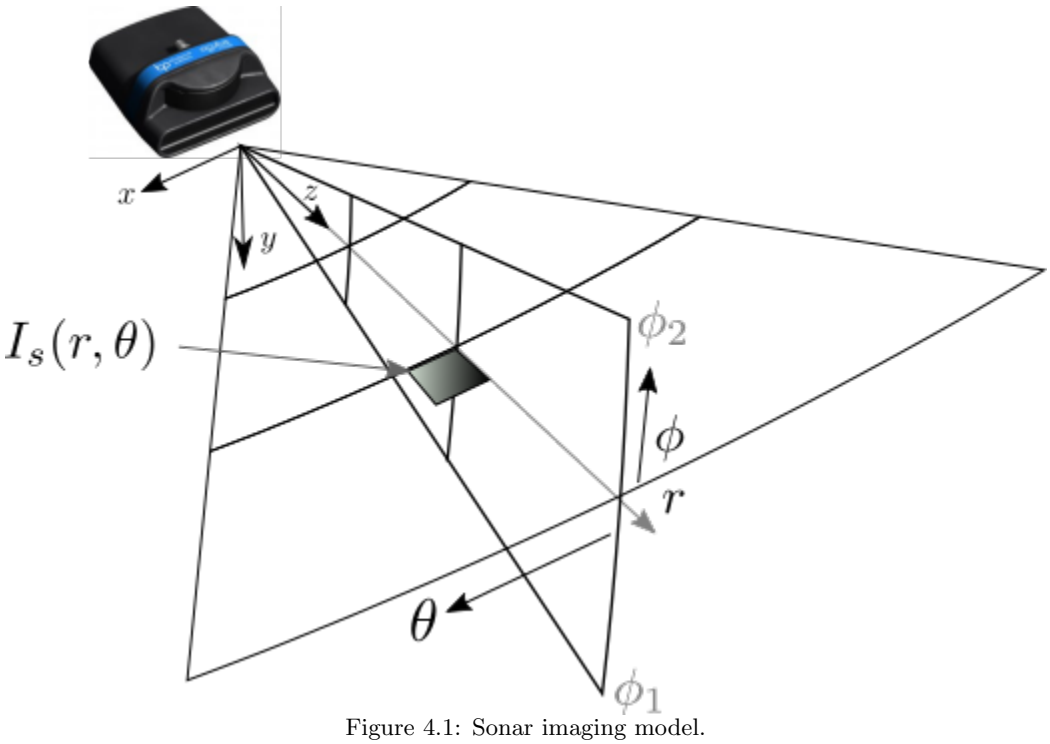


Figure 4.1: Sonar imaging model.

dependent gains. Moreover, calculation of the range r from the timing of the acoustic returns requires knowledge of the local speed of sound in water, which is typically estimated external to the sonar. For simplicity, this project deals only with the final image $\hat{I}(r, \theta)$ and assumes no error in the calculation of r.

4.2 Camera-Sonar Calibration

An important feature of joint optical-acoustic reconstruction is the registration between the optical and acoustic sensor frames. However, the extrinsic estimation between a camera frame and sonar frame is typically difficult for two reasons: 1) calibration objects are not typically visible and unique in both the optical and acoustic field of view; and 2) the decimation between the sensors varies, i.e. camera senors decimate range while sonar sensors decimate along the elevation axis. To that end, a core activity is the creation of a calibration procedure to estimate the rigid body relationship between the camera and sonar systems.

4.2.1 Calibration Equations

Our calibration process leverages work by Hurtós et al. (2010), which describes a calibration procedure between a monocular camera and multi-beam sonar. In their paper, extrinsic camera-sonar calibration utilizes a checkerboard target, similar to what is utilized for intrinsic and extrinsic camera calibration (see Fig. 4.2).

This extrinsic calibration process aims to calculate rigid body transformation ${}^{s}[\mathbf{R}|t]_c$ between the sonar s and camera c frames. Points in the sonar frame can then be transformed in to the camera frame by

$$X_c = {}^{s}\mathbf{R}_c^{-1}(X_s - {}^{s}t_c) \tag{4.1}$$

where X_c and X_s describe points in the camera and sonar frame, respectively.

To calculate the ${}^s[\mathbf{R}|t]_c$ transformation, this approach assumes a known rigid body transformation, ${}^c[\mathbf{R}|t]_w$, between the calibration target in the world (i.e. w) frame and camera frame. From this transformation, the calculation of N, a vector parallel to the world frame, is trivial:

$$N = R_3(R_3^T(-t)) \tag{4.2}$$

where R_3 is the third column of the rigid body transformation rotation ${}^c\mathbf{R}_w$. If X_c lies on the calibration target, then:

$$N \cdot X_c = |N|^2 \tag{4.3}$$

as N is defined as parallel to the calibration targets normal vector.

From equations 4.1 and 4.3

$$N \cdot R_c^{-1}(X_s - {}^s\, t_c) = |N|^2 \tag{4.4}$$

and by defining a 3×3 matrix H as:

$$\begin{aligned} \mathbf{H} &= {}^s\, R_c^{-1}\hat{t} \\ \hat{t} &= \begin{bmatrix} 1 & 0 & -t_1 \\ 0 & 0 & -t_2 \\ 0 & 1 & -t_3 \end{bmatrix}, \end{aligned} \tag{4.5}$$

equation 4.4 reduces to:

$$N\mathbf{H}X_s = |N|^2. \tag{4.6}$$

Given that the y axis is completely decimated in the sonar frame, X_s is described as

$$X_s = [x_s, z_s, 1] \tag{4.7}$$

and equation 4.6 can be further deconstructed to:

$$ah = |N|^2, \tag{4.8}$$

where $a_{1X9} = [n_1x_s, n_1z_s, n_1, n_2x_s, n_2z_s, n_2, n_3x_s, n_3z_s, n_3]$ and $h_{9X1} = [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}]^T$ if $\mathbf{H}$ is decomposed as

$$\mathbf{H} = [h_{ij}]_{3\times3}. \tag{4.9}$$

Equation 4.8 describes a linear constraint for corresponding points between the two sensor frames. After a series of $n > 9$ lineally independent observations, h can be solved as:

$$h = \mathbf{A}^\dagger B \tag{4.10}$$

where $\mathbf{A}_{n\times9} = [a1, a2, ..., an]^T$, $B_{n\times1} = [|N|_1^2, |N|_2^2, ..., |N|_n^2]^T$ (where $|N|$ is not necessarily unique between

observations), and † is the matrix psuedo-inverse. $\mathbf{H}$ is then trivially solved by casting h back to $\mathbf{H}$ via. equation 4.9.

Finally, $\mathbf{H}$ can be cast into ${}^{s}\mathbf{R}_{c}$, ${}^{s}t_{c}$. Decomposing $\mathbf{H}$ into $[H_1, H_2, H_3]$, it is clear from equation 4.5 that H_1 is rotation around the x-axis, H_2 is the rotation around the z-axis and H_3 is a rotated translation between the two frames. Given that ${}^{s}\mathbf{R}_{c}$ is necessarily orthonormal, and given that equation 4.5 describes a negative translation, ${}^{s}[\mathbf{R}|t]_{c}$ can be reconstructed as:

$$\begin{aligned} {}^{s}\mathbf{R}_{c} &= [H_1, -H_1 \times H_2, H_2] \\ {}^{s}t_{c} &= -{}^{s}\mathbf{R}_{c} H_3. \end{aligned} \tag{4.11}$$

4.3 Calibration Procedure

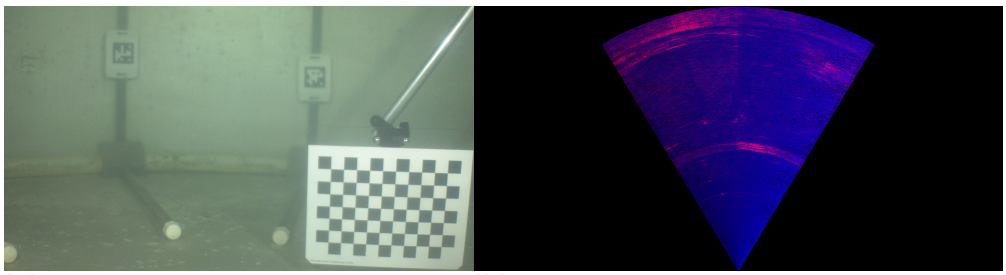
Using the mathematical framework described in section 4.2.1, our camera-sonar calibration procedure required moving a calibration target of known size through a series of unique poses throughout the camera and sonar field of view. At each pose, we would first calculate the transformation of the checkerboard frame to the left camera frame i.e., ${}^{c}[\mathbf{R}|t]_{w}$. This transformation was estimated by manually identifying the four corners of the checkerboard, and identifying a transformation from the current frame pose to the origin pose that satisfied an $SE(3)$ transformation constraint. We utilized this transformation to estimate the projection error between the calibration target and its reprojection to the origin, and only added points with a significantly low reprojection error. After ${}^{c}[\mathbf{R}|t]_{w}$ was estimated (and if the reprojection error was deemed significantly low), we identified the location of the checkerboard frame origin in the acoustic image, i.e. X_{mb} from equation 4.7, through manual correspondence. Once both ${}^{c}[\mathbf{R}|t]_{w}$ and X_{mb} were known for the specific calibration target pose, the $\mathbf{A}$ and B matrices from equation 4.10 were concatenated with the appropriate new pose information, and the checkerboard target was moved to the next pose.

A description of our calibration procedure is shown in Algorithm 1.

4.3.1 Calibration challenges

There were a number of challenges with the described calibration procedure, which made calibration both difficult and tedious. These issues, with some proposed mitigation strategies, are listed below.

1. *Lack of automatic optical image calibration target corner identification.* While the checkerboard target is ubiquitous for the calibration of computer vision cameras, that process depends on the automatic



(a) (b)

Figure 4.2: Appearance of the calibration target in (a) the left stereo-camera optical frame and (b) in the sonar frame

Algorithm 1 Camera-sonar calibration procedure

Inputs:
  Stereo calibration parameters (K,P)
  Re-projection error threshold max_proj_{err}
Initialize:
  Initialize $\mathbf{A}$ matrix from eqn 4.10
  Initialize B vector from eqn 4.10
while true **do**
  Move calibration target
  if calibration target visible **then**
    calculate ${}^{c}[\mathbf{R}|t]_{w}$ from stereo calibration
    find X_{mb} from acoustic image
    calculate re-projection error e
    if $e < max\_proj_{err}$ **then**
      calculate a from eqn 4.5
      calculate b from eqn 4.5
      concatenate $\mathbf{A}$
      concatenate B
    end if
  end if
  if terminate **then**
    calculate ${}^{s}[\mathbf{R}|t]_{c}$ from eqns 4.10 and 4.11
    break
  end if
end while

extraction of internal (printed) corners of the checkboard through the detection of intersecting gradients. Unfortunately, the sonar views the physical extents of the checkboard, for which there are far fewer robust turnkey algorithms. Instead the physical corners of the checkerboard were identified in each image manually. This is both a) tedious and b) a source of error, as manual corner detection is typically less accurate than computer vision-based point identification methods, if an appropriate fiducial is present. To reduce error in the final calibration calculation, we estimated the reprojection error of the calibration target back to the origin based on the calculated transformation, and discarded transformations with high reprojection errors (in practice, we utilized a threshold of $0.2 - 0.3$m for the presented calibration target of size $27.6 \times 21.2\text{cm}^2$).

