Feature-based SLAM for Imaging Sonar with Under-constrained Landmarks

Eric Westman, Akshay Hinduja, and Michael Kaess

***Abstract*— Recent algorithms have demonstrated the feasibility of underwater feature-based SLAM using imaging sonar. But previous methods have either relied on manual feature extraction and correspondence or used prior knowledge of the scene, such as the planar scene assumption. Our proposed system provides a general-purpose method for feature-point extraction and correspondence in arbitrary scenes. Additionally, we develop a method of identifying point landmarks that are likely to be well-constrained and reliably reconstructed. Finally, we demonstrate that while under-constrained landmarks cannot be accurately reconstructed themselves, they can still be used to constrain and correct the sensor motion. These advances represent a large step towards general-purpose, feature-based SLAM with imaging sonar.**

I. Introduction

Imaging sonar sensors provide rich measurements in underwater scenes which can be useful for the robotic tasks of localization and mapping. However, the ambiguity in the elevation angle of the measurements has been a major obstacle to fully utilizing the data provided by these sensors. Various efforts have been made to mitigate this problem, including assuming a planar environment or using other prior knowledge of the scene.

Monocular cameras and imaging sonars are analogous in that both measure the 3D environment in the form of a 2D image. In the monocular case, using the pinhole camera model, each pixel corresponds to a ray in 3D space. Using the spherical point parameterization of bearing (or azimuth), elevation, and range, the ray describes the bearing and elevation angles of the 3D point measured by the pixel. However, the *range* of the point along the ray is lost in the projection. Analogously, while the bearing angle and range are measured by a pixel in a sonar image, the *elevation* angle of a point measurement is lost in the projection.

Using point correspondences between monocular images of the same environment, bundle adjusment may be applied to disambiguate the range of point measurements and generate full 3D reconstructions of feature points [17]. Acoustic structure-from-motion (ASFM) applies the bundle adjustment framework from visual SLAM to sonar images in order to recover the elevation angle of sparse feature points using observations from multiple viewpoints [6, 7]. In doing so, no assumptions are made about the structure of the imaged scene, and the 3D position of landmarks may be reconstructed when sufficient constraints are present.

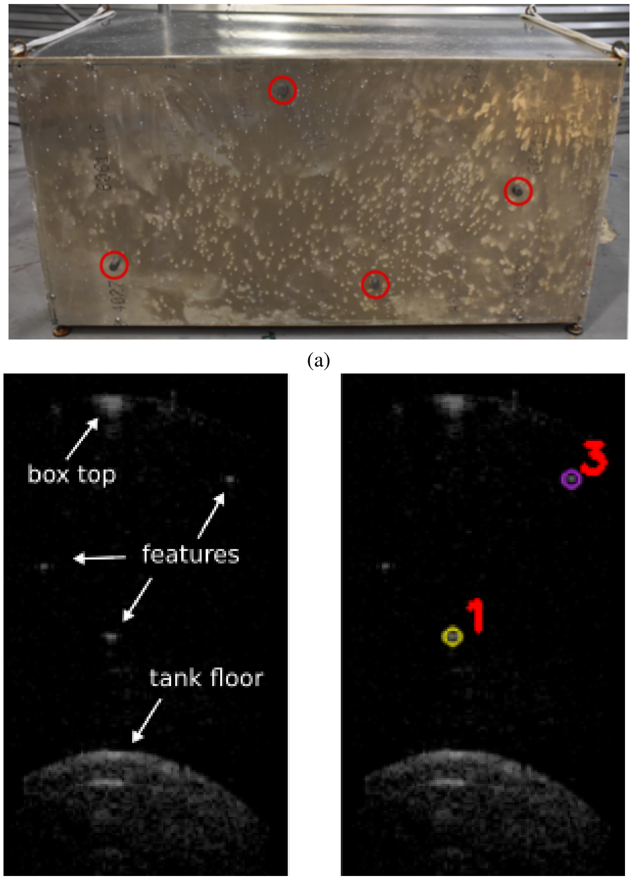

(a)
(b)

Fig. 1: (a) The box used in our test tank experiments. Magnets are placed on the sides of the box to provide features for proposed SLAM algorithm and are circled in the image. The smaller spots are corrosion and do not appear in the sonar images. (b) A sample imaging sonar frame from one of our experimental datasets. The left image labels relevant structures in the scene, and the right image shows the detected features and their corresponded landmarks.

The original ASFM algorithm suffers from several notable deficiencies limiting its applicability in real-world scenarios. First, the algorithm relies on manually extracted and associated features. Second, the algorithm models the belief of all variables as normally distributed. This model is accurate when the measurements of a landmark provide sufficient constraints to disambiguate the elevation angle. However, the elevation angle of a landmark may often be under-constrained by the measurements, in which case the belief of the elevation angle is closer to a uniform distribution over the feasible range. The unconstrained nonlinear optimization underlying ASFM may estimate such landmarks to lie out-

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

This work was partially supported by the Office of Naval Research under award N00014-16-1-2365.

side of the field of view of the sensor, which can in turn increase the error of the optimized sensor poses. Accurately modeling landmark positions is crucial to constraining the sensor motion and reducing error in localization.

In this work we address these short-comings of ASFM to develop a fully-automatic, feature-based acoustic SLAM algorithm that can be used to reduce localization error underwater and accurately reconstruct landmarks when possible. Specifically, the new contributions of this work are:

1) A fully-automatic acoustic SLAM pipeline for *general-purpose* use with imaging sonar (i.e. no assumptions made about scene geometry), including feature extraction and association
2) A method of detecting *well-constrained* landmarks, which may be explicitly modeled and accurately reconstructed without negatively influencing pose estimation
3) A novel formulation of the feature-based SLAM problem which appropriately handles *under-constrained* landmarks to constrain and correct the sensor motion.

