Towards Acoustic Structure from Motion for Imaging Sonar

Tiffany A. Huang and Michael Kaess

***Abstract*— We present a novel approach, entitled acoustic structure from motion (ASFM), for recovering 3D scene structure from multiple 2D sonar images, while at the same time localizing the sonar. Imaging sonar or forward looking sonar (FLS) is commonly used for autonomous underwater vehicle (AUV) navigation. An FLS provides bearing and range information to a target, but the elevation of the target is unknown within the sensor's field of view. Hence, current state-of-the-art techniques commonly make a flat surface (ground) assumption so that the FLS data can be used for navigation. Unlike other methods, our solution does not require a flat surface assumption and is capable of utilizing information from many frames, as opposed to pairwise methods that can only gather information from two frames at once. ASFM is inspired by structure from motion (SFM), the problem of recovering 3D structure from multiple camera images, while also recovering the position and orientation from which the images were taken. In this paper, we formulate and evaluate the optimization of several AUV sensor readings of the same scene from different poses, the sonar equivalent of bundle adjustment. We evaluate our approach on both simulated data and FLS sonar data with the assumption that feature extraction and data association have been completed. The acoustic equivalents of those two important features of SFM are left for future work.**

I. Introduction

In recent years, society has observed an increasing need for autonomous vehicles that can operate underwater. Autonomous underwater vehicles (AUVs) have many applications in performing tedious and potentially dangerous tasks such as monitoring marine structures, inspecting ship hulls, and exploring deep ocean depths.

However, much of the current technology developed for AUVs does not yet match the vision for a vehicle that can navigate, inspect, and explore all on its own. One area that requires further development is long-term autonomy; current localization methods suffer from unbounded drift, making it increasingly difficult to precisely determine the whereabouts of the vehicle as the duration of the mission increases. A second significant hurdle for full autonomy is perception underwater. Beneath the water, optical cameras are only of limited use due to water turbidity. Without optical cameras, sonar becomes the sensor of choice.

Towards real-time autonomous navigation and creating a faster and more accurate 3D map with sonar, we introduce the concept of acoustic structure from motion (ASFM), using multiple, general sonar viewpoints of the same scene to reconstruct the 3D structure of select point features while minimizing the effects of accumulating error (Fig. 1). In this paper, we formulate much of the theoretical basis of the approach and focus on its integration with measurements received from other onboard sensors. We assume for now that feature extraction and data association can be sequentially and incrementally completed, and we explore the feasibility of using this information to localize the AUV and map its surroundings. In essence, this paper explores the acoustic equivalent of bundle adjustment [14], the geometric optimization in traditional structure from motion for cameras. The nontrivial tasks of automatic feature extraction and data association for sonar are topics of ongoing work in our group.

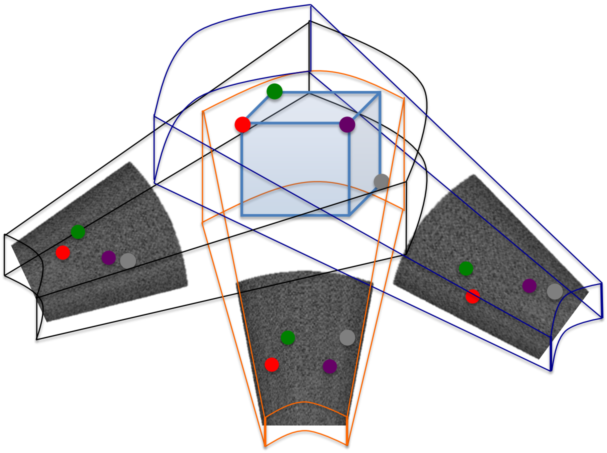


Fig. 1: Multiple imaging sonar views of a scene allow recovery of 3D position of point features, even though the individual views do not provide elevation information about the features.

ASFM has applications in real-time navigation for AUVs in general 3D environments. The simultaneous localization and mapping (SLAM) capabilities provided by our method can significantly reduce drift for long-term operation and facilitate mapping and inspection tasks. Unlike previous approaches, our solution does not make any assumptions about the planarity of the environment in order to localize the sonar. Additionally, ASFM is able to use information gathered from multiple sonar images to better constrain the 3D geometry of the scene and the motion of the vehicle. Much of the current work regarding 3D reconstruction from sonar uses a pairwise approach that cannot exploit constraints obtained from multiple viewpoints for a more accurate solution.

In this paper, forward-looking sonar (FLS) is used, but ASFM is not limited to this type of sonar. FLS is an obvious choice because it is one of the less expensive types of sonar and its larger field of view allows for faster imaging of

This work was partially supported by ONR grants N00014-14-1-0373 and N00014-12-1-0093.

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {tahuang,kaess}@cmu.edu

an environment. However, as an interesting topic for future work, ASFM can be extended to other types of 2D sonar including side-scan sonar [5]. Currently, beam-steering 3D forward-looking sonar sensors are available (e.g. Blueview 3DFLS), but they are both more expensive and slower to image a given volume (because of the low speed of sound in water), requiring up to 4 seconds for a single sweep at a short 6 m range, and more time for larger ranges. Thus, for many applications, it is advantageous to apply a 3D reconstruction technique with an FLS rather than utilize a 3D sonar directly.