Mitigation Strategy: add a fiducial markers which can be used to extrapolate the physical corner location.

2. *Lack of automatic correspondence between the acoustic and optical frame.* While the calibration target emitted strong intensity in both the optical and acoustic frame, as shown in Fig. 4.2, the precise location of the checkerboard origin point (the lower left corner in Fig. 4.2(a)) is not clear. This correspondence becomes more challenging as the checkerboard moves through more unique poses, e.g. turned orthogonal to the sonar frame.

   Mitigation Strategy: Add an acoustically reflective marker to the origin point of the calibration target, to help with manual origin verification.

3. *High number of required corresponding points* While nine points are required for the mathematical description to be both solvable and unique, we typically found that accurate calibration results required

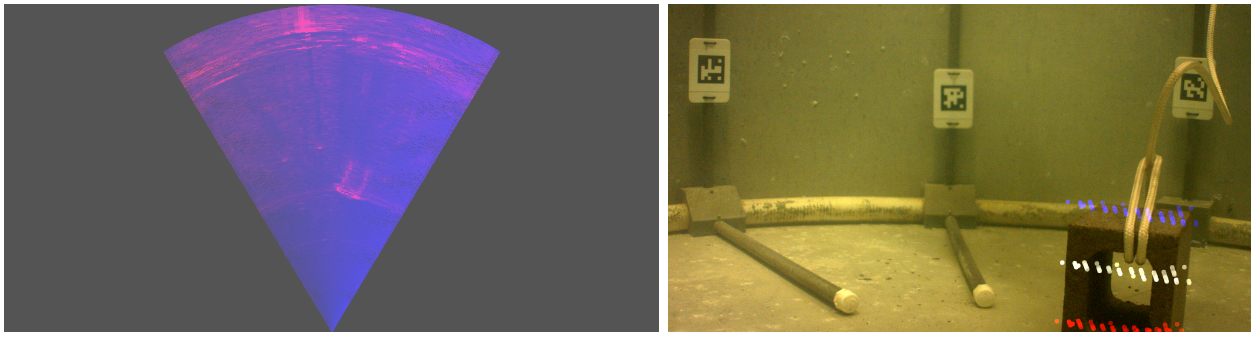


Figure 4.3: **Sonar projection 1**. (a) View of the cinder block target in the sonar frame. (b) Projected sonar points in the left camera image, with gray corresponding to the true calculated projection, and blue and red corresponding the upper and lower bounds, respectively, of the projection based on an 11 degree field of view.

many more points, at times more than thirty. As this process is not yet automaaed, the backtracking of this many calibration target positions is tedious.

From the above described calibration procedure, we were able to calculate an estimated rigid body transform between these two sensor frames, despite the challenges listed in section 4.3.1. The rotation between the two frames was close to the identity, as was expected given the mechanical configuration of the sensor head, with a small translation across all three axis. This estimated extrinsic was consistent with the expected result given the mechanical configuration of the sensor head, although given both the camera and sonar centers do not (necessarily) correspond to a physical point on the sensors themselves, precise mechanical validation is non-trivial.

For additional verification of the calibration results, we transformed sonar points into the left image frame using ${}^{s}[\mathbf{R}|t]_{c}$ and projected them onto the left image using calibrated left camera intrinsic matrix, $\mathbf{K}$. An example of this projection is shown in Fig 4.3(b), where projected sonar points are shown in white. Given that the sonar decimates across its vertical aperture, this projection can only show a planar 'slice' of the cinder block target. However, the upper and lower bounds based on a 11 degree vertical field of view are shown in blue and red, respectively, to illustrate a the effective bounds over which the sonar image is integrated. Furthermore, in this display, sonar points are only shown if they are less than 0.9m distance from the sonar (to avoid showing the projections of the back wall) and have a significantly high intensity value (to avoid displaying excessive noise and/or non-cinderblock targets).

A consistent problem with all of the data shown here is the relatively small overlap between the fields of view of the cameras and sonar at the short ranges ($\sim$ 1m) available in our test tank, with the sonar field of view occupying at most the lower half of the camera images. In an operational scenario featuring such short sensor-to-target distances, the sensors should be verged (tipped) towards each other to converge at an expected target range.

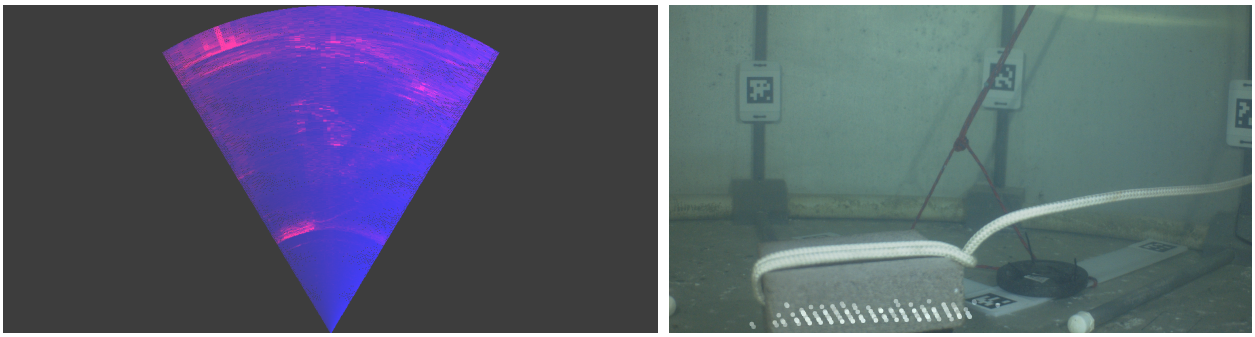


Figure 4.4: **Sonar projection 2**. This sonar projection shows the cinder block represented as a planar target in the sonar field of view, which is projected to the approximate correct location in the optical field of view.

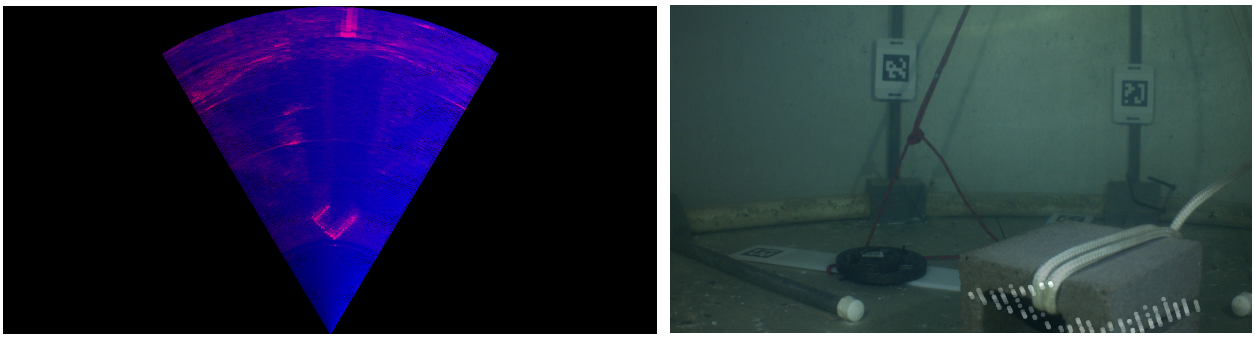


Figure 4.5: **Sonar projection 3**. In this sonar projection, three of the four cinder block sides are registered in the sonar image. This shape is then clearly projected to the optical cinder block image.

4.4 Calibration Results

From Fig 4.3(b), the projected sonar points appear to be projected to the approximate center of the cinder block image, particularly in the x direction. The y direction appears to be less centered, which is consistent with calibration expectations as accurate estimation of the y axis is challenging, given the axis decimation in the sonar frame. Further, as we can only reconcile a slice of the target, the y location of the projected sonar points will vary based on the number of strong vertical returns from the sonar.

Two additional projections are shown in Figs. 4.4 and 4.5, where projected sonar points are shown in gray. In all three of these provided registration examples, it is clear that the camera-sonar extrinsic calculation is approximately estimating camera-sonar transformation.

Chapter 5

3D Reconstruction

The ultimate goal of this project is the development of an algorithm which can process and fuse perception data collected from both optical and acoustic sensors in realtime. As discussed in the introduction, we start from the foundation of visual simultaneous localization and mapping (SLAM) with the goal of extending the core SLAM engine to opportunistically include both sonar and visual data as conditions permit. In the project, post-processed visual reconstruction is also employed as a tool for generating ground truth for scene structure and camera motion; while this approach may seem circular (using reconstruction to test reconstruction), there are a number of key differences between the two reconstruction approaches. Foremost, Metashape, the reconstruction tool used to estimate ground truth is not designed to run in a realtime manner, taking minutes to produce a sparse estimate of scene structure and hours to generate a detailed dense point cloud. Moreover, it focuses on batch rather than incremental processing, weighting all data equally to arrive at an optimal solution for world structure. Because of this, the cost of performing this post-processed reconstruction grows with the quantity of data available / length of mission. In contrast, SLAM algorithms focus on realtime operation and most are fundamentally incremental, designed to accept a continuous stream of new data to improve and extend an existing map. As an additional factor, as it relies on a commercial product, there was no practical path to extending the post-processed reconstruction pipeline to include sonar information.

5.1 Ground Truth Reconstruction through Post-Processed Photogrammetry

An essential element of SLAM-based reconstruction is that it estimates both the 3D structure of the scene *and* the trajectory of the camera relative to the scene from the data itself, and does not require precise navigation information as an input. In doing so, it inherently provides the necessary information to plan and execute interactions between the robot and the scene. Evaluating the performance of any algorithm, however, depends on collection of ground truth of the scene contents and camera motion. The former can be constrained by the use of objects of known geometry, or a direct structure sensor could be employed to a gather a reference model before data collection (given speed of data collection and update rate are not critical when collecting ground truth). The latter we constrain through the use of the camera gantry.