II. Related Work

Johannsson et al. used imaging sonar to aid localization by finding high-gradient pixel clusters and using the Normal Distribution Transform (NDT) to register sonar images [10]. This method assumes the imaging target is a flat seafloor. This assumption of global or local planarity has been utilized to simplify the problem in other related works as well [5, 13]. While this assumption is reasonable in the case of mapping the seafloor or the non-complex area of a ship-hull, it does not apply to general environments where structures may be complex and non-planar, such as the running gear of a ship, coral reefs, or underwater pipelines.

Acoustic structure from motion (ASFM) [6] was introduced as a method of recovering the 3D position of sparse point features from a moving imaging sonar. Taking inspiration from visual structure from motion and bundle adjustment, the SLAM problem is formulated as a nonlinear least squares optimization and solved using Levenberg-Marquardt. This work demonstrated the ability to disambiguate 3D structure by utilizing multiple viewpoints. An evaluation of several degenerate motion cases is performed, which highlights how certain types of motion do not provide sufficient constraints to accurately estimate the elevation angle of observed landmarks. However, this work does not examine the effect that mapping such under-constrained landmarks has on the pose estimates within the SLAM framework. As we demonstrate in this paper, naively applying the ASFM framework to under-constrained landmarks degrades the sonar state estimate and results in *higher* localization error. In this work, we take special care to address the degeneracy of the under-constrained landmarks in order to improve the sonar state estimate in a SLAM framework.

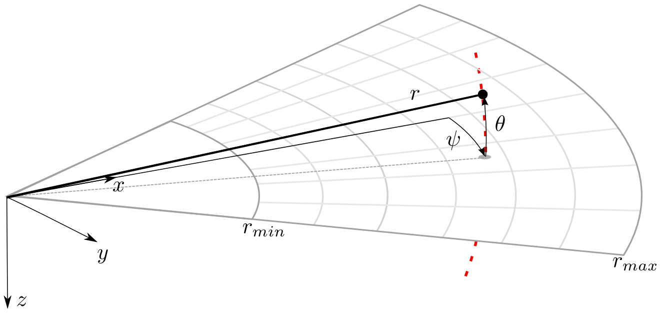


Fig. 2: Imaging sonar geometry. The gray grid shows the Cartesian image projected on the xy plane. Any sound reflected by points along the elevation arc (red dashed line) will contribute to the corresponding pixel intensity. The pixel's location within the image specifies the measurement's bearing angle and range.

Several recent works have developed methods to automatically extract and associate feature points for use in SLAM. Ji et al. [9] utilize the Hough transformation to extract line segments from sonar images. The endpoints are taken as point features, which are automatically associated based on geometric criteria. While the efficacy is demonstrated by reconstructing corner points on a triangular prism placed in a test tank, using lines to determine feature points does not generalize to other types of structure which may entirely lack straight edges. [14] applied KAZE / A-KAZE (accelerated KAZE) features for automatic feature detection. A RANSAC homography estimation method is provided to reject correspondence outliers. Like other works, though, this method assumes that all detected points lie on a flat seafloor or test tank floor. High-dimensional features have been extracted from sonar images and used for underwater localization with promising results [12], although this is done under the framework of performing loop closures relative to a locally planar surface, rather than explicitly tracking and mapping specific 3D points. In contrast, we extract and associate geometric point features, which can be reconstructed accurately when sufficient constraints are present and may be useful for mapping.

III. ASFM Background

In this section we describe the geometry of imaging sonar measurements and the basics of the fundamental ASFM algorithm.

A. Imaging Sonar Geometry

Consider a point $C = \begin{bmatrix} x & y & z \end{bmatrix}^\top$ parameterized in the local Cartesian sonar coordinate frame. The point may also be expressed as S using the spherical parameterization, and the conversion between the two is

$$C = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = r \begin{bmatrix} \cos\psi\cos\theta \\ \sin\psi\cos\theta \\ \sin\theta \end{bmatrix} \tag{1}$$

$$S = \begin{bmatrix} \psi \\ r \\ \theta \end{bmatrix} = \begin{bmatrix} \operatorname{arctan2}(y, x) \\ \sqrt{x^2 + y^2 + z^2} \\ \operatorname{arctan2}\left(z, x^2 + y^2\right) \end{bmatrix} \tag{2}$$

where ψ is the bearing angle, r is the range, and θ is the elevation angle. An imaging sonar generates partial spherical measurements by using a one-dimensional array of transceivers to send out acoustic signals into a target volume, called the frustum, and receive the reflected signals. r is

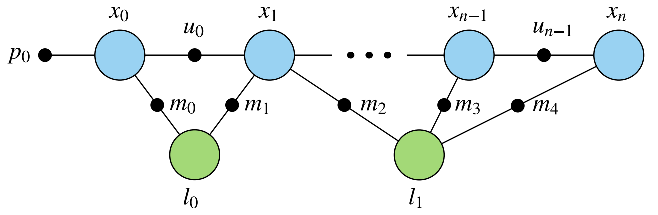


Fig. 3: Factor graph representing the basic ASFM algorithm. A two-dimensional bearing-range measurement relates a sensor pose node x_t to a point landmark node l_j. This type of factor provides a two-dimensional error function in the bearing-range space, which is the difference between the raw measurement and the predicted bearing-range measurement. The prediction is based on the estimate of the 3D position of landmark l_j, which is parameterized in spherical coordinates relative to its base pose.

determined by the time of flight and the speed of sound in water. The array of transceivers allows the bearing angle ψ of a received reflection to be calculated to within $< 1°$ of accuracy. However, the measurements do not provide any information about the elevation angle θ. Therefore, detected sonar returns that are reflected from surface patches that lie on the same elevation arc will project to the same pixel in the resulting imaging sonar image, as seen in Figure 2. Compiling all measurements within the sensor field of view results in a grayscale polar coordinate image, where columns correspond to discrete bearing angle bins and rows correspond to discrete range bins. For a pixel p, let $m = M(p) = (\psi, r)$ denote the transformation from pixel space to bearing-range space. The intensity of a pixel corresponds to the intensity of the sound reflected from the elevation arc at the specified bearing angle and range.