II. Related Work

Much of the previous work in sonar image processing has focused on image registration, or finding matching features from pairs of sonar images. Rooted in the image registration problem is the need to recover the motion between frames. To solve this problem, Johannsson et al. [9] and Hover et al. [7] extract points with high gradients from the sonar image and cluster the points to use as features. Next, a Normal Distribution Transform (NDT) algorithm is applied to serve as a model for image registration. The entire trajectory of the AUV is put into a pose-graph formulation, and the optimized trajectory shows significant improvements over dead reckoning from the Doppler velocity log (DVL). However, to solve the ambiguity in elevation of the points presented by sonar, the points are assumed to lie on a plane that is level with the vehicle. This planar assumption works well for the noncopmlex areas of a ship hull, the main application of their work, but induces large errors for many other environments. ASFM does not require this assumption, making it useful for a wider range of environments. Hurtos et al. [8] explore a different approach, using Fourier-based techniques instead of feature points for registration. However, the authors primarily focus on applications in 2D mapping, so do not address 3D geometry in detail.

Aykin and Negahdaripour [2] relax the planar assumption for pairwise matching of sonar frames but still assume a locally planar surface in order to include shadow information. They show improvements over Johannsson [9] by instead applying a Gaussian Distribution Transform to the images. Negahdaripour [12] extends this work to feature tracking and visual odometry in sonar video. Various other works have explored different ways to recover 3D geometry from sonar images. Babaee and Negahdaripour [3] use a stereo imaging system composed of one sonar and one optical camera where the centers of the two cameras' coordinate systems and their axes align. The trajectory of the stereo system is calculated using opti-acoustic bundle adjustment. However, the use of an optical camera reduces this system's range due to water turbidity. In contrast, ASFM requires only one sensor and water turbidity is not an issue because no optical cameras are involved.

Assalih [1] once again exploits the stereo idea, but instead uses two imaging sonars placed one on top of the other. Our work is more similar to Brahim et al. [4] where point-based features are used with evolutionary algorithms to recover 3D geometry from pairs of sonar frames. Unlike Assalih and Brahim however, ASFM is capable of using information from multiple viewpoints as opposed to only pairs of images. Multiple viewpoints add more information and can further constrain the problem to result in more accurate reconstruction than pairwise matching.

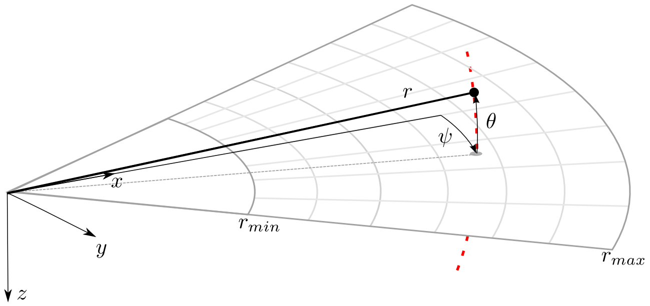


Fig. 2: Imaging sonar geometry. Any 3D point along the dashed red arc will appear as the same image point in the $x - y$ plane. Range r and bearing angle ψ are measured, but the elevation angle θ is lost in the projection process.

A great deal of previous work has been done in computer vision on optical structure from motion [6, 14], and we show in this paper that many of the same ideas can be applied to imaging sonar with a couple of important differences. For instance, optical cameras return bearing and elevation of a scene point, but not depth, while sonar returns bearing and depth, but no elevation. Consequently, sonar projection functions differ from those of optical cameras and present several challenges that need to be addressed before structure from motion can be applied to sonar. In this paper, we present our solution to these challenges including an analysis of degenerate cases, special situations in which reconstruction is not possible without additional information. Such additional information could be provided by other sensors on the AUV such as an inertial measurement unit.

In summary, ASFM's main advantages over related work are 1) the lack of a planar assumption and 2) the ability to use information from more than two sonar images to automatically recover 3D structure and motion more accurate than those recovered with only pairwise comparisons.

III. Acoustic Structure From Motion

We address in this work how to recover the 3D position of features from multiple observations of the same scene, while at the same time constraining the sensor poses. This is a challenging problem because a single sonar image is not sufficient to recover the 3D geometry of the scene. As seen in Fig. 2, the sonar only provides partial information about a feature (bearing ψ and range r) and does not provide its elevation angle θ.

A. Problem Formulation

We represent the ASFM problem as a factor graph [11] (Fig. 3). A factor graph is a bipartite graph with two node types: variable nodes that represent the poses x_i and point features l_j to be estimated, and factor nodes that represent odometry u_i and point feature measurements m_k. An edge in the factor graph connects one factor node with one variable

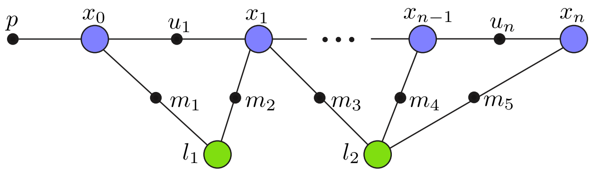


Fig. 3: Factor graph representation of the acoustic structure from motion problem. Variable nodes consist of the underwater vehicle poses x_i and the point features l_j. The black dots represent factor nodes, which are derived from odometry measurements u_i and feature observations m_k. The unary factor p represents a prior on the first pose that defines the reference frame.

node. Here, almost all factors are binary, i.e. they connect only two variables. Only one factor, p, is unary, and defines a reference frame, eliminating otherwise unconstrained degrees of freedom.

