Wide Aperture Imaging Sonar Reconstruction using Generative Models

Eric Westman and Michael Kaess

***Abstract*— In this paper we propose a new framework for reconstructing underwater surfaces from wide aperture imaging sonar sequences. We demonstrate that when the leading object edge in each sonar image can be accurately triangulated in 3D, the remaining surface may be "filled in" using a generative sensor model. This process generates a full three-dimensional point cloud for each image in the sequence. We propose integrating these surface measurements into a cohesive global map using a truncated signed distance field (TSDF) to fuse the point clouds generated by each image. This allows for reconstructing surfaces with significantly fewer sonar images and viewpoints than previous methods. The proposed method is evaluated by reconstructing a mock-up piling structure and a real world underwater piling, in a test tank environment and in the field, respectively. Our surface reconstructions are quantitatively compared to ground-truth models and are shown to be more accurate than previous state-of-the-art algorithms.**

I. Introduction

The ability of acoustic waves to propagate through turbid waters makes sonar sensors the de facto option for exteroceptive underwater sensing in poor visibility conditions. Side-scan sonars have been used widely for many years on autonomous underwater vehicles (AUVs) to image the seafloor in order to perform large-scale localization [13], mapping [9], and object tracking [38]. A newer class of higher frequency sonars called *imaging* or *forward looking sonars* (FLS) have been developed for sensing on a smaller scale. Examples include the SoundMetrics ARIS [1] and DIDSON [2] sensors. Like side-scan sonars, they have been used in seafloor scenarios for image registration [3, 18, 21], mosaicing [14, 33], mapping [20, 22, 30, 31, 39, 44], and tracking [17]. However, unlike side-scan sonars, imaging sonars are not restricted to the configuration of pointing downward towards the seafloor. They have been mounted on AUVs in a variety of configurations that allow for inspection of more complex environments than seafloors [8, 19, 28].

The focus of this work is using imaging sonar for underwater mapping with known poses. Specifically, we aim to accurately reconstruct the surfaces of objects in underwater scenes. A growing body of work has emerged in which imaging sonars are used for this very purpose. The main difficulty these algorithms must overcome is the sonar's elevation ambiguity. Each pixel in a sonar image corresponds to a specific bearing (or azimuth) angle and range, but does not measure the elevation angle, similar to a monocular camera's range ambiguity. Most previous approaches do not attempt to resolve the elevation ambiguity, but rather assume that a pixel measurement applies to the entire volume lying along the elevation arc.

This work was partially supported by the Office of Naval Research under grants N00014-16-1-2103 and N00014-16-1-2365.

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

(a) (b)

Fig. 1: (a) The above-water portion of a pier piling, with the HAUV preparing to scan. (b) A mesh model of the 3D reconstruction generated by our proposed algorithm. The coordinate axes represent the vehicle and sonar poses.

In this work we take significant steps towards the ultimate goal of autonomous underwater mapping of arbitrary structures using multibeam imaging sonar. Specifically, we present:

- a general framework for reconstructing 3D objects from a single sonar scan, by explicitly estimating the missing elevation angle using a generative sensor model;
- a method of fusing surface measurements from multiple viewpoints to create a cohesive object model;
- experimental results demonstrating the ability of our proposed algorithm to accurately reconstruct underwater piling structures using a robotic AUV platform.

The remainder of this paper is organized as follows: Section II discusses previous approaches to the imaging sonar mapping problem and the advantages that our proposed method provides over them. Section III provides an introduction to imaging sonar sensing and defines the problem statement. Sections IV and V describe our proposed frontend image processing and backend surface reconstruction modules, respectively. We present results from our experiments in Section VI. Lastly, we summarize the contributions of this work in Section VII and discuss the next steps in this line of research in Section VIII.

II. Related Work

A variety of algorithms have been proposed in recent years to generate 3D surface models using underwater imaging sonar. Teixeira et al. utilize a concentrator lens on a DIDSON [2] sonar to narrow the sensor's elevation aperture from 14° to 1° and assume that each point has zero elevation [41]. In this case, the sonar is treated as a single-line scanner. While this technique has been used to generate large-scale reconstructions, we aim to increase the accuracy and reduce the time required to create such maps by utilizing much richer images generated with a wide elevation aperture.

The principle of space carving (SC) has been applied to generate dense 3D models from known poses [5, 7]. Rather than using high intensity pixel measurements that correspond to surface observations, space carving uses the low intensity pixel measurements that correspond to free space to "carve out" the unoccupied regions of the scene, ideally generating a surface model that is an outer bound of the actual object surfaces. Due to the fact that surfaces along a particular bearing angle may occlude other surfaces lying at the same angle but a longer range, this requires using the feasible object region mask (FORM image), rather than the raw polar coordinate sonar image. The FORM image segments the image into the region from the sonar's minimum range up until the leading object edge as "free space", and the region from the leading object edge to the maximum range as "possibly occupied". Aykin et al. produce the surface model by using the intersection of α-shapes corresponding to the leading object edges. This framework has proven effective at accurately reconstructing simple shapes when precisely known poses are available in a laboratory environment. However, the SC paradigm is incapable of reconstructing a wide variety of complex shapes and discards most of the information available from the sonar images.