This type of sonar may be used with a lens that results in an elevation field of view as small as 1°, which is often referred to as "profiling mode." This significantly reduces the ambiguity in the elevation angle of measurements, which can simply be estimated as 0° with little error. However, the imaged volume is so small that very little overlap between volumes is attained as the vehicle moves. For this reason a spreader lens, which provides a 28° elevation field of view, is often utilized to image larger volumes. The ASFM algorithm is designed to address the problem of reconstructing points detected in this configuration with a large elevation field of view, often referred to as "imaging mode".

B. ASFM Algorithm

We follow Huang et al. [6, 8] in formulating ASFM as a nonlinear least-squares factor-graph optimization, a framework that is commonly used to solve the SLAM problem [11]. A factor graph is a bipartite graph in which *variable* nodes, which represent the unknown variables to be optimized, connect to *factor* nodes, which represent the measurements. An example factor graph depicting the ASFM problem is shown in Figure 3.

At each timestep t, the pose x_t is added as a new node to the factor graph along with the odometry measurement u_{t-1}, which provides a motion estimate between x_{t-1} and x_t. A bearing-range measurement m_k of the jth landmark is added to the graph, connecting the 3D point node l_j to the pose from which it was observed. Two different parameterizations of the point landmarks may be used: Cartesian with respect to the global frame and spherical with respect to the frame of the "base pose" (the first pose at which the landmark was observed). In this work, we use the spherical parameterization, as it is more amenable to our proposed methods and generally results in a more linear system, as described in [8]. The initial estimate for a point landmark may be generated either by assuming 0° elevation or utilizing the linear triangulation method described in [18].

The factor graph is solved as a nonlinear least squares optimization using the Levenberg-Marquardt algorithm. The overall objective function that is minimized is is the sum of all of the costs defined by each factor:

$$\underset{X}{\operatorname{argmin}} \sum_i \|h_i(X) - z_i\|^2_{\mathbf{\Sigma}_i} \tag{3}$$

where the state vector $X = [x_0, x_1, \ldots, l_0, l_1, \ldots]^T$ contains all unknown variables: the poses and landmarks. The ith factor specifies a prediction function $h_i(X)$, a measurement z_i, and a measurement uncertainty $\mathbf{\Sigma}_i$. In the case of a bearing-range measurement of landmark j from pose x_t, the prediction function is $h_i(X) = \pi(x_t, l_j)$, which transforms the estimated 3D landmark position into the sonar coordinate frame of pose x_t, and projects the point into the 2D bearing-range space according to Equation 2. The corresponding backprojection function $\pi^{-1}(x_b, m_b, \theta)$ computes a 3D landmark position based on the base pose x_b, a corresponding bearing-range measurement m_b, and a provided elevation angle θ. See [6] for additional technical details.

IV. Automatic Feature Extraction and Correspondence

The low signal-to-noise ratio and nonlinear geometry of sonar sensors present significant challenges to the tasks of identifying and corresponding environmental features. A-KAZE features [1] are specifically designed to handle images with high speckle noise by utilizing diffusion in nonlinear scale spaces. This type of diffusion smooths the image but maintains high-gradient boundaries. Such features were used for underwater sonar mapping in [14] under the assumption that all feature points lie on a flat seafloor. A sample sonar image with A-KAZE features is shown in Figure 1.

As in [14], we extract A-KAZE features from the sonar images. We associate features based only on geometric criteria, rather than descriptor similarity, due to the similar appearance of features in our experiments and the relatively low bandwidth of sonar images. We find feature correspondences from the current image to a database of all previously identified features. Algorithm 1 details the projective data association algorithm used to correspond a feature measurement $m_i = (\psi_i, r_i)^\top$. All detected features from the current frame are processed by the algorithm in arbitrary order, with no duplicate assignments being accepted. The algorithm is conservative—a detected feature will only be corresponded with a previously identified feature (or identified as a new