The factor graph captures the dependence structure of the ASFM estimation problem. The goal of ASFM is to find the maximum probability set of point features and vehicle poses $\Theta = \{x_i, l_j\}$ given all measurements $Z = \{u_i, m_k\}$. The set Θ^* that satisfies this criteria is defined as

$$\begin{aligned}\Theta^* &= \underset{x}{\operatorname{argmax}}\ p(\Theta|Z) \\ &= \underset{x}{\operatorname{argmax}}\ p(\Theta)p(Z|\Theta) \\ &= \underset{x}{\operatorname{argmax}}\ p(x_0)\prod_{k=1}^{m} p(m_k|x_{i_k}, l_{j_k}) \\ &\quad \cdot \prod_{i=1}^{n} p(u_i|x_{i-1}, x_i). \qquad (1)\end{aligned}$$

Here we have used Bayes Theorem to obtain a maximum a posteriori (MAP) solution for Θ^*. We have also exploited the factorization defined by the factor graph, where each term in Eq. (1) corresponds to one of the factors in Fig. 3.

B. Sonar and Odometry Models

To evaluate the probability of a sensor measurement for a given variable configuration, we need to define a generative sensor model. The generative model consists of a geometric prediction given a configuration of poses and points, in combination with a noise model. As is standard in the literature, we assume a Gaussian noise model.

The generative model for odometry measurements is

$$g(x_{i-1}, x_i) + \mathcal{N}(0, \Lambda_i) \qquad (2)$$

where $g(x_{i-1}, x_i)$ predicts the odometry measurement between poses x_{i-1} and x_i.

Similarly, we define the generative model for sonar measurements by

$$h(x_{i_k}, l_{j_k}) + \mathcal{N}(0, \Xi_k) \qquad (3)$$

where $h(x_{i_k}, l_{j_k})$ predicts the sonar measurement (ψ, r). $h(x_{i_k}, l_{j_k})$ first transforms the landmark $l_{j_k} = (x, y, z)$ into the sonar frame based on pose x_{i_k}, obtaining the local coordinates (x_s, y_s, z_s). Bearing ψ and range r are then obtained by

$$\begin{aligned} r &= \sqrt{x_s^2 + y_s^2 + z_s^2} \qquad (4) \\ \psi &= \operatorname{atan2}(y_s, x_s). \qquad (5)\end{aligned}$$

C. Nonlinear Least-Squares

Under the assumption of Gaussian noise, the MAP problem of Eq. (1) simplifies to a nonlinear least-squares problem. Here we use Mahalanobis distance notation defined as:

$$\|x\|_\Sigma^2 = x^T \Sigma^{-1} x. \qquad (6)$$

The nonlinear least-squares problem becomes:

$$\begin{aligned}\Theta^* &= \underset{x}{\operatorname{argmin}}[-\log p(x_0)\prod_{k=1}^{m} p(m_k|x_{i_k}, l_{j_k}) \\ &\quad \cdot \prod_{i=1}^{n} p(u_i|x_{i-1}, x_i)] \\ &= \underset{x}{\operatorname{argmin}}[\,\|x_0\|_\Lambda^2 + \sum_{k=1}^{m} \|h(x_{i_k}, l_{j_k}) - m_k\|_{\Xi_k}^2 \\ &\quad + \sum_{i=1}^{n} \|g(x_{i-1}, x_i) - u_i\|_{\Lambda_i}^2]. \qquad (7)\end{aligned}$$

Here we have made use of the monotonicity of the logarithm function.

We find an initial estimate for the feature points by backprojection of the sonar measurements. We use the first observation of each feature, consisting of a range r and bearing ψ measurement. We apply the backprojection function

$$\begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = r \begin{bmatrix} \cos\psi\cos\theta \\ \sin\psi\cos\theta \\ \sin\theta \end{bmatrix} \qquad (8)$$

where we set the unknown elevation angle θ to 0. The sonar pose x_{i_k} is then used to convert the point from sonar Cartesian coordinates (x_s, y_s, z_s) to world Cartesian coordinates (x, y, z), which serve as initial guesses for the 3D position of the features.

Starting from this initial estimate, the nonlinear least-squares problem is solved by iterative linearization. For nonlinear measurement functions, nonlinear optimization methods such as Gauss-Newton iterations or the Levenberg-Marquardt algorithm solve a succession of linear approximations in order to approach the minimum. A brief overview of the nonlinear least-squares solution is given below. For a more detailed derivation, see [10]. At each iteration of the nonlinear solver, we linearize around the current estimate Θ to get a new, linear least-squares problem in $\mathbf{\Delta}$

$$\underset{\Delta}{\operatorname{argmin}} \|A\mathbf{\Delta} - b\|^2, \qquad (9)$$

where $A \in \mathbb{R}^{M \times N}$ is the measurement Jacobian consisting of $M = 6n + 2m$ measurement rows, and $\mathbf{\Delta}$ is an N-dimensional vector, where $N = 6n + 3m$. Note that the covariances Σ_i, which represent covariances such as Λ_i and Ξ_k in Eq. (7), have been absorbed into the corresponding block rows of A, making use of

$$\|\mathbf{\Delta}\|_\Sigma^2 = \mathbf{\Delta}^T \Sigma^{-1} \mathbf{\Delta} = \mathbf{\Delta}^T \Sigma^{-\frac{T}{2}} \Sigma^{-\frac{1}{2}} \mathbf{\Delta} = \left\|\Sigma^{-\frac{1}{2}} \mathbf{\Delta}\right\|^2. \qquad (10)$$

Once $\mathbf{\Delta}$ is found, the new estimate is given by $\Theta \oplus \mathbf{\Delta}$, which is then used as the linearization point in the next iteration of the nonlinear optimization. The operator $\oplus$ is often simple addition, but for overparametrized quantities such as 3D rotations, an exponential map is used instead to locally obtain a minimal representation.