As a further source of structure information, we use Metashape, a photogrammetric reconstruction tool. Metashape is designed for the generation of 3D models from large sets of images but in contrast to the

algorithms described below it operates strictly in a post-processed mode. Given a set of images of sufficient quality, it produces a 3D model and camera pose estimate for each image. Naturally, this approach is only effective in low turbidity water.

As with any photogrammetric solution, the resulting model has no reference of absolute orientation, nor can Metashape set the overall scale of the model solely based on the input imagery. To address this issue, Metashape provides a ground truth / ground control point (GCP) toolkit for the labelling of known reference points in the model. If these points are then placed in a fixed coordinate system (e.g., assigned UTM northing/eastings for a landscape-scale model), the model can be transformed and scaled into a metric form.

Examples of Metashape output are provided in the data analysis section (Chapter 6).

5.2 Realtime Reconstruction with LSD-SLAM

The reconstruction algorithm selected is based on Large-Scale Direct SLAM (LSD-SLAM) as written by Dr. Jakob Engels at the Technical University of Munich and subsequently published in both monocular 2014 and stereo 2015 forms. Dr. Engels subsequently released the monocular version to Open Source under the GNU Public License version 3 (GPLv3). LSD-SLAM was selected as the basis for this research for its technical properties, as described below, as well as its relatively rational algorithmic structure (as described in Engel, Schöps, et al., 2014), which we believed made it amenable to extension.

Within the taxonomy of visual SLAM algorithms, LSD-SLAM can be classified based on three technical elements:

- LSD-SLAM processes visual data in a *direct semi-dense* manner. Unlike approaches which decimate incoming images into a reduced set of visual features (e.g., Klein and Murray, 2007; Mur-Artal et al., 2015), LSD-SLAM works directly with the intensity information from a large number of the pixels in each image. In this way, LSD-SLAM incorporates a greater fraction of the incoming image information, and natively tracks a large number of image points at any point in time, producing a depth estimate for those points without further processing.

  While LSD-SLAM initially considers the entirety of each image, it focuses on pixels lying near areas of strong image gradient, as these edges are both easier to track photometrically and often correspond to real physical edges in the scene. As each new image is passed to LSD-SLAM, it calculates a gradient (finite differences) image, then tracks only those pixels corresponding to regions of high gradients.

  The semi-dense approach does increase the computational cost of processing each new image. The open source LSD-SLAM implementation uses multi-threaded SIMD operations to accelerate core calculations.

- The fundamental data structure in LSD-SLAM is a *keyframe.* A keyframe corresponds to a single image pulled from the image stream, and stores an estimated depth for each high-gradient pixel in that image, as well as an estimate of the camera position for that keyframe relative to the first frame in the SLAM map.

  Newly arrived images update the depth map on the most recent keyframe through variable baseline stereo matching. When the distance from the previous keyframe reaches a threshold, LSD-SLAM promotes a new image to be the next keyframe, estimating the new keyframe's camera pose through

visual odometry, and initializing the depth map through reprojection of the prior keyframe's depth map. In this way, keyframes spatial downsample incoming data. Incoming imagery is integrated into the latest keyframe then discarded, leaving only a sparse set of keyframes based on the actual spatial coverage of the video, rather than the mission duration.

- These keyframes are stored within a *pose graph*, where the vertices in the graph are the keyframes, and the edges are the estimated change in pose between keyframes. These edges are initialized with visual odometry when new keyframes are added. Side-links within the graph are formed when two non-temporally-sequential keyframes are found to overlap, with the change in camera pose between those keyframes also estimated through visual odometry. LSD-SLAM searches for these side links continuously in a lower-priority thread, consuming spare cycles when the system is not otherwise performing live processing.

  At intervals, the vertex states within the pose graph (the keyframe poses) are optimized, subject to the odometric constraints stored in the graph edges.

Given these three key structures, LSD-SLAM operates as four semi-independent processes:

1. As new data (a new image) arrives, LSD-SLAM first pre-processes the imagery, forming an image pyramid and calculating the gradient image, thus isolating the pixels of interest. A *tracking* process then estimates the pose of the new image relative to the current keyframe through a dense matching of the high-gradient pixels. Tracking occurs immediately to provide a low-latency estimate of camera position relative to the existing LSD-SLAM model.

   If the pose change from the preceding keyframe is sufficiently large, the current frame is promoted to be the next keyframe. Its depth map is initialized by reprojection of the previous keyframe's depth map to the new keyframe's coordinate frame and the new keyframe is inserted into the pose graph with an odometry edge based on the tracking result.

2. If the new frame is not promoted to be a keyframe, it is passed to a *mapping* process which uses the epipolar constraint introduced by the visual odometry to calculate a stereo depth estimate for each pixel in the current keyframe's depth map. If successful, the pixel's depth and associated variance are updated based on that new depth estimate. Having completed this update, the new image frame is discarded.

3. In parallel with the above, the system is running two lower-priority processes which opportunistically fit into spare CPU cycles. The first performs photometric comparisons between keyframes, searching for opportunities to add cross-links (or loop closures) within the graph. The list of potential image pairings to evaluate is prioritized by the inter-vertex distance estimated by traversing the graph.

4. The second low-priority process optimizes the pose graph. The optimization process alternates between local optimization in a neighborhood around the current keyframe and global optimization of the full graph.

Starting from the existing Open Source implementation of LSD-SLAM, a number of substantial extensions were written to support this project. Foremost was the reintroduction of stereo visual processing to the LSD-SLAM codebase. While the original authors also built a stereo version of LSD-SLAM (Engel, Stückler, et al., 2015), its code was never published. Our implementation is relatively straightforward, with stereo depth

calculated independently using a standard dense stereo algorithm before the new images are sent to LSD-SLAM. This depth image is then treated as a supplemental input associated with the new image data within LSD-SLAM and used to update the current keyframe's depth map in a manner identical to the motion-parallax-derived depth estimate between the new image and the keyframe, which is also still calculated. The stereo-derived depth map is particularly useful during the early stages of reconstruction before there has been sufficient motion to derive an accurate depth from motion parallax, and before the depth map has converged, an important consideration for low-motion scenarios, as discussed below.

As described for the sensor data acquisition software in Chapter 3, during the course of the project we also transitioned to heavier use of the Robot Operating System (ROS) as a software development environment. To accommodate this, a ROS-compliant wrapper was written around our version of LSD-SLAM. This version is significantly simpler than the standalone version as it does not need to handle any sensor data ingestion and can instead utilize the standard ROS data subscription pattern, leaving sensor handling to a separate, standalone node. Similarly, in this version, the standard ROS stereo depth estimation node is used to produce the initial dense stereo depth map, which again is passed into LSD-SLAM as a ROS message.

Although we started with the Open Source version of LSD-SLAM from the original author, we made significant structural changes to the software, both to introduce the changes detailed above and to improve the overall quality of the codebase. As delivered we found the code to be poorly structured and exceptionally brittle, particularly in the synchronization of the major processing threads. Re-engineering and refactoring the code proved to be a time-consuming but necessary step which resulted in a more tractable codebase and also helped us to understand the technical details of the software. Our heavily modified LSD-SLAM codebase has been re-published as Open Source, maintaining the original authors' GPLv3 license.

Through our investigations, we gained further insights into the behavior of LSD-SLAM. In addition to the general issues with code quality, we discovered a number of properties related to the keyframe data structure which are less consequential when using LSD-SLAM to map large areas but which become a handicap when performing small-area mapping with limited range of motion, as in this case. First, while the keyframes are independent data structures, each storing an image and associated depth map, the depth maps themselves are subject to a long-term convergence with a settling time longer than the typical active period of a given keyframe. That is, on system startup, the first image is used to initialize the first keyframe, but in a monocular system its depth map is necessarily uninitialized. As the camera moves, the motion allows updates to the first keyframe's depth map, but there is typically insufficient parallax to fully converged its depthmap before it is supplanted by the second. The partially-converged depth map is propagated to the second keyframe, where it continues to improve and is propagated to future keyframes. Even in data sequences with significant motion, a converged, accurate depth map can take 5-10 keyframes and several minutes worth of data to emerge, simply due to the relationship between the average spatial lifespan of a keyframe and the diversity of viewpoints required to produce an accurate 3D reconstruction. If there is limited motion during the early phases of model creation this convergence is further delayed. This is remedied significantly through the use of stereo, as the initial depth estimates provided by stereo are available immediately from the first frame, regardless of camera motion.

The second is that while LSD-SLAM's output appears to be a global point cloud of the full scene, it is in fact a concatenation of the discrete point clouds belonging to each keyframe, and creation of a "global" point cloud (or a global point cloud subset spanning multiple keyframes) requires an independent step of combining (and potentially resampling) the many point clouds. In one sense this is an inherent feature of the system as the disjoint nature of the point cloud representation allows the underlying pose graph to be

updated trivially, as no expensive global point cloud update is required within normal processing. However, the efficiency gained by not carrying a global point cloud data structure is immediately lost when subsequent processing requires a unified point cloud, for example for object search or even for presentation to human observers.

A final criticism is that, while the depth maps are connected at the time of their initialization through propagation, there is no mechanism for updating past keyframes' depthmaps using information found in future keyframes, even in cases where the two keyframes have a high degree of overlap. This is of greatest consequence when combined with the long convergence period of the depth map noted previously. When mapping a large area, the relative inaccuracy of the first few keyframes is of little consequence, and in many cases those can be dropped (or not included in renderings of the point cloud) trivially. When viewing a relatively small object, where the total number of keyframes may be small, the inaccuracy of early depth maps becomes far more noticeable.

We propose two approaches to resolving this issue. First is to develop a mechanism for propagating and updating point cloud information between non-sequential keyframes where there is significant overlap. This could occur either as an independent point-cloud based search or as part of the process of identifying cross-links between keyframes (which uses image information, rather than depth information). To avoid spatial errors, this point cloud update should occur only after the inter-keyframe geometry within the pose graph has converged.

The second mechanism is to examine keyframe re-activation, the ability to return to previous keyframes when revisiting previously explored regions rather than continuously adding new keyframes. This approach is also beneficial for long-term observation of compact scenes as it leads to the number of keyframes being proportional to the spatial extent (accounting for changes in pose) rather than the length of the camera trajectory. However, it requires additional care in how the current depth map is propagated *back* to the re-activated keyframe, as it will contain some prior depth map information.

5.3 Paths to Integration of Sonar Data

As detailed in the introduction, this program architecturally starts with visual SLAM and then integrates sonar as a complementary data source. This development path was originally selected based on the suitability of visual SLAM for scenarios where water clarity allows some degree of visual reconstruction, even if it is unevenly available in time or space. A well-designed algorithm could then opportunistically include sonar data as required when optical data became unavailable.