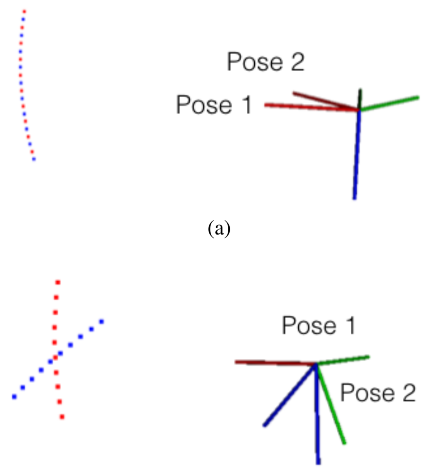
Algorithm 1 Projective data association algorithm to match detected feature f_i with a previously observed feature. D_1 and D_2 are empirically selected thresholds in units of pixels.

```
Input: the feature measurement to be associated $m_i = (\psi_i, r_i)^\top$, the current estimated pose $x_t$, the first measurement of all features in the database $g_1, \ldots, g_{n_f}$, their corresponding base pose estimates, $x_{g_1}, \ldots, x_{g_{n_f}}$, and the set of database features $\mathcal{J}$ that have already been matched to features from the current image
Output: index $j^*$ of the matched feature ($-1$ if no match, $n_f + 1$ if identified as a new feature)
$j^* \leftarrow -1$
$d_{min} \leftarrow \infty$
$n_v \leftarrow 0$
$\Theta \leftarrow \{\theta_{min}, \theta_{min} + \Delta\theta, \ldots, \theta_{max} - \Delta\theta, \theta_{max}\}$
for $j \in \{1, \cdots, n_f\}$ do
    $d_j \leftarrow \min_{\theta\in\Theta} \|M^{-1}(m_i) - M^{-1}(\pi(x_t, \pi^{-1}(x_{g_j}, g_j, \theta)))\|_2$
    if $d_j < d_{min}$ then
        $d_{min} \leftarrow d_j$
        $j^* \leftarrow j$
        if $d_j < D_1$ then
            $n_v \leftarrow n_v + 1$
        end if
    end if
end for
if $n_v = 0$ AND $d_{min} > D_2$ then
    $j^* \leftarrow n_f + 1$
else if $n_v \neq 1$ OR $j^* \in \mathcal{J}$ then
    $j* \leftarrow -1$
else
    $\mathcal{J} \leftarrow \mathcal{J} \cup j^*$
end if
```

feature to track) when there is little ambiguity. Otherwise, the feature measurement is disregarded.

V. ASFM with Under-constrained Features

Utilizing the automatic feature extraction and correspondence method described in Section IV with the basic ASFM algorithm will actually degrade the overall sensor motion estimate (see Section VI for examples using simulated and experimental data). This is because a large number of point landmarks observed by an imaging sonar from multiple views will be under-constrained. Consider the case in which a point feature is observed by a sonar wherein the sensor is purely rotated about the z-axis between two poses (pure yaw rotation), as in Figure 4. The elevation arcs along which the point may lie according to each measurement are exactly the same—the elevation angle of the point is completely ambiguous despite having measured the point from multiple poses. Under the standard ASFM framework, a landmark constrained by such measurements would make the overall optimization rank-deficient and degenerate. With noise in the measurements, the optimization may be solvable, but the landmark's estimated elevation angle would be determined entirely by the noise, not by constraints provided by the measurements. This would result in large errors in landmark estimation, which will negatively impact the accuracy of the pose estimation as well.



(a)

(b)

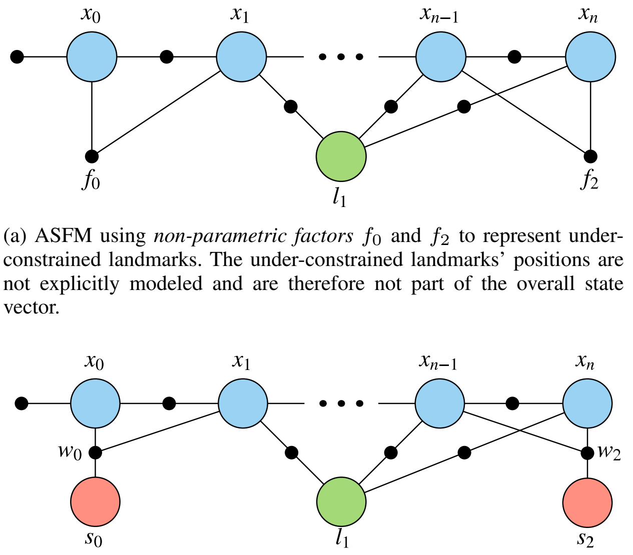
Fig. 4: (a) The elevation arcs corresponding to measurements of the same point from pose 0 (red points) and pose 1 (blue points) separated by pure yaw motion are exactly aligned and the point's elevation angle is entirely ambiguous. (b) The elevation arcs have minimal overlap when the poses are separated by pure roll rotation—the observed point's elevation angle is well-constrained.

We utilize a two-tiered system that categories observed feature points as either *under-constrained features* or *well-constrained features*. After a feature is detected and corresponded, it is tested to see if the measurements sufficiently constrain its elevation angle, by a process described in the following subsection. If so, it is added to the factor graph as a well-constrained landmark using the standard parameterization from [6]. Otherwise, it is added as a under-constrained landmark. We propose two different parameterizations of under-constrained landmarks that model the elevation angle non-parametrically, rather than assuming a normally-distributed belief, as detailed in sections V-B and V-C. Whenever a under-constrained landmark is observed, the same check is performed to determine if it may be upgraded. Upon upgrading a under-constrained landmark, its corresponding node and / or factor are removed from the graph, and the landmark is re-added using the standard 3-DOF spherical parameterization, along with its corresponding measurements.

A. Identifying Well-constrained Features

In order to determine if a point landmark is well-constrained, we formulate an additional nonlinear optimization in which the state consists of only the landmark l_j, using the 3-DOF spherical parameterization:

$$l_j^* = \underset{l_j}{\operatorname{argmin}} \sum_i \|h_i(l_j) - z_i\|^2_{\mathbf{\Sigma}_i} \tag{4}$$



(a) ASFM using *non-parametric factors* f_0 and f_2 to represent under-constrained landmarks. The under-constrained landmarks' positions are not explicitly modeled and are therefore not part of the overall state vector.