The minimum of the linear system $A\mathbf{\Delta} - \mathbf{b}$ is obtained by Cholesky factorization. By setting the derivative in $\mathbf{\Delta}$ to zero we obtain the normal equations $A^T A\mathbf{\Delta} = A^T\mathbf{b}$. Cholesky factorization yields $A^T A = R^T R$, and a forward and backsubstitution on $R^T\mathbf{y} = A^T\mathbf{b}$ and $R\mathbf{\Delta} = \mathbf{y}$ first recovers $\mathbf{y}$, then the actual solution, the update $\mathbf{\Delta}$. See [11] for an efficient incremental solution in a recursive setting.

D. Existence of Solution

We discuss under which conditions the system of equations is solvable by analyzing the number of feature points that need to be observed to fully constrain the system. Let n be the number of poses, and m be the number of points to reconstruct. For every pose, there are 6 unknowns $(x, y, z, yaw, pitch, roll)$ and for every point there are 3 unknowns (x, y, z). The first pose is fixed using a prior, so there are 0 degrees of freedom for the first pose. In case all features are visible from each pose, there are $2n$ equations for each point, and the system is fully constrained iff:

$$6(n-1) + 3m \leq 2mn \tag{11}$$

Since we are not restricted to pairs of sonar views, our simulated examples below use information from 3 sonar viewpoints. From Eq. (11) we see that for 3 sonar views, a minimum of 4 points are needed to fully constrain the estimation problem. In our real sonar data below, features from 5 poses are used; thus, a minimum of 4 points are needed to make 3D reconstruction possible.

E. Degenerate Cases

As is the case for optical structure from motion, there are situations in which a unique solution does not exist. We now discuss three such cases that we have also included in the next section's simulation evaluation.

One of these cases is pure translation in the x-direction. This scenario does not allow us to recover elevation of the point features because the circular arc containing the set of possible 3D points in the sonar geometry for the first pose will intersect the circular arc of the same point seen in the next pose, which differs only in x, at two points. These two intersections cause an ambiguity in the elevation of the points symmetric about the zero plane.

Another case is pure pitch rotation. Since all points lying along a circular arc map to the same point in a sonar image, all of the images from this case would be the same. Consequently, we would not have enough information to recover elevation. However, if the sonar pitched so much that the vehicle would have to translate in the z-direction as well to see the same scene, this motion would be able to recover the points well because the overlapping arcs would overlap in a small region.

The third situation that results in multiple solutions is pure yaw and y-translation. Like the other two cases discussed, in this kind of trajectory, the elevation arcs would have a large overlap region. The correct elevation of the feature point could lie anywhere in this overlapping region.

TABLE I: Simulated Data Experimental Design

	Value
Number of Monte Carlo samples	1000
Orientation: stddev (rad)	$\frac{\pi}{180}$
Translation: stddev (m)	0.01
Minimum range of sonar (m)	0.375
Maximum range of sonar (m)	9.375
Bearing FOV of sonar (degrees)	28.8
Elevation FOV of sonar (degrees)	28
Number of bearing bins	96
Number of range bins	512