Due to time constraints, the integration of acoustic data into LSD-SLAM was not addressed. The proposed method would implement a space-carving method similar to that describe in Guerneve et al. (2018). In this method, the current LSD-SLAM keyframe structure would be supplemented with an voxel occupancy grid which corresponds with the field of view of the sonar. Given the sonar-to-camera extrinsics from Chapter 4, each depth map point within the keyframe could be associated with a cell within the occupancy grid. Each cell would store a probability of occupancy based on the strength of the associated sonar return. As the sensors move, data from subsequent sonar images is reprojected into the current keyframe and each cell's occupancy estimate updated probabilistically. A subsequent integration step would be required to fuse the sonar-based occupancy data with the vision-based depth map.

Chapter 6

Data Collection and Analysis

Test data was collected during two test sessions. In November 2018, the first data was collected coincident with a test for SERDP project MR-2734 in the University of Washington School of Oceanography salt water test tank. The second session spanned August-September 2019 and made use of a freshwater acoustic test tank at APL. Despite the relatively lightweight nature of the test apparatus, the logistics required to deploy to the Oceanography test tank underlined the challenges inherent in collecting data in the water, even in highly controlled circumstances.

6.1 UW School of Oceanography Test Tank, Nov 2018

This data collection was scheduled to piggyback on planned testing for MR-2734. For that project, a 8ft x 8ft x 12ft galvanized steel "test frame" was constructed to hold a Schilling Titan 4 hydraulic ROV manipulator along with a $\sim$ 5 ft by 5 ft plastic storage container containing washed gravel as a simulated seafloor. Though not a perfect stand-in for sandy or silty seafloor, the larger particulates were far more acceptable in the shared test tank than a finer sand.

The gantry was mounted on the top of the sensor frame such that the linear rail allowed motion towards / away from the arm. This allowed the radial arm to be pulled out of reach of the arm during MR-2734 testing, then moved over the target box while gathering data for this project.

In total, six minutes of data were collected in three combined stereo/sonar data files at a rate of approx 61MB/sec. In each run, the sensor head was manually orbited around the UXO in the gravel-filled testbed, with sample stereo imagery shown in Figure 6.2 and sonar imagery in Figure 6.3. In addition to the UXO, the scene included a number of visual fiducials as well as sets of brightly colored steel spheres. The goal of the latter was to provide an automatically identifiable target in both optical and acoustic modalities for subsequent registration of the data and potentially calibration. In practice the spherical targets proved to be difficult to identify in the sonar data.

Subsequent analysis of data tracks revealed a number of flaws in the data collection process. Most critically, the timing information for data frames within the video/sonar data files was improperly defined such that the sonar frames' timestamp reflected the time at which data was committed to the video file, not the time when it was collected. Due to a related bug, the sonar data was prioritized after the video frames and were written as a block at the end of the recording session. Thus, precise information on the timing of the sonar data relative to the video was lost. An estimated timing could be recovered by evenly distributing

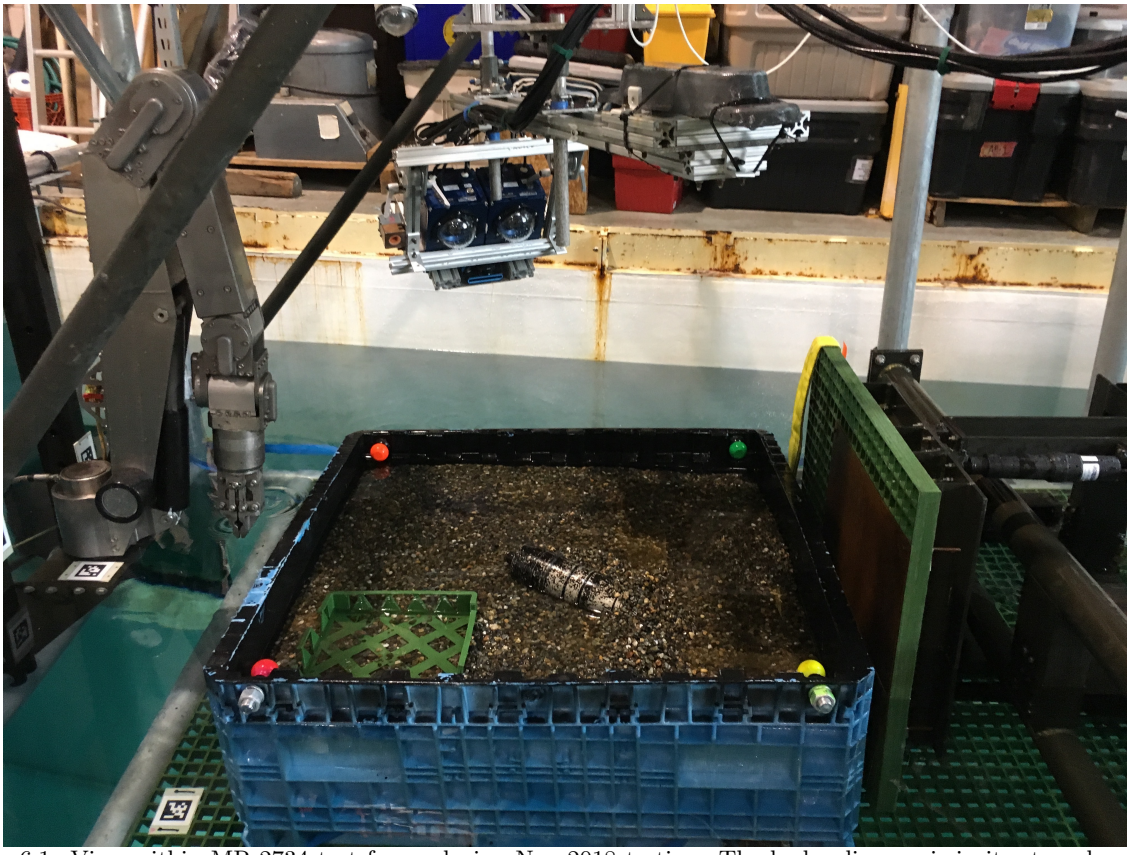


Figure 6.1: View within MR-2734 test frame during Nov 2018 testing. The hydraulic arm is in its stowed position to left. The sensor gantry has been translated such that it is centered over the target area and the linear trolley locked in place. The sensor head (at center) can now move radially around the target area, which contains a gravel substrate, and a UXO-stand-in. The four brightly colored fiducial spheres can be seen in the corners of the plastic container.

Figure 6.2: Stereo image pair from UW School of Oceanography test tank showing the simulated UXO in situ in the testbed.

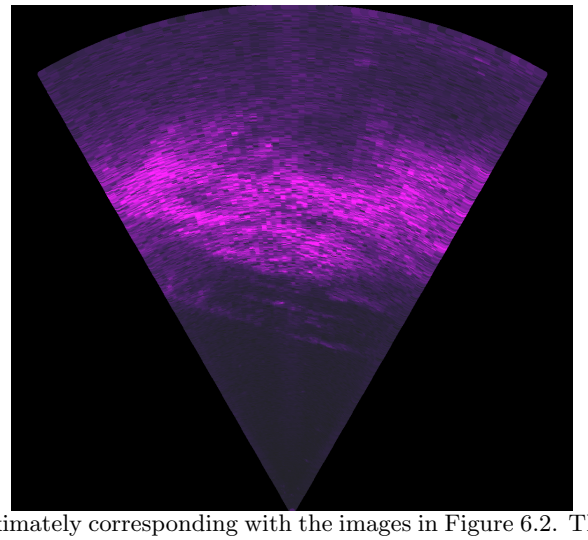


Figure 6.3: Sonar image approximately corresponding with the images in Figure 6.2. The edge of the seabed container closest to the sonar is visible in the middle of the image. As per text, data synchronization was not maintained in these data files, making accurate video-to-sonar data synchronization impossible.

sonar frames throughout the time extent of the video, but these timestamps would be approximate at best.

Despite the problems with the correlation of the acoustic data, the visual data was suitable for further processing. As described previously, to establish a ground truth on the scene structure and camera trajectories, the COTS photogrammetric package Metashape was used to produce a 3D reconstruction from a subset of the visual data. Based on the rate of camera motion, one frame was extracted per second, with redundant frames from periods of no camera motion removed. This processing was completed for each camera independently, as well as for a combined left-right image set, noting that the photogrammetry software does not take advantage of the fixed geometry of the stereo pair. The results from this reconstruction were quite successful (as per Figure 6.4), offering a detailed, 3D model of the UXO in situ.

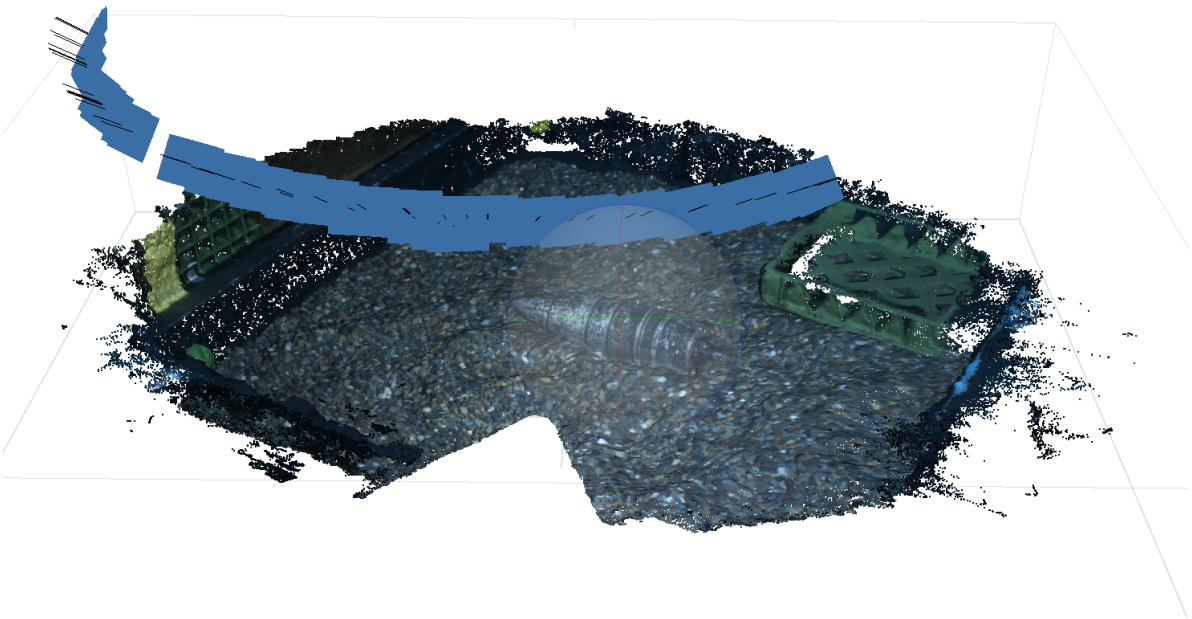


Figure 6.4: Ground truth 3D reconstruction of test UXO. Blue squares correspond to camera positions for each image (103 each per left and right cameras) used to produce the reconstruction.