(b) ASFM using a *semi-parametric* representation of under-constrained landmarks. Nodes s_0 and s_2 represent the bearing and range components of the corresponding landmarks, which are explicitly modeled as normally distributed variables in the overall optimization.

Fig. 5: The three different ASFM frameworks evaluated. In this example, landmarks l_0 and l_2 are under-constrained, while l_1 is well-constrained.

Since the sensor poses are not state variables, they are considered constant and the prediction function $h_i(l_j)$ uses the latest estimates available from the overall factor graph state estimate. Note that this formulation requires a prediction of the most recent pose, which may come from any sensor that estimates odometry, such as an inertial measurement unit (IMU), or a motion model. Taking an initial estimate of the landmark l^0 as a linearization point, we use the Taylor series expansion of the measurement function

$$h_i\left(l_j\right) = h\left(l_j^0 + \Delta_j\right) \approx h\left(l_j^0\right) + \mathbf{H}_i\Delta_j \tag{5}$$

$$\mathbf{H}_i := \left.\frac{\partial h_i\left(l_j\right)}{\partial l_j}\right|_{l_j} \tag{6}$$

to simplify the optimization as a linear least squares problem

$$\begin{aligned} l_j^* &\approx \underset{\Delta_j}{\operatorname{argmin}} \sum_i \left\| h_i(l_j^0) + \mathbf{H}_i\Delta_j - z_i \right\|_{\mathbf{\Sigma}_i}^2 & (7)\\ &= \underset{\Delta_j}{\operatorname{argmin}} \left\| \mathbf{A}\Delta_j - b \right\|^2 & (8)\end{aligned}$$

where $\Delta_j = l_j - l_j^0$ is the state update vector [3] . $\mathbf{A}$ and b are obtained by stacking the blocks

$$\begin{aligned} \mathbf{A}_i &= \mathbf{\Sigma}_i^{-1/2}\mathbf{H}_i & (9)\\ b_i &= \mathbf{\Sigma}_i^{-1/2}\left(z_i - h_i\left(l_j^0\right)\right). & (10)\end{aligned}$$

The linearization point l_j^0 is taken to be the first bearing-range measurement backprojected at zero elevation (the linear triangulation method from [18] may also be used here, although it may not provide a good estimate if sufficient constraints are not present).

As discussed in detail in [19], examining $\mathbf{A}^T\mathbf{A}$ is the key to determining if the optimization is well-constrained by the measurements. The 3×3 matrix $\mathbf{A}^T\mathbf{A}$ will be rank deficient if the elevation angle is entirely unconstrained, as in the case of pure yaw rotation with noiseless measurements, depicted in Figure 4. As the elevation angle becomes more constrained, the smallest eigenvalue of $\mathbf{A}^T\mathbf{A}$, λ_3, will increase in magnitude relative to the first two eigenvalues λ_1 and λ_2. Landmarks are therefore required to meet the criterion $\frac{\lambda_2}{\lambda_3} < \rho$ in order to be considered *well-constrained*, where ρ is a user–defined tunable threshold. Other criteria based on the eigenvalues may be used here instead, such as the degeneracy factor, inverse maximum covariance eigenvalue, or inverse condition number [19, 2]. All of these criteria perform the same basic function, and we do not compare the performance of these criteria in this work. If the criterion is not met, the landmark is classified as *under-constrained*. Note that this optimization is never fully solved - we only examine the eigendecomposition of $\mathbf{A}^T\mathbf{A}$ at the linearization point.

B. Method 1: Non-parametric Representation

The first method we propose is to remove under-constrained landmarks from the state vector entirely, so that their position is not explicitly modeled in the optimization. The measurements corresponding to the landmark l_j are collected into one combined non-parametric factor, f_j, as shown in Figure 5a. This factor treats the landmark's first bearing-range measurement m_b, taken from its base pose x_b, as constant, fixing two of the landmark's spherical coordinates. At every iteration in the optimization, the factor performs a search in the feasible elevation range by sampling elevation angles at uniform increments, and selects the elevation angle with the lowest total reprojection error as the current estimate:

$$\begin{aligned}\theta^* = &\underset{\theta \in \Theta}{\operatorname{argmin}}\\ &\sum_k \left\| \pi\left(x_k, \pi^{-1}\left(x_b, m_b, \theta\right)\right) - m_k \right\|_{\mathbf{\Sigma}_k}^2 .\end{aligned} \tag{11}$$

Where $\Theta = \{\theta_{min}, \theta_{min} + \Delta\theta, \ldots, \theta_{max} - \Delta\theta, \theta_{max}\}$. The reprojection error is computed as the Mahalanobis distance between the projection of the 3D point into the frame of pose x_k and the measurement m_k, using the measurement uncertainty Σ_k. The cost function for this factor is then the total reprojection error evaluated at the optimal elevation angle:

$$h_i\left(X\right) - z_i = \sum_k \left\| \pi\left(x_k, \pi^{-1}\left(x_b, m_b, \theta^*\right)\right) - m_k \right\|_{\mathbf{\Sigma}_k}^2 . \tag{12}$$

The nonlinear optimization using these non-parametric factors is solved using Levenberg-Marquardt, as in the original ASFM algorithm.

This formulation is advantageous because it disassociates the under-constrained landmarks' elevation angles from the nonlinear optimization's Gaussian model — it essentially treats the belief of the elevation angle as a *uniform distribu-*

tion. This has the benefit of being able to freely update the elevation estimate to the best elevation angle at any iteration in the optimization, in addition to preventing the optimization from estimating the elevation angle to lie outside of the feasible range. At first glance, one main drawback of this formulation is that is takes a single bearing-range measurement as constant and does not refine the estimate. We next present an alternative formulation to address this concern.