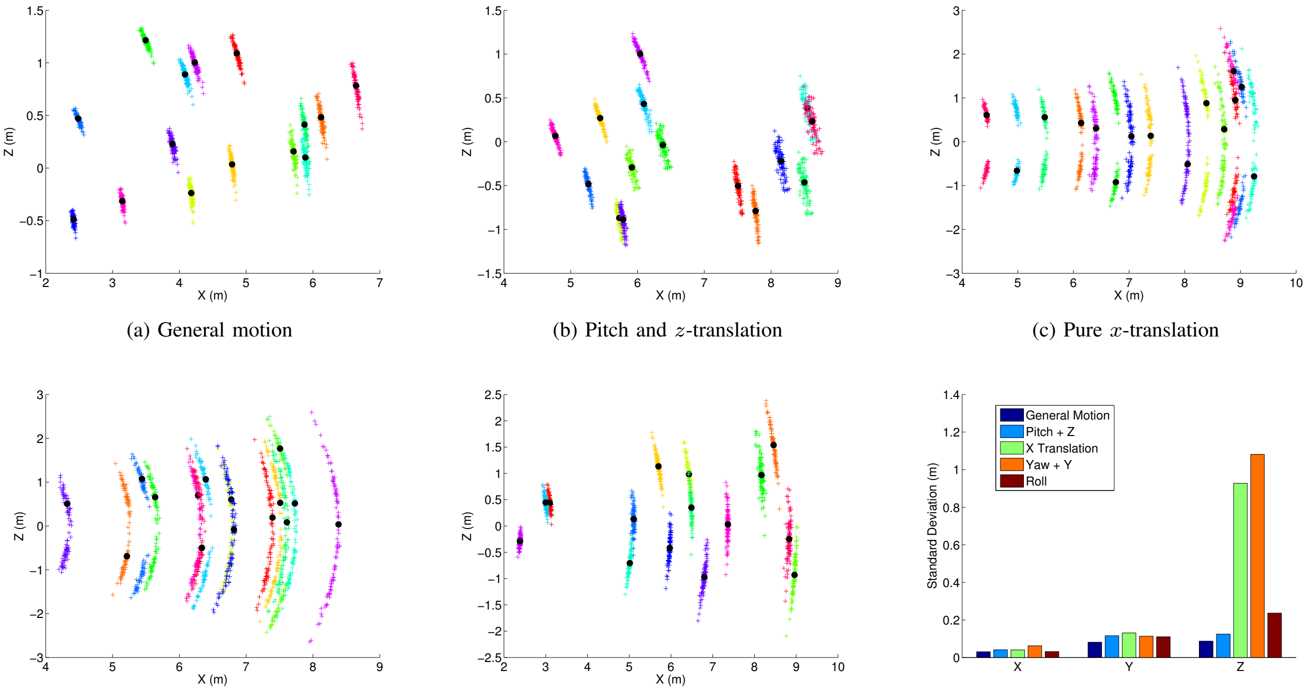
IV. Experimental Results

A. Simulation

We present statistical results for multiple types of vehicle motion using simulated data. The simulation data was generated by selecting three sonar poses containing overlapping regions in their fields of view and randomly creating 3D points until at least 15 points were visible in all three sonar frames. Gaussian noise was added to the bearing ($\sigma = 0.2°$) and range ($\sigma = 0.005$ m) components of the ground truth sonar measurements. Similarly, Gaussian noise was added to both rotational ($\sigma = 1°$) and translational components ($\sigma = 0.01$ m) of the odometry between consecutive poses. The simulated sonar and environment specifications are listed in Table I. Five different sonar trajectories were analyzed:

1) *General Motion*: In this trajectory, the sonar undergoes an x, y, and z-translation as well as changes in yaw, pitch, and roll.
2) *Pitch + Z*: To represent a well-constrained case, we have the sonar go through purely pitch and z-translation motion. This configuration is particularly well-constrained because the different arcs along which a point could lie intersect with very small overlapping regions.
3) *Forward Motion*: One degenerate case is shown through this trajectory of pure x-translation (2 m total). For this motion, the arcs along which the points could lie intersect in two regions, which creates an ambiguity as to whether the point lies in an elevation above or below the zero elevation plane.
4) *Yaw + Y*: Another degenerate case is explored using a pure yaw and y-translation trajectory. The elevation arcs in this case have a large overlapping region, making the z-coordinate of feature points difficult to recover accurately.
5) *Roll*: For this trajectory, the sonar undergoes pure roll motion, 45° in total. This case is fairly well-constrained because the motion rotates the elevation arc about the actual elevation point.

We use Monte Carlo sampling to compare the variations in recovered point features for each motion type. Each sonar trajectory was simulated 1,000 times with the same 3D



(a) General motion

(b) Pitch and z-translation

(c) Pure x-translation

(d) Yaw and y-translation

(e) Roll

(f) Summary: Standard deviations

Fig. 4: Monte Carlo simulation results for the different motion sequences. (a-e) Cluster of point estimation results from 100 random noise trials of the five different simulations. The black dots denote ground truth. (f) Standard deviation of the error for the recovered x, y, and z coordinates over 1000 Monte Carlo simulations, clearly indicating degenerate motion cases for pure x translation as well as for yaw + y motion.

points and noise randomly sampled each time from the same Gaussian distribution with $\mu = 0$ and σ as described above. The standard deviation for the recovered points for each case over the 1,000 runs can be seen in Fig. 4f. The variation in z is greater than the variations in x and y for all the cases as expected. The ambiguity in elevation that is not resolved from the information gained in the degenerate cases causes the variation in z for those situations to be much greater than the other two trajectories. Variations in x and y-coordinates can be attributed to the optimization changing the odometry slightly to meet the measurement constraints.

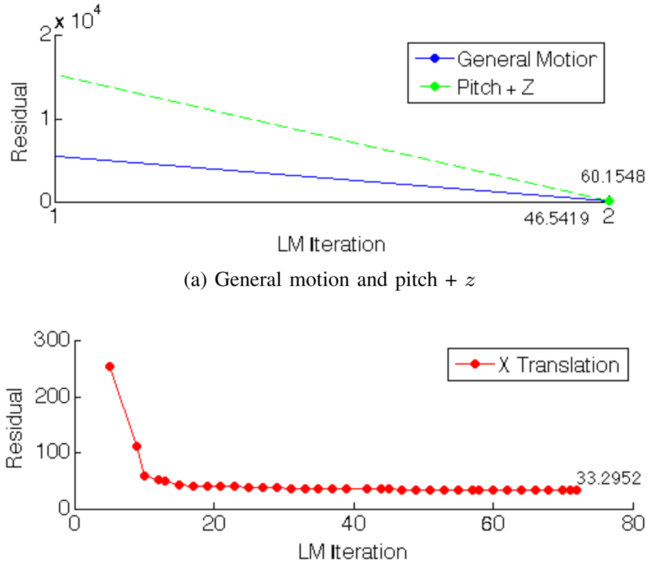
A visualization of the variation for each individual motion provides further insights. The x, z-coordinate distributions for 100 runs of each of the 15 points in each simulation are shown in Fig. 4 with the ground truth marked for comparison. Only 100 runs are shown to avoid cluttering the graph. As seen by the thin bands, the elevation varies much more than the x-coordinate for each point. Note also that the bands are not vertical, but rather trace an arc, which is the elevation arc along which all the points would map to the same point in the sonar image. For the degenerate cases of pure x-translation and pure yaw and y-translation, the symmetric ambiguity about the zero plane is clearly seen. Points were equally likely to appear at the correct elevation or at the same elevation on the opposite side of the zero plane.