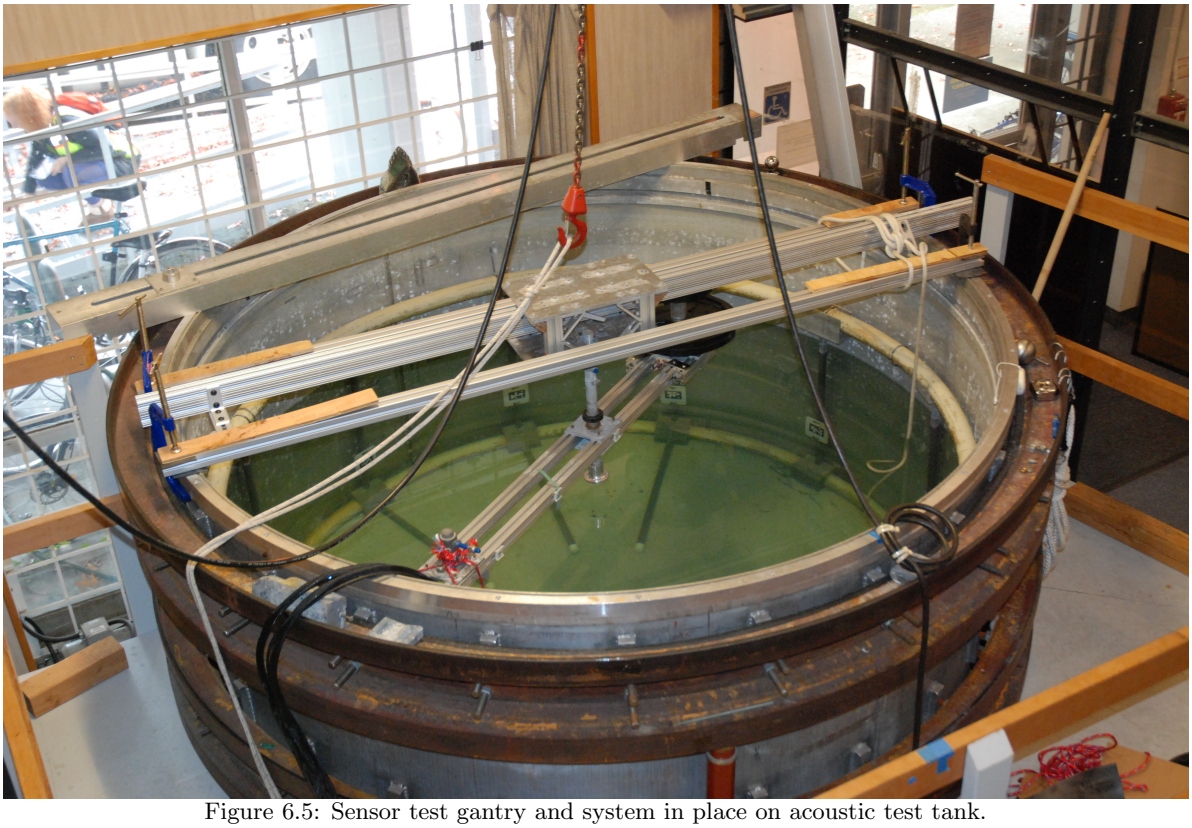


Figure 6.5: Sensor test gantry and system in place on acoustic test tank.

6.2 UW-APL Acoustic Test Tank, Aug-Sept 2019

Following the November 2018 tests, the system was stood down to allow for data analysis and investigation of issues discovered during initial testing and to make mechanical and software improvements. The system was subsequently re-deployed to a $\sim$ 2m diameter fresh water acoustic test tank at UW-APL (Figure 6.5). While more space constrained, this tank was available for long blocks of time such that the sensor system could remain in place allowing efficient data collection and analysis. A further limitation of this test tank was that an artificial seafloor could not be used in the tank, and a simpler non-UXO object was used for preliminary data collection. Sample imagery from this testing is shown in Figure 6.6, with sample sonar imagery in Figure 6.7. Ultimately, the ready access to the test tank allowed much needed iteration through testing scenarios and refinement of procedures.

As before, the first step in data processing was to generate ground truth geometry and camera motion using Metashape (Figure 6.8). This process was repeated for each data set to establish a baseline trajectory for both cameras. Within Metashape, each camera (left / right) was processed independently, as well as a merged left+right data set. As discussed previously, Metashape does not make use of the known stereo geometry between the two cameras and instead treats the data from each camera as a collection of fully independent images.

This imagery shows a number of further process improvements. An L-shaped fiducial marker of known size and geometry was placed on the tank floor to provide a ground control points for subsequent reconstructions. This fiducial is spatially more compact than the colored spheres used in the previous test and offers more consistent control points in any given image. Within Metashape the center of each fiducial marking could be manually identified as a ground control point using the tools provided by the software. These ground

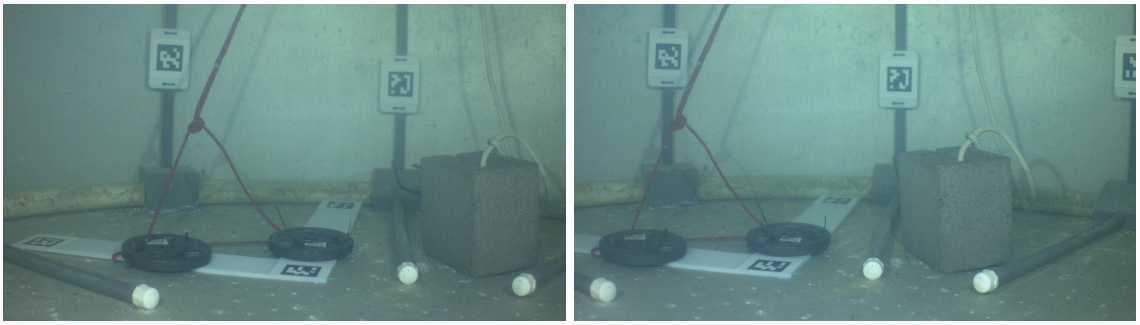


Figure 6.6: Stereo image pair from APL test tank showing a sample object (a cinderblock) as well as fiducial used for establishing scale and coordinate frame in ground truth reconstructions.

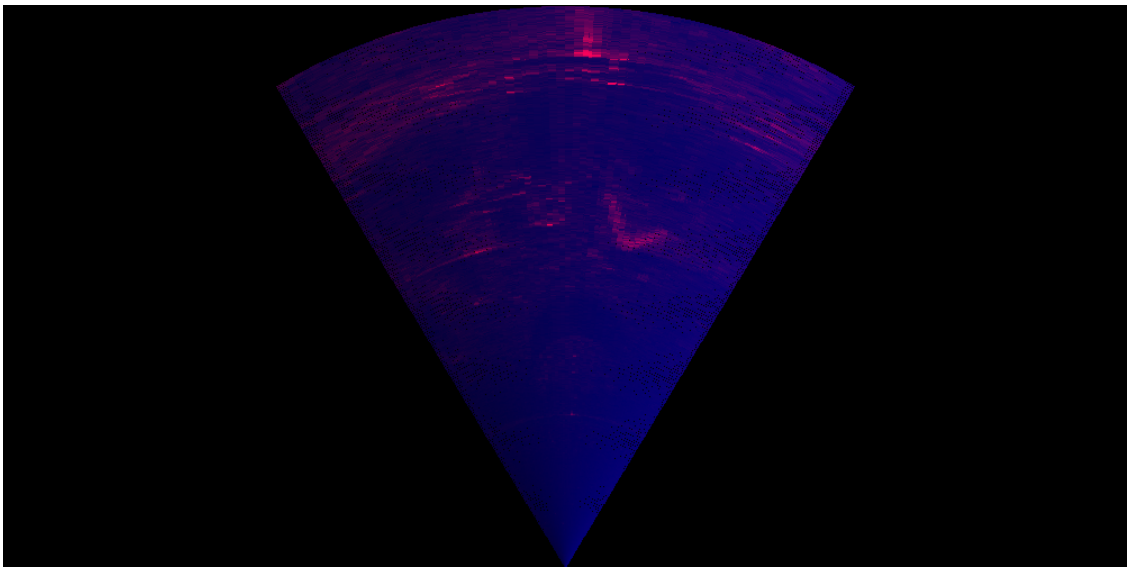


Figure 6.7: Sonar imagery from APL test tank corresponding to Figure 6.6. The corner of the cinderblock as well as strong reflections from the weights on the L-shaped fiducial marker are visible, as well as a reflection from either the pipe manifold or fiducial marker on the back wall of the test tank.

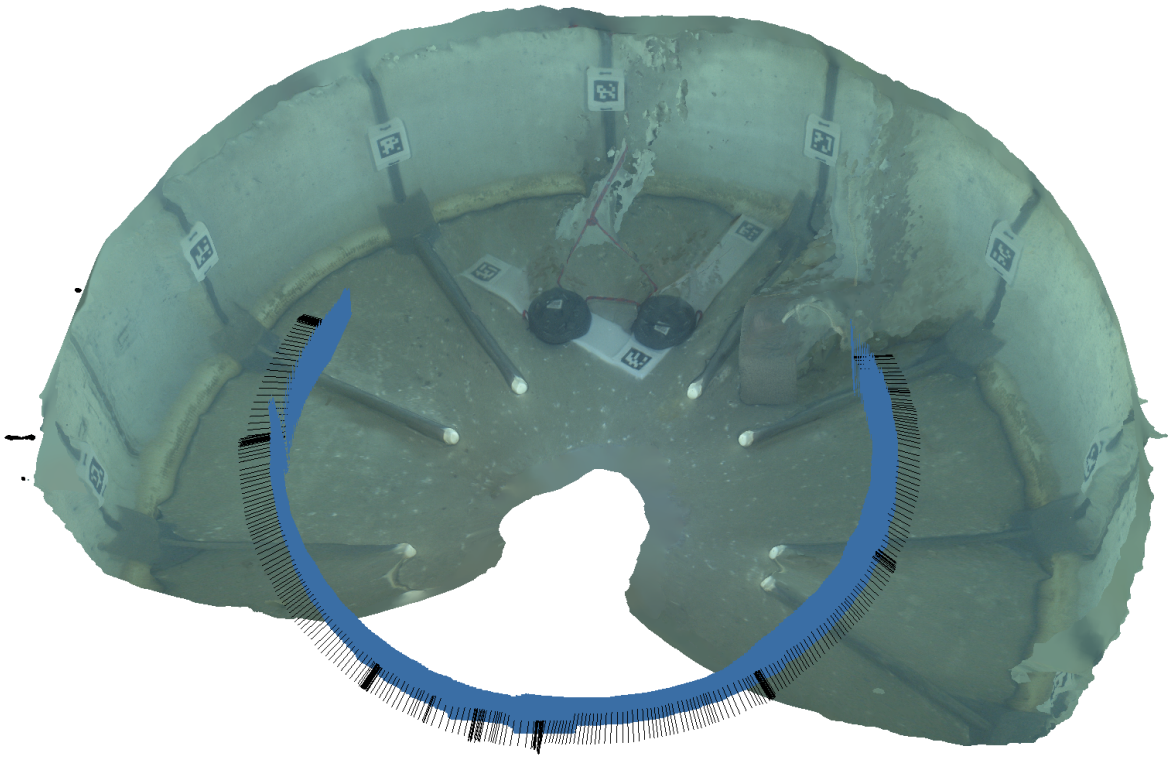


Figure 6.8: 3D Reconstruction of left camera images (as seen in Figure 6.6 from Metashape showing cinderblock as well as L-shaped fiducial object. Note failure of reconstruction near the taglines on both the fiducial and cinderblock which may be due to the relatively low texture on the tank wall near the taglines, or may be due to motion in the lines during data collection.