A similar framework presents a slightly different implementation of space carving using voxel grids [15, 16]. Each pixel in the sonar image is projected into the volume along its elevation arc. Each voxel that it intersects tracks the minimum pixel value observed. An observation corresponding to a low intensity pixel carves that voxel out of the model. Occlusions are handled with a post-processing step, which attempts to generate a model consisting only of points on the object's surface, discarding points from the object's interior. This approach, called "min-filtering", suffers from the same limitations as the SC method of Aykin et al. – namely that generating an accurate model is highly dependent on having a multitude of good viewing angles.

Similar to min-filtering, voxel grids have been used to model the likelihood of occupancy under the occupancy grid mapping (OGM) framework [43, 42]. Projecting the pixels into the voxel grid, voxels are updated using an inverse sensor model, which encodes how a pixel intensity measurement corresponds to the likelihood of a voxel's occupancy. A threshold may be selected to classify the voxels as free or occupied based on their filtered values. Like space carving, this framework is dependent on having a variety of viewpoints which may not be possible to obtain. In fact, if the FORM image is used to account for possible occlusions, occupancy grid mapping can be seen as a generalization of the space carving framework.

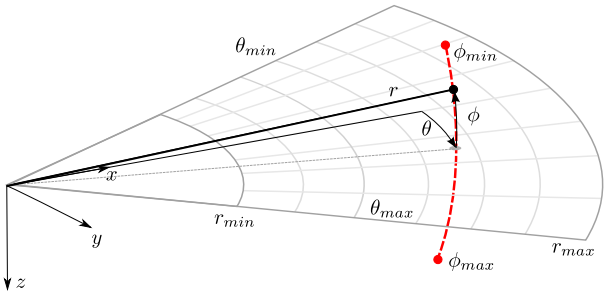


Fig. 2: Simple sonar geometric sensor model. A point at location (θ, r, ϕ) is projected along the red, dotted elevation arc into the zero elevation imaging plane. However, *all* surfaces lying along the elevation arc may reflect sound back towards the sensor and contribute to the intensity measured at the corresponding pixel.

Several more recent algorithms have been presented that attempt to infer directly from a generative sensor model. Guerneve et al. [16] propose a linear approximation to the nonlinear elevation aperture, which presents an efficient solution to solve for 3D occupancy using blind deconvolution with a spatially-varying kernel. However, this method requires precise motion from the sensor – pure translation along the z-axis. Additionally, the linear approximation holds well for sonars with a narrow elevation aperture, but the quality of the reconstruction degrades as the elevation aperture widens.

A body of work by Aykin et al. [4, 6] is, to the best of our knowledge, the only work that aims to directly estimate the elevation angle of each pixel in a sonar image. This method is constrained to the scenario of objects lying on the seafloor. Upon detecting the seafloor plane and manually extracting object and shadow edges, the bottom and top 3D contours of the object of interest may be easily computed. With these edges bounding the object's surface, the interior of the object is iteratively "filled-in" based on the generative sensor model and the actual sonar image intensities. This method has laid the groundwork for our proposed algorithm, in which we seek to apply generative model-based surface reconstruction to arbitrary scenes, not just seafloor mapping.

In this work we seek to develop a method that does not place restrictions on the sensor motion (e.g. pure z-axis translation only, no motion), environment (e.g. seafloor mapping only), or make linearization assumptions that break down for wide aperture sensors. Our proposed algorithm may be divided into two steps which are discussed in the following sections: frontend image processing and backend model-based surface reconstruction.

III. Problem Statement and Background

The problem we wish to solve in this work is as follows. Given a set of polar coordinate sonar images, the poses from which they were taken, and a generative sensor model,

produce a three-dimensional reconstruction of the imaged surfaces. In this section, we describe the fundamentals of the imaging sonar sensor and define the generative sensor model that we use in this work.

To precisely define the sonar sensor model, consider a 3D point in the frame of the sonar sensor:

$$\mathbf{p} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = r \begin{bmatrix} \cos\theta\cos\phi \\ \cos\theta\sin\phi \\ \sin\phi \end{bmatrix} \tag{1}$$

where (r, θ, ϕ) denote the spherical coordinates: range, bearing (azimuth), and elevation. As shown in Fig. 2, this point projects into the polar coordinate sonar image $I(r, \theta)$ at the discrete pixel bin that contains the real-valued range and bearing measurements:

$$r = \sqrt{X^2 + Y^2 + Z^2} \tag{2}$$

$$\theta = \operatorname{atan2}(Y, X). \tag{3}$$

This simple projection accurately models the *location* of a projected point or