Another insight into how well-constrained a trajectory and set of landmarks are is the number of Levenberg-Marquardt (LM) iterations needed until convergence, and the resulting residual. Over 1,000 simulations of each trajectory, the general motion case converged in an average of 2 LM iterations, the pitch and z case converged in an average of 2 LM iterations, and the pure x translation case converged in an average of 74 LM iterations. The yaw and y example converged in an average of 95 LM iterations and the roll case converged in an average of 3 LM iterations. A representative example of the residuals after each LM iteration for the first three simulation trajectories can be seen in Fig. 5. The residuals for the first two simulation cases started off high, but quickly dropped after just one LM iteration to 46.5419 for the general motion case and 60.1548 for the pitch and z case. This indicates that the cost functions for these two trajectories are close to quadratic. The high number of LM iterations needed for the pure x-translation trajectory, which eventually reaches a residual of 33.2952, as well as the yaw and y case implies that the optimization function is not quadratic, but presumably close to flat in at least one direction. The flatness is due to the degenerate geometric configuration, which leads to much slower convergence.

The poses and overall error (geometric distance from the estimated point to the true point) for each simulation can be found in Table II. Note that the point errors for fully constrained situations are less than 0.23 m and general motion has the smallest geometric error of only about 0.11 m. The point errors are much larger in the degenerate cases because the z-coordinates of the recovered points were symmetric about $z = 0$. The odometry is recovered well,

TABLE II: Monte Carlo Simulation Results

	General Motion	Pitch + z	x	Yaw + y	Roll
Feature mean error (m)	0.1090	0.1551	0.9425	1.0549	0.2266
Feature stddev (m)	0.0662	0.0888	0.8339	0.8120	0.1586
Pose 1 (m, m, m, rad, rad, rad)	(0, 0, -1, 0, -0.4, 0)	(0, 0, -2, 0, -0.4, 0)	(0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0)	(0, 0, 0, 0, 0, 0)
Pose 2 (m, m, m, rad, rad, rad)	(-1, 0, 0, 0, 0, 0.3)	(0, 0, 0, 0, 0, 0)	(1, 0, 0, 0, 0, 0)	(0, 2, 0, -0.3, 0, 0)	(0, 0, 0, 0, 0, 0.4)
Pose 3 (m, m, m, rad, rad, rad)	(-0.5, 2, 2, -0.4, 0.4, 0)	(0, 0, 3, 0, 0.5, 0)	(2, 0, 0, 0, 0, 0)	(0, 4, 0, -0.4, 0, 0)	(0, 0, 0, 0, 0, 0.8)
Pose position mean error (m)	0.0144	0.0153	0.0133	0.0132	0.0103
Pose position stddev (m)	0.0062	0.0065	0.0062	0.0062	0.0052
Pose orient. mean error (rad)	0.0141	0.0135	0.0211	0.0239	0.0232
Pose orient. stddev (rad)	0.0082	0.0085	0.0097	0.0110	0.0102
Avg. number of LM iterations	2	2	74	95	3



(a) General motion and pitch + z

(b) x-translation

Fig. 5: Sample residuals for simulation data after each Levenberg Marquardt iteration for a representative run of (a) the general motion case and the pitch + z trajectory and (b) the pure x translation case. The final residual for each case is labeled. Iterations without a residual in (b) indicate a rejected LM step due to an increase in error.

largely due to a good initial guess (perfect odometry with added Gaussian noise (σ listed in Table I)). A promising result is that the general motion simulation performs very well, suggesting that ASFM could work well for inspection and surveying applications.

B. Imaging Sonar Sequence

We demonstrate 3D structure recovery from several imaging sonar frames recorded with an underwater robot. This sequence of a ladder on a dock was taken on a Bluefin Hovering Autonomous Underwater Vehicle (HAUV) (Fig. 6) in Boston, Massachusetts. Five sonar frames were selected from the dataset to perform ASFM, three of which can be seen in Fig. 7. Corresponding features were manually selected and matched from all five sonar frames.

The imaging sonar used is a SoundMetrics DIDSON 300m forward-looking sonar. It has a $\psi_{max} = 28.8°$ bearing field of view (FOV) and a 28° vertical FOV (using a spreader lens). Note that the theory behind ASFM would not be affected by a sonar with a narrower vertical FOV. The only problem that may arise with a smaller FOV is that it may be more challenging to find enough matching features between sonar frames. The DIDSON discretizes returns into $w = 96$ bearing bins and $h = 512$ range bins. The DIDSON mode used for this dataset provides a minimum range of $r_{min} = 0.75$ m and a maximum range of $r_{max} = 5.25$ m. Let (u, v) be the image coordinates of a feature in the Cartesian sonar image, and γ be a constant describing the number of pixels per meter in the Cartesian image. Then, the bearing ψ and range r are obtained using:

$$\begin{aligned} \gamma &= \frac{w}{2r_{max}\text{sin}(\frac{\psi_{max}}{2})} && (12)\\ x_s &= \frac{u - \frac{w}{2}}{\gamma} && (13)\\ y_s &= r_{max} - \frac{v}{\gamma}. && (14)\\ r &= \sqrt{x_s^2 + y_s^2} && (15)\\ \psi &= \frac{180}{\pi}\text{atan2}(x_s, y_s). && (16) \end{aligned}$$

Next, we partition the sonar field of view into discrete range and bearing bins and find the bins that contain the desired point using the following expressions:

$$\begin{aligned} n_r &= \frac{h(r - r_{min})}{r_{max} - r_{min}} && (17)\\ n_b &= M_4(w, \psi) && (18) \end{aligned}$$

where n_b is the bearing bin, n_r is the range bin, and $M_4(w, \psi)$ is a third-order polynomial (with 4 coefficients determined by w) given by the sonar manufacturer that accounts for lens distortion.

The manually selected feature points were placed into the factor graph optimization using the mapping described in Eq. (13)-(18). Odometry readings from the vehicle were also used in the optimization to further constrain the problem. Because of the short time elapsed between the sonar frames, drift in vehicle navigation is minimal, and the odometry readings from the vehicle are very accurate. We chose uncertainties of $\sigma = 1°$ for odometry rotation and $\sigma = 0.1$ m for odometry translation. For bearing and range measurements from the DIDSON sonar we use $\sigma = 0.2°$ and $\sigma = 0.005$ m respectively.