C. Method 2: Semi-parametric Representation

The second method we propose is to only remove the under-constrained landmarks' elevation angle from the state vector. A under-constrained landmark is then explicitly modeled as a two-dimensional bearing-range point in the factor graph, as shown in Figure 5b. As in Method 1, all of the measurements of under-constrained landmark l_j are combined into a single joint measurement factor s_j. This joint measurement factor is identical to the non-parametric factor in Method 1, except that it uses the landmark's explicitly modeled bearing and range estimate in computing reprojection error, rather than the base pose's measurement. In this framework, the bearing and range of the landmark are explicitly modeled as normally-distributed and are able to be updated by the nonlinear optimization algorithm.

VI. Results and Discussion

A. Simulation Results

The proposed methods are evaluated quantitatively in simulation. A sequence of 50 sonar poses is generated with landmarks uniformly distributed throughout the environment near the sensor. In each frame, artificial measurements of point landmarks that fall within the sonar's viewable frustum are generated and taken as inputs to the ASFM system. We use a $28.8°$ bearing field of view, $28°$ elevation field of view, and a range of $1-3$m, which are comparable to the operating characteristics of the sensor in our real-world experiments. Varying levels of isotropic Gaussian noise are added to the odometry measurements and the simulation is repeated for 200 independent trials for each level of odometry noise. A constant level of Gaussian noise is also added to the bearing-range measurements ($\sigma_\psi = 1°$, $\sigma_r = 0.01m$), which are consistent with the capabilities of the sonar used in our real-world experiments. Two different types of trajectories are tested: pure roll rotation and pure y translation, using increments of 0.1 radians and 0.1 m between each timestep, respectively. Roll rotation is a motion type that constrains the point landmarks well, while pure y translation does not constrain the landmarks well [6, 18]. In these simulations we use an elevation discretization of $\Delta\theta = \frac{\theta_{max}-\theta_{min}}{60}$ and $\rho = 20$.

We evaluate the absolute trajectory error (ATE, as defined in [16]) and average landmark error (ALE) for the original ASFM framework and our two proposed frameworks. The ATE aligns the estimated trajectory with the ground truth trajectory, and computes the average translation error between corresponding poses - it is a measure of the error accumulated along the entire trajectory. The ALE simply computes the average Euclidean distance between all well-constrained reconstructed landmarks and their ground truth locations. Our implementations of all frameworks use the GTSAM library [4] and the Levenberg-Marquardt algorithm for optimization.

Dataset	DR	ASFM	Method 1	Method 2
Stationary 1	1.1	1.2	0.67	**0.65**
Stationary 2	**0.9**	1.4	1.5	1.1
Stationary 3	1.0	1.3	0.60	**0.57**
Stationary 4	0.9	4.1	1.3	**0.8**
y-axis 1	**4.3**	7.2	**4.3**	**4.3**
y-axis 2	3.1	27.0	3.2	**2.8**
y-axis 3	3.3	6.0	**3.2**	3.3
y-axis 4	3.2	4.4	**2.8**	3.0

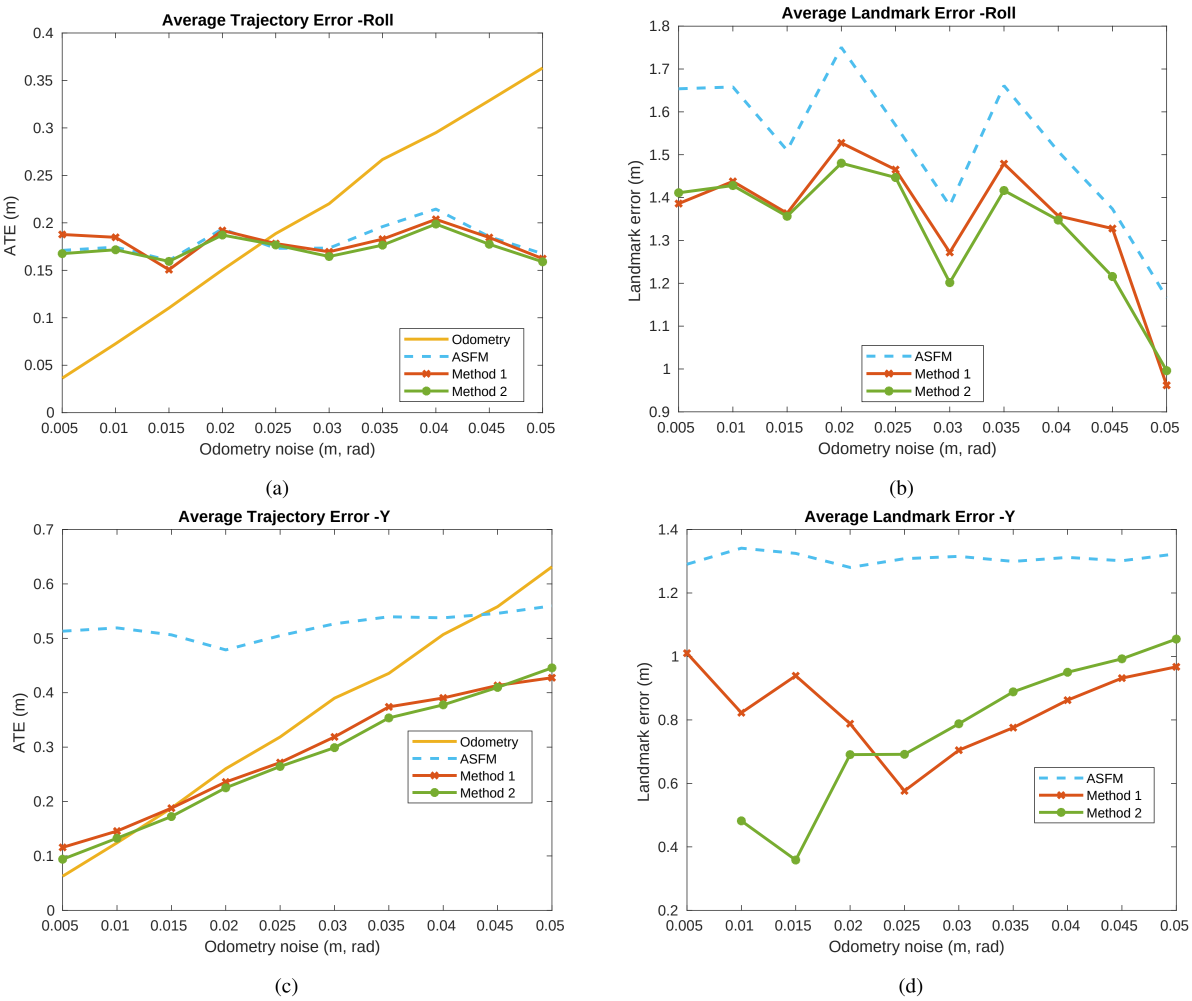
TABLE I: Absolute trajectory error (ATE) in cm of the proposed methods on test tank datasets. The lowest ATE on each dataset is bolded.