control point can then be associated with a known coordinate frame (we define the two legs of the marker as the world $X-$ and $Y-$ axes). This also fixes the model scale, which is otherwise unobservable from imagery.

The collected data was then processed with LSD-SLAM using the camera intrinsic and extrinsic calibrations estimated previously. As shown in Figure 6.9, the stereo-video-based reconstruction algorithm was able to build a model of the test tank in realtime, correctly capturing the overall shape of the tank as well as the fiducial marker and cinderblock. As LSD-SLAM concentrated on image portions with high gradients, the reconstruction is denser on object boundaries than in untextured areas, hence the missing reconstruction for the tank walls and floor. Similarly, the "fuzziness" in the reconstruction is due to the convergence effects of LSD-SLAM discussed previously.

6.2.1 Trajectory Analysis

For the data set shown in Figures 6.8 and 6.9, the camera trajectories (attitude and position) were exported from both LSD-SLAM and Metashape. Both trajectories are indexed to by frame number in the original input video, allowing time alignment of the two trajectories. Further, the Metashape-derived trajectory can be considered *metric* as the L-shaped fiducial had been located within the model as set of ground control points, setting an absolute coordinate origin and defining the model scale. This alignment *could* be done in the LSD-SLAM model but at present detection of the fiducial has not been implemented. Absolute attitude could also be determined by an external input e.g., an IMU to measure the local gravity vector. As our LSD-SLAM implementation uses stereo video, it should calculate world scale inherently.

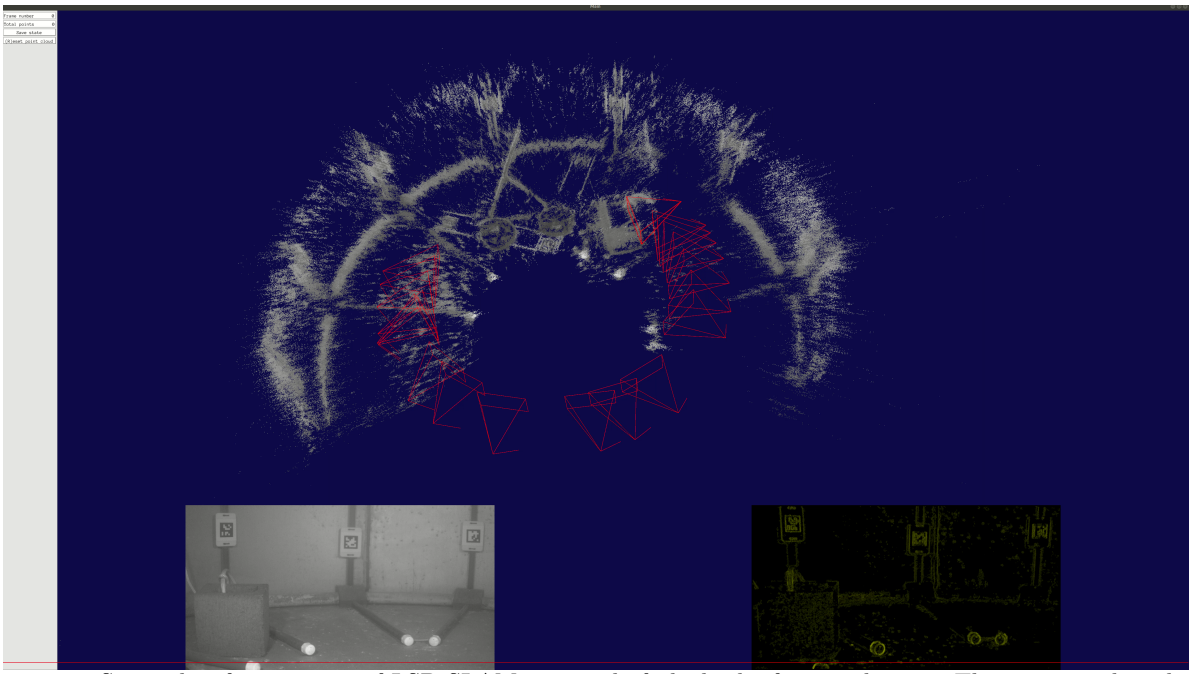


Figure 6.9: Screenshot from output of LSD-SLAM near end of playback of a test dataset. The main window shows the resulting point cloud (white) and the position of each of the keyframes (red). The lower left window shows the current frame of video input, while the lower right shows the associated gradient image.

To allow comparison between the Metashape and LSD-SLAM derived trajectories, the LSD-SLAM trajectory is subject to a rigid body transformation which aligns its first camera position with the corresponding image in the Metashape trajectory. The X-Y projection of the two trajectories is shown in Figure 6.10. The two trajectories show very similar shape but a clear difference in scaling. Given the known radius of the camera gantry of 0.75m, the Metashape trajectory is approximately accurate, while the LSD-SLAM trajectory is undersized, despite use of the stereo imagery. The source of this bias has not been determined.

To allow further analysis, the LSD-SLAM output was assumed to be subject to a single fixed scaling error, due to e.g., a software bug. A scaling factor of 1.145 was found to minimize the 3D RMS position error between each LSD-SLAM point and its corresponding Metashape point. This approach is understood to not be an accurate representation of the realworld performance of LSD-SLAM, but allows constructive comparison in the short term, particularly on the assumption that scale estimation from stereo vision is relatively reliable and robust, and the resulting the scale error is due to a bug in LSD-SLAM.

Having applied this scale correction, the X-Y projection of the two trajectories, as well as a 0.75 m-radius circle of the known gantry geometry, are shown in Figure 6.11. The X-, Y- and Z-errors as a function of frame number (effectively time) are shown in Figure 6.13, and the camera roll, pitch and yaw are shown in Figure 6.14.

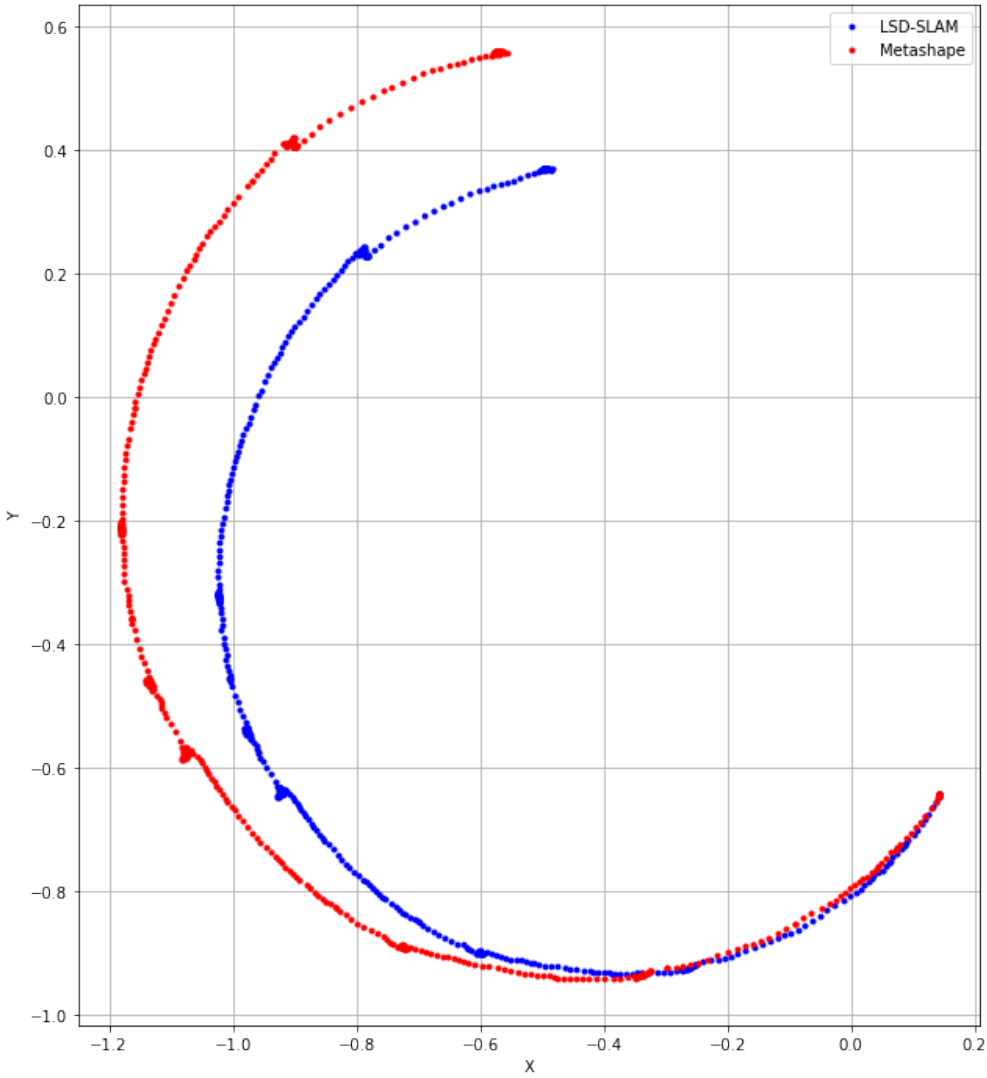


Figure 6.10: X-Y projection of Metashape and LSD-SLAM derived trajectories after aligning first LSD-SLAM camera position with corresponding camera in Metashape track. The world origin is defined by the center of the "corner" fiducial marker in the L-shaped ground truth fiducial object.