Fig. 8 shows how the optimization reduces errors in the location of the point features initially very quickly over a few iterations. Near the minimum, each iteration reduces errors

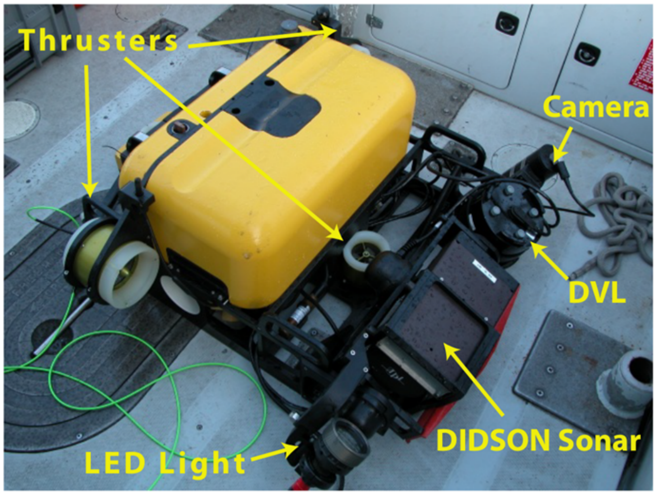


Fig. 6: Bluefin Hovering Autonomous Underwater Vehicle (HAUV) used in our real data experiments. The DIDSON sonar is pictured attached to the front of the vehicle.

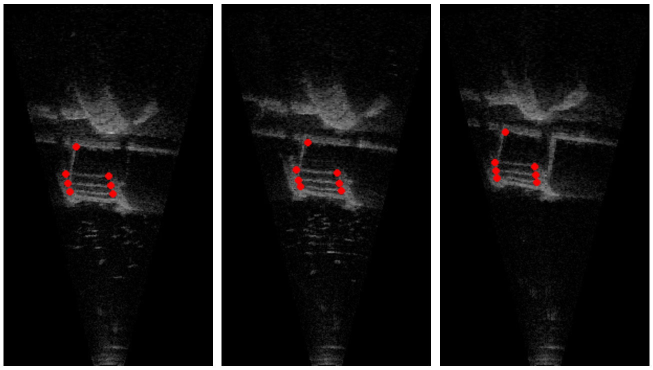


Fig. 7: Manually marked features (red circles) for three of the five raw sonar frames that were used to reconstruct the ladder geometry.

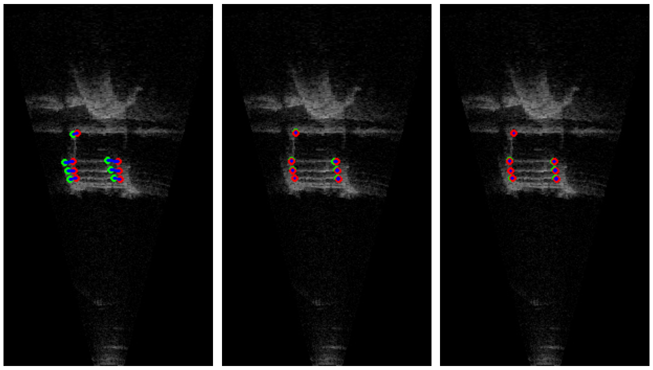
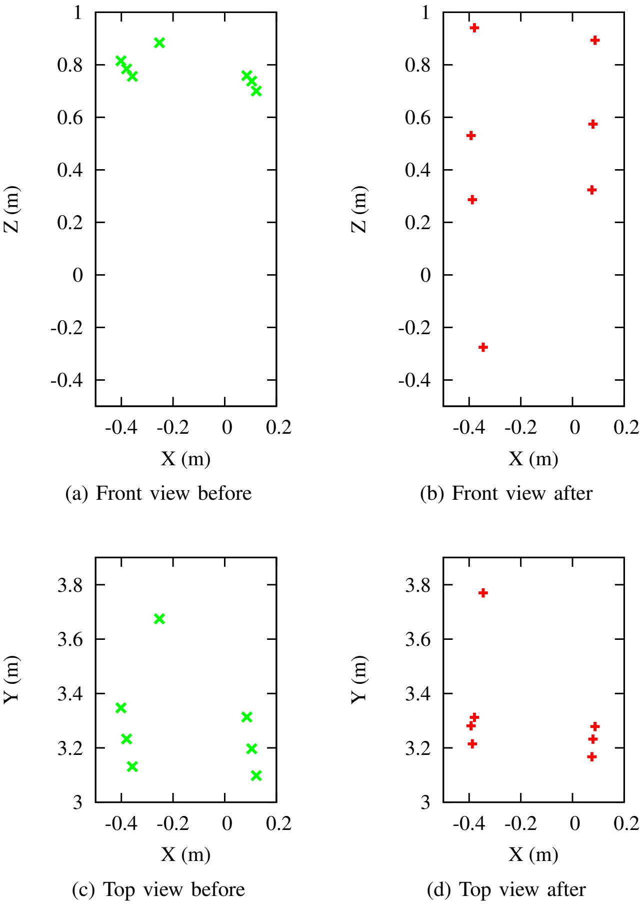


Fig. 8: Reprojection error for the last sonar frame (left) from initialization, (center) after 5 LM iterations, and (right) after a solution was found (27 LM iterations). The red circles indicate the manually selected features and the green circles indicate the reprojected features. The blue lines show the reprojection error used in the ASFM optimization.