Figure 6 shows the results of the simulations. Under pure roll rotation, the proposed methods result in similar errors in localization (ATE) as the original ASFM algorithm since this motion provides good constraints on the landmarks' elevation angles. Likewise, the average landmark error improves slightly with the proposed approaches, as only landmarks that are sufficiently well-constrained are reconstructed. Under pure y translation, the original ASFM algorithm significantly increases localization error compared dead-reckoning due to the prevalence of under-constrained landmarks. The proposed methods actually improve on the dead-reckoning localization errors when there is significant noise present in the odometry measurements. Likewise, landmarks are reconstructed more accurately by the proposed methods as well, as only landmarks with sufficient constraints are explicitly reconstructed.

B. Experimental Results

The proposed methods are also evaluated experimentally on real-world data recorded in a test tank. A SoundMetrics DIDSON imaging sonar [15] was used in 1.8 MHz mode with a spreader lens to achieve $28°$ elevation field of view. The sonar is mounted on a Bluefin hovering autonomous underwater vehicle (HAUV), shown in Figure 7, along with a 1.2MHz Teledyne/RDI Workhorse Navigator Doppler Velocity Log (DVL) and a stereo camera (not used in these experiments). The sonar and DVL are fixed pointing down, toward the bottom of the tank. Measurements from the DVL and a tactical grade, ring-laser Honeywell HG1700 IMU are fused to estimate vehicle odometry in the DVL frame. While dead-reckoning with these odometry measurements will inevitably drift during long-term operation, the odometry estimates are highly accurate for short-term operation. Manually measured DVL-sonar extrinsics are used to obtain ground truth sonar poses for our evaluation.

A 4ft × 2ft × 2ft aluminum box was placed in the test tank with several stacks of small, round magnets (2.5 cm diameter) attached to one of the faces. The magnets on the vertical sides of the box are used as features for SLAM, as shown in Figure 1. We use minimum and maximum range



(a)

(b)

(c)

(d)

Fig. 6: Under pure roll rotation, (a) the absolute trajectory error (ATE) and (b) average landmark error (ALE). Subfigures (c) and (d) show the ATE and ALE for pure y translation, respectively. The noise in odometry measurements is isotropic, using units of meters for the translation directions and radians for the rotation directions.

thresholds to disregard spurious features detected on the top of the box or the bottom of the test tank. Note that while the features lie on a planar surface, we do not use any planar scene assumptions at any point in the proposed algorithms.

We artificially add noise to the ground-truth poses to generate a noisy odometry trajectory, which is used to simulate a dead-reckoning trajectory estimate using a lower quality IMU and no DVL. We evaluate the same methods as in our simulations: dead reckoning (DR) using the noisy odometry and the three SLAM frameworks. The three SLAM methods all incorporate the noisy odometry measurements in the optimization, as depicted in Figures 3, 5a, and 5b.

Precisely evaluating the accuracy of our underwater SLAM algorithms is difficult in a real-world experiment, due to the lack of a motion-capture system and uncertainty in the extrinsics relating the vehicle's odometry frame to the sonar frame. Because of this, we carefully select two types of trajectories in order to minimize the effect of the imprecise DVL-sonar extrinsics on ground-truth pose estimates: (1) no motion and (2) translation along the y-axis of the sonar. The stationary datasets average 25 sonar frames each, while the y-axis datasets average over 100 sonar frames. To highlight the advantage of our proposed landmark parameterizations, we consider all landmarks as under-constrained in these datasets. In these experiments, we use $D_1 = 10\,\text{pixels}$, $D_2 = 25\,\text{pixels}$, and $\Delta\theta = \frac{\theta_{max} - \theta_{min}}{60}$.

Table I shows the ATE for the four localization methods, evaluated on the eight test tank datasets (four stationary, four translation along the y-axis). The original ASFM algorithm *increases* the localization error compared to dead-reckoning in all of the datasets, due to the incorrect estimation of the under-constrained points. Both of the proposed formulations improve localization error compared to dead-reckoning and ASFM on almost all of the datasets as well, with the semi-parametric Method 2 generally outperforming the non-parametric Method 1. The proposed algorithms achieve this enhanced localization accuracy despite some sub-optimal feature correspondences (most of the features are duplicated and represented by multiple landmarks in the SLAM framework due to the conservative nature of the data association algorithm).