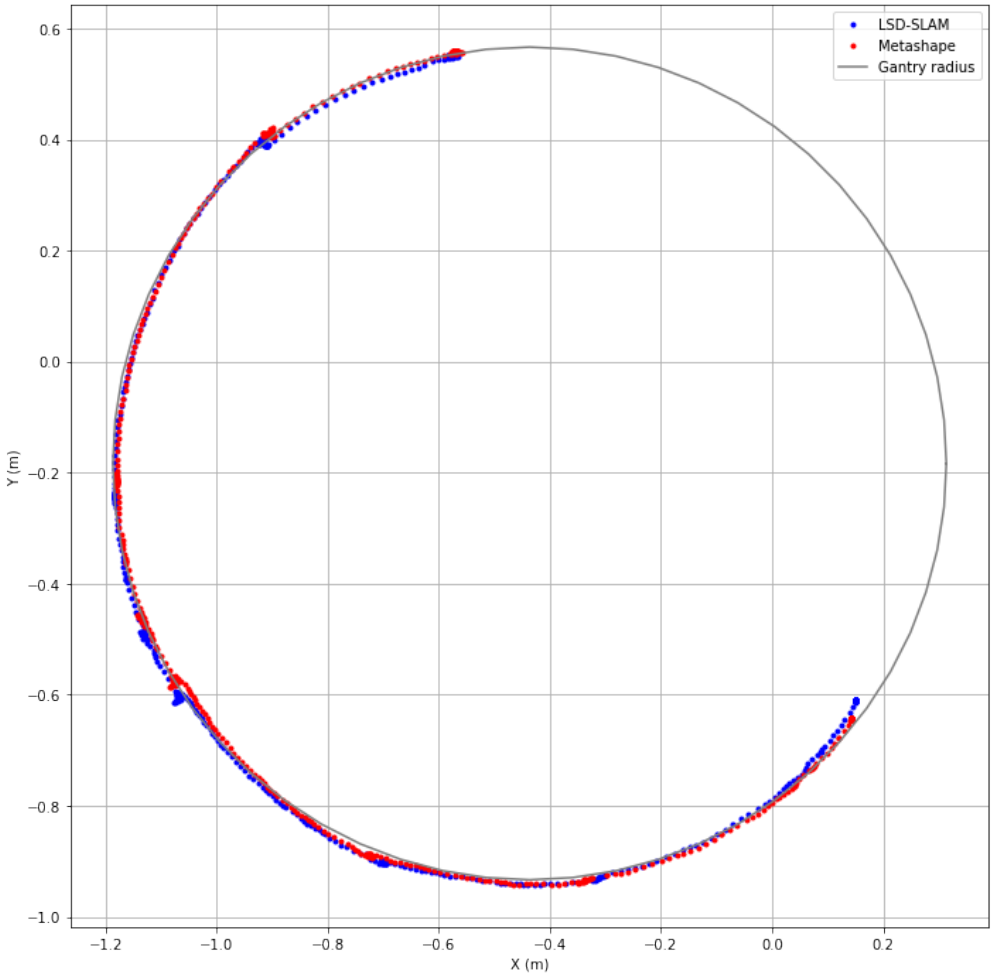


Figure 6.11: X-Y projection of Metashape and LSD-SLAM trajectories after aligning the first LSD-SLAM camera position with corresponding camera in Metashape track and applying the calculated scale correction.

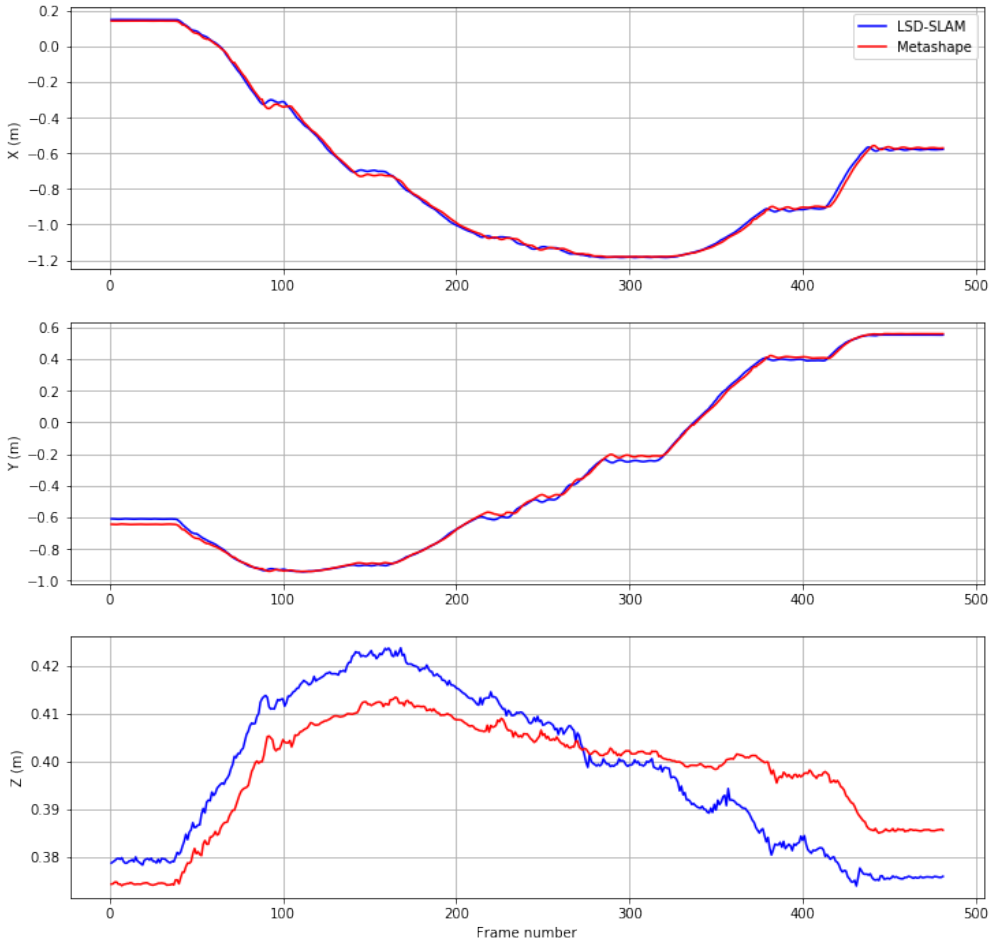


Figure 6.12: X, Y and Z components for Metashape and scale-corrected LSD-SLAM trajectories.

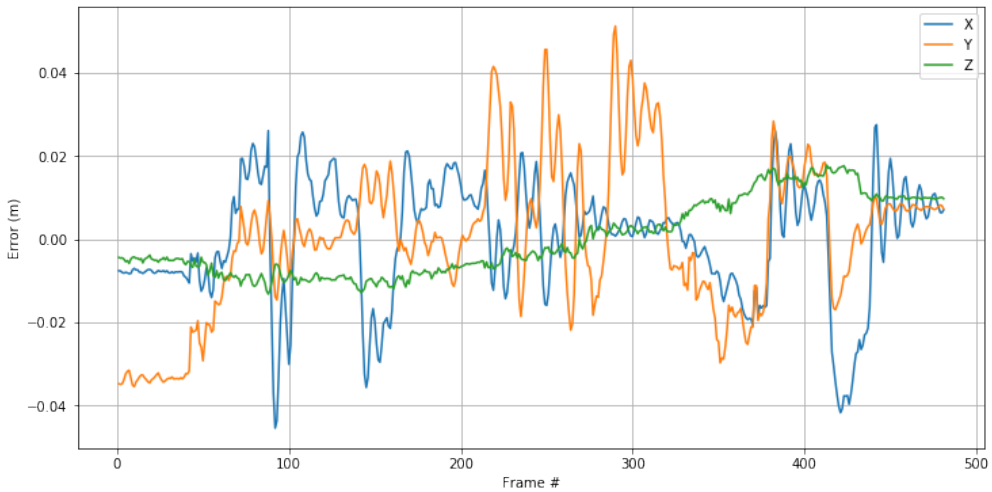


Figure 6.13: X, Y and Z errors between Metashape and scale-corrected LSD-SLAM trajectories.

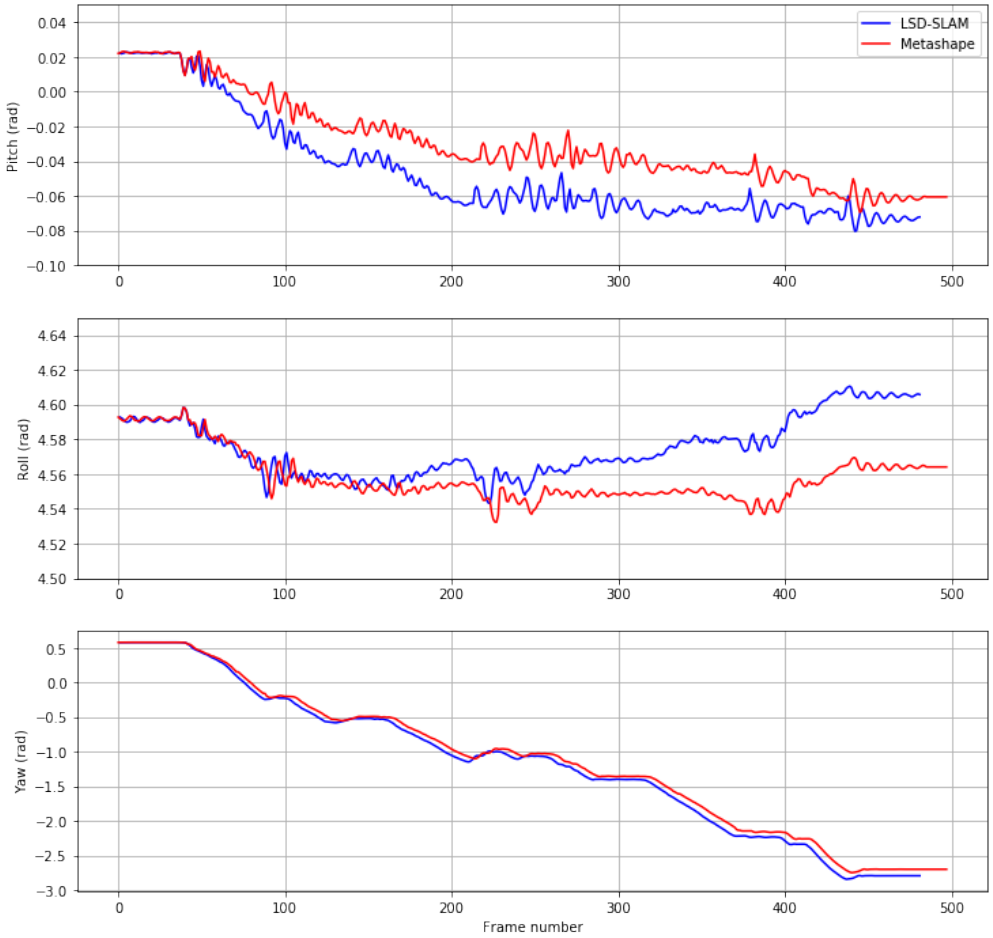


Figure 6.14: Roll, pitch, and yaw of Metashape and scale-corrected LSD-SLAM trajectories.