(a) Front view before

(b) Front view after

(c) Top view before

(d) Top view after

Fig. 9: (a, b) Front and (c, d) top views of 3D ladder structure (a, c) before and (b, d) after optimization from five imaging sonar frames.

more slowly. The final reprojection error is shown in the last frame. Since no ground truth is available for this dataset, we use reprojection error on the Cartesian image as one indicator for ASFM's performance. As seen in Fig. 8, each recovered point is only off from the manually selected point by one or two pixels. The optimization for this imaging sonar sequence took 27 LM iterations and had an ending residual of 52.8032.

The 3D geometry of the ladder in the imaging sonar sequence was recovered as shown in Fig. 9. Before optimization, the ladder is initialized as a flat object lying in the $x-y$ plane. The structure in the $x-y$ plane looks convincing, but from the $x-z$ view, it is clear that the initialization does not capture the reality that the ladder's rungs are at different z elevations. An important note is that the Cartesian sonar images are flipped up-down, meaning that the top of the image is in fact at a lower elevation in the world than the bottom of the image. Because the sonar is tilted down (about $15°$), the first returns encountered are from the top of the ladder, while the last returns (top of the image) are from a piling. The recovered structure of the ladder shows the distinct ladder rungs. Without ground truth, it is difficult to determine the geometric error between the recovered points and the true 3D points. Going off the assumption that the steps are spaced evenly on the ladder, we can estimate our maximum error to be about 0.2 m given that the top point on the left side of the ladder is spaced about 0.2 m farther than the spacing between the other points.

V. Conclusion

We have presented a novel algorithm for recovery of 3D point features from multiple sonar views, while also constraining the poses from which the images are taken. In contrast to previous solutions, we do not make any planar surface assumption. Simulations of several types of sonar trajectories show the ability of ASFM to recover 3D structure with low uncertainty for general trajectories. They also show a limitation of ASFM in its failure to recover elevation of points for motions that provide poor constraints such as in the case of pure x-translation. An experiment with real sonar data and manually extracted feature points further demonstrates ASFM's 3D reconstruction capabilities.

From the imaging sonar sequence, we note that good initialization is needed for reliable reconstruction. Consequently, of great interest for future work is automatic feature extraction and incremental data association. These problems present several challenges, including the characteristically low number of feature points in a sonar image and sonar's typical low signal to noise ratio. In addition, due to the elevation ambiguity present in sonar, another challenge is how to determine whether a point seen in one sonar image is really the same point seen in another sonar image. For instance, two different points on a vertical, arced object would map to the same point if seen straight on, but would map to two different points if seen from a non-zero yaw or roll angle.

A possible solution to automatic data association is the use of Joint Compatibility Branch and Bound [13], which takes into account geometric relationships between feature points and is computationally feasible for sonar where few feature points are available. With data association, ASFM becomes a potential method for registering loop closures and improved navigation, even for AUVs whose primary motion is in the forward x-direction. For such AUVs, the loop closure trajectory will most likely not be a degenerate case, making ASFM feasible. Further investigation is also needed into different parametrizations with the goal of reducing nonlinearity to improve convergence properties. Currently, good initial pose estimates are needed to ensure convergence. Future work also includes applying acoustic structure from motion to side-scan sonar.

Acknowledgment

The authors would like to thank Dr. Jason Stack for his support, Pedro Vaz Teixeira for recording the ladder sequence and the anonymous reviewers for their helpful and detailed comments.

References

[1] H. Assalih, "3D reconstruction and motion estimation using forward looking sonar," Ph.D. dissertation, Heriot-Watt University, 2013.

[2] M. Aykin and S. Negahdaripour, "On feature matching and image registration for two-dimensional forward-scan sonar imaging," *J. of Field Robotics*, vol. 30, no. 4, pp. 602–623, Jul. 2013.

[3] M. Babaee and S. Negahdaripour, "3-D object modeling from occluding contours in opti-acoustic stereo images," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, Sep. 2013.

[4] N. Brahim, D. Gueriot, S. Daniel, and B. Solaiman, "3D reconstruction of underwater scenes using DIDSON acoustic sonar image sequences through evolutionary algorithms," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, Santander, Spain, Jun. 2011.

[5] E. Coiras, Y. Petillot, and D. Lane, "Mutliresolution 3-D reconstruction from side-scan sonar images," *IEEE Trans. on Image Processing*, vol. 16, no. 2, pp. 382–390, Feb. 2007.

[6] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003, second Edition.

[7] F. Hover, R. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *Intl. J. of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, Oct. 2012.

[8] N. Hurtos, X. Cufi, Y. Petillot, and J. Salvi, "Fourier-based registrations for two-dimensional forward-looking sonar image mosaicing," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2012.

[9] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, Oct. 2010.

[10] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Trans. Robotics*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.

[11] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *The International Journal of Robotics Research*, vol. 31, pp. 217–236, Feb. 2012.

[12] S. Negahdaripour, "On 3-D motion estimation from feature tracks in 2-D FS sonar video," *IEEE Trans. Robotics*, vol. 29, no. 4, pp. 1016–1030, Aug. 2013.

[13] J. Neira and J. D. Tardos, "Data association in stochastic mapping using the joint compatibility test," *IEEE Trans. Robotics and Automation*, vol. 17, no. 6, pp. 890–897, Dec. 2001.

[14] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. LNCS, W. Triggs, A. Zisserman, and R. Szeliski, Eds., vol. 1883. Springer Verlag, 2000, pp. 298–372.