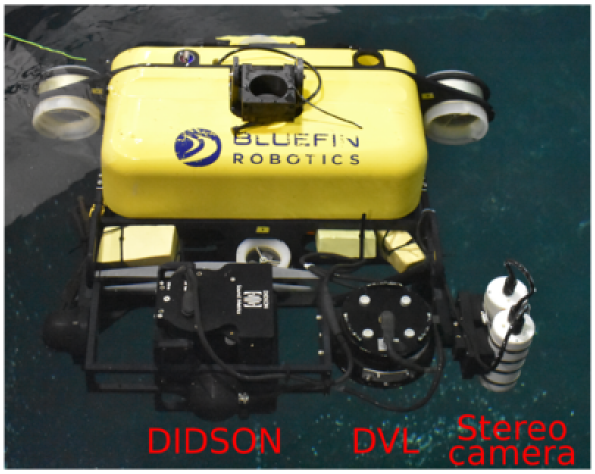


Fig. 7: Bluefin HAUV robot used to gather experimental data in a test tank.

VII. Conclusion

In this work, we have addressed several of the shortcomings of the basic ASFM algorithm and proposed a fully-automatic pipeline that enables real-time ASFM for underwater localization and mapping. Our method automatically extracts and corresponds features using a nonlinear scale space to reduce the effect of the speckle noise which is present in sonar images. We provide an algorithm for identifying well-constrained features whose 3D positions may be accurately estimated for mapping purposes. Finally, we describe two viable frameworks for integrating under-constrained features in order to constrain the sensor motion. We have demonstrated that while under-constrained features can be used to improve the sensor localization estimate, it is detrimental to explicitly model their elevation angle, and may also be detrimental to include the bearing and range of such measurements in the state vector of the optimization.

The ASFM algorithm would greatly benefit from features that are more viewpoint-invariant than the A-KAZE feature points used in this framework, although the nonlinear projection of the sensor presents great challenges for this task. The data association process could potentially be improved by implementing a JCBB framework, at the expense of greater computational cost, or by tracking features frame-to-frame. In future work, we hope to implement a more robust data association framework and evaluate our methods more rigorously in real-world experiments by using a more accurate ground-truth trajectory than the on-board odometry.

Acknowledgment

The authors would like to thank Tiffany Huang and Ming Hsiao for insightful discussions, as well as Chuck Whittaker for help preparing and carrying out the test tank experiments.

References

[1] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, “Fast explicit diffusion for accelerated features in nonlinear scale spaces,” in *British Machine Vision Conference, BMVC 2013, Bristol, UK, September 9-13, 2013*, 2013.

[2] E. W. Cheney and D. R. Kincaid, *Numerical Mathematics and Computing*, 6th ed. Pacific Grove, CA, USA: Brooks/Cole Publishing Co., 2007.

[3] F. Dellaert and M. Kaess, “Factor graphs for robot perception,” *Foundations and Trends in Robotics*, vol. 6, no. 1-2, pp. 1–139, Aug. 2017, http://dx.doi.org/10.1561/2300000043.

[4] F. Dellaert, “Factor graphs and GTSAM: A hands-on introduction,” GT RIM, Tech. Rep. GT-RIM-CP&R-2012-002, Sept 2012. [Online]. Available: https://research.cc.gatech.edu/borg/sites/edu.borg/files/downloads/gtsam.pdf

[5] 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.

[6] T. A. Huang and M. Kaess, “Towards acoustic structure from motion for imaging sonar,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2015, pp. 758–765.

[7] T. Huang and M. Kaess, “Incremental data association for acoustic structure from motion,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, IROS*, Daejeon, Korea, Oct. 2016, pp. 1334–1341.

[8] T. Huang, “Acoustic structure from motion,” Master’s thesis, Carnegie Mellon University, Pittsburgh, PA, May 2016.

[9] Y. Ji, S. Kwak, A. Yamashita, and H. Asama, “Acoustic camera-based 3d measurement of underwater objects through automated extraction and association of feature points,” in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Sept 2016, pp. 224–230.

[10] 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, pp. 4396–4403.

[11] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, “iSAM2: Incremental smoothing and mapping using the Bayes tree,” *Intl. J. of Robotics Research, IJRR*, vol. 31, no. 2, pp. 216–235, Feb. 2012.

[12] J. Li, P. Ozog, J. Abernethy, R. M. Eustice, and M. Johnson-Roberson, “Utilizing high-dimensional features for real-time robotic applications: Reducing the curse of dimensionality for recursive bayesian estimation,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1230–1237.

[13] 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.

[14] Y. S. Shin, Y. Lee, H. T. Choi, and A. Kim, “Bundle adjustment from sonar images and SLAM application for seafloor mapping,” in *OCEANS 2015 - MTS/IEEE Washington*, Oct 2015, pp. 1–6.

[15] Sound Metrics Corporation, “SoundMetrics Didson 300 specifications,” http://www.soundmetrics.com/products/DIDSON-Sonars/DIDSON-300-m/.

[16] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of RGB-D SLAM systems,” in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012, pp. 573–580.

[17] 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.

[18] Y. Yang and G. Huang, “Acoustic-inertial underwater navigation,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2017, pp. 4927–4933.

[19] J. Zhang, M. Kaess, and S. Singh, “On degeneracy of optimization-based state estimation problems,” in *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Stockholm, Sweden, May 2016, pp. 809–816.