Chapter 7

Conclusions and Further Steps

Overall, this project made significant progress towards the collection and analysis of simultaneous optical and acoustic data of objects on the seafloor, including construction of a novel testbed for collecting consistent test datasets and independently estimating ground truth scene structure and sensor head motion. As shown, stereo visual construction was performed using data collected by the system, although this required far greater development than initially anticipated due to the need to explicitly re-introduce stereo processing to the selected SLAM algorithm, and to improve SLAM system performance in low-motion underwater images as found in this test scenario, as well as low quality in the original software implementation. Furthermore, we demonstrated a successful, if unoptimized approach to camera-sonar calibration allowing mapping of sonar data into the video frame. Ultimately, there was insufficient project scope to achieve the final step of leveraging acoustic data within the visual reconstruction framework.

Despite this, there is significant value in the ability to capture time-synchronized data from both optical and acoustic sensors in a controlled setting which will directly benefit a range of problems related to fine-scale sensing and reconstruction in underwater environments. At a trivial level, the data sets collected within the project, and similar data sets can be immediately applied to address monocular- and stereo visual, sonar-only, and fused vision-sonar reconstruction. Moreover, while we did not structure our scenario to require realtime position information during reconstruction, the ability to develop an accurate sensor track in post-processing allows synthesized ROV positioning information during playback. This in turn enables testing of algorithms (in a playback mode) which rely on external positioning information for reconstruction (as in e.g., Guerneve et al., 2018). This is particularly powerful as positioning on ROVs typically relies on a combination of an IMU and acoustic sensors which are both complex and difficult to emulate in a constrained tank environment.

Similarly, the capacity to efficiently generate testing data is particularly powerful given the rise of machine-learning based data processing approaches. Such supervised training algorithms require large sets of labelled data for training. When the cost of collecting sample data is high (e.g., with underwater video), the paucity of exemplar data handicaps algorithm development. As an example, data captured during this project has been used to test a proof of concept in pre-training object recognition algorithms before operating in novel environments. For example, given an object, data captured in the test tank can be used to train a network to detect that object in those particular conditions. How well does that network function when searching for the object in open water, potentially with different levels of particulates, ambient lighting, etc? Can the process be improved either through artificial augmentation of the training data (simulating different underwater imaging conditions) or through in situ recalibration of the object detector, recognizing and adapting to the

differences between the training circumstances (test tank) and the operating conditions (open water) without affecting the underlying object detection system?

As such, there are multiple steps forward which leverage the existing investment in testing infrastructure and expertise. Continued refinement of the optical and opti-acoustic reconstruction techniques investigated in this project is the clearest avenue for continued work. In addition to the core problem of integrating sonar data into the optical SLAM framework, a closely related problem is quantifying the degradation of the optical reconstruction as water clarity declines. Not only is this important for defining the boundaries of usability for optical reconstruction underwater, it also addresses the problem of determining when optical reconstruction is likely to fail (based on input data), or has failed (based on output data), which in turn is critical to any algorithm which opportunistically switches between the modalities based on the relative strengths or qualities of the data at any point in time.

An alternative is to redouble efforts to support autonomous manipulation solely with acoustics. While clearly more challenging, any such approach would naturally be applicable to a broader range of turbidities. A natural extension would continue to investigate approaches to reconstruction from imaging sonars. While there have been a number of successes in this space in recent years, strong caveats remain about the use of many such algorithms, including the availability of precise positioning information, or travelling in particular trajectories or sensor motions. As previously outlined, the existing infrastructure allows straightforward collection of large quantities of imaging sonar data which can be correlated to sensor trajectories using video.

Bibliography

Bruno, F et al. (2011). "Experimentation of structured light and stereo vision for underwater 3D reconstruction". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.4, pp. 508–518.

Choi, Hyun-Taek et al. (2015). "New concepts for smart ROV to increase efficiency and productivity". In: *2015 IEEE Underwater Technology (UT)*. IEEE, pp. 1–4.

Davis, Anthony and Angus Lugsdin (2005). "High speed underwater inspection for port and harbour security using Coda Echoscope 3D sonar". In: *Proceedings of OCEANS 2005 MTS/IEEE*. IEEE, pp. 2006–2011.

Engel, Jakob, Thomas Schöps, and Daniel Cremers (2014). "LSD-SLAM: Large-scale direct monocular SLAM". In: *European conference on computer vision*. Springer, pp. 834–849.

Engel, Jakob, Jörg Stückler, and Daniel Cremers (2015). "Large-scale direct SLAM with stereo cameras". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1935–1942.

Guerneve, Thomas, Kartic Subr, and Yvan Petillot (2018). "Three-dimensional reconstruction of underwater objects using wide-aperture imaging SONAR". In: *Journal of Field Robotics* 35.6, pp. 890–905.

Hildebrandt, Marc et al. (2008). "A practical underwater 3D-Laserscanner". In: *OCEANS 2008*. IEEE, pp. 1–5.

Hogue, Andrew, Andrew German, and Michael Jenkin (2007). "Underwater environment reconstruction using stereo and inertial data". In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, pp. 2372–2377.

Hurtós, N., X. Cufí, and J. Salvi (2010). "Calibration of optical camera coupled to acoustic multibeam for underwater 3D scene reconstruction". In: pp. 1–7. DOI: 10.1109/OCEANSSYD.2010.5603907.

Inglis, Gabrielle et al. (2012). "A pipeline for structured light bathymetric mapping". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 4425–4432.

Johnson-Roberson, Matthew et al. (2010). "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys". In: *Journal of Field Robotics* 27.1, pp. 21–51.

Klein, Georg and David Murray (2007). "Parallel tracking and mapping for small AR workspaces". In: *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, pp. 1–10.

Lagudi, Antonio et al. (2016). "An alignment method for the integration of underwater 3D data captured by a stereovision system and an acoustic camera". In: *Sensors* 16.4, p. 536.

Massot-Campos, Miquel and Gabriel Oliver-Codina (2015). "Optical sensors and methods for underwater 3D reconstruction". In: *Sensors* 15.12, pp. 31525–31557.

McGlamery, BL (1975). "Computer analysis and simulation of underwater camera system performance". In: *Technical Report; Visibility Laboratory, University of California, San Diego and Scripps Insitution of Oceanography*. 75, p. 2.

Mur-Artal, Raul, Jose Maria Martinez Montiel, and Juan D Tardos (2015). "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE transactions on robotics* 31.5, pp. 1147–1163.

Negahdaripour, Shahriar and Pezhman Firoozfam (2006). "An ROV stereovision system for ship-hull inspection". In: *IEEE Journal of oceanic engineering* 31.3, pp. 551–564.

Oleari, Fabio et al. (2015). "An underwater stereo vision system: from design to deployment and dataset acquisition". In: *OCEANS 2015-Genova*. IEEE, pp. 1–6.

Rydén, Fredrik, Andrew Stewart, and Howard Jay Chizeck (2013). "Advanced telerobotic underwater manipulation using virtual fixtures and haptic rendering". In: *Proceedings of 2013 OCEANS-San Diego*. IEEE, pp. 1–8.

Taylor, Russell H et al. (2016). "Medical robotics and computer-integrated surgery". In: *Springer handbook of robotics*. Springer, pp. 1657–1684.

Tetlow, Stephen and John Spours (1999). "Three-dimensional measurement of underwater work sites using structured laser light". In: *Measurement Science and Technology* 10.12, p. 1162.

Zhang, Zhengyou (1999). "Flexible camera calibration by viewing a plane from unknown orientations". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 1. Ieee, pp. 666–673.

Appendix A

Software Packages

All software packages developed within this program are written in C++. Software created as part of this project are released as Open Source under the MIT license, with the exception of the packages derived from the original authors' version of LSD-SLAM which inherit the GPLv3 license. This section describes the major software components written specifically for this project, although many of these packages are dependent on other package shared amongst different UW-APL programs; and some are dependent on libraries or APIs provided by hardware vendors.

All packages are posted online at the *github.com* software repository hosting site.

libbmsdiprotocol

`https://github.com/apl-ocean-engineering/libbmsdi-protocol`

Provides tools for generating packets compliant with the Blackmagic Designs SDI Camera Control Protocol. This package is written with minimal external dependencies (i.e., it is not dependent on the Blackmagic DecklinkAPI) to allow reuse in a variety of projects.

liblackmagic

`https://github.com/apl-ocean-engineering/libblackmagic`

Interface with Blackmagic Design Decklink SDI capture card and DecklinkAPI software, which must be downloaded separately from Blackmagic. Contains tools for capturing mono and stereo streams, and for sending Blackmagic Design SDI Camera Control Protocol messages to the cameras.

liboculus

`https://github.com/apl-ocean-engineering/liboculus`

Interfaces for reading and controlling the Oculus sonar, as well as C++ classes for unpacking and parsing Oculus sonar messages. Based on sample code provided by Blueprint Subsea.

libvideoencoder

`https://github.com/apl-ocean-engineering/libvideoencoder`

Code and interfaces for capturing multiple video streams and non-video (sonar) data in a video file container. Uses the open source *ffmpeg* software library.

serdp_recorder

`https://github.com/apl-ocean-engineering/serdp\_recorder`

Ties all packages together, providing a standalone application which records stereo video and sonar in a multi-track video file format.

serdp_player

`https://github.com/apl-ocean-engineering/serdp\_player`

Code to open the multi-track video file format and extract the video/sonar data while maintaining timing information.

serdp_common

`https://github.com/apl-ocean-engineering/serdp\_common`

Common code shared by *serdp_recorder* and *serdp_player*.

lsd-slam

`https://github.com/apl-ocean-engineering/lsd-slam`

Primary LSD-SLAM development repository, forked from the original author's code.

lsd-slam-pangolin-gui

`https://github.com/apl-ocean-engineering/lsd-slam-pangolin-gui`

Wrapper around the LSD-SLAM library which provides a graphical display using the `Pangolin` GUI toolkit, as seen in Figure 6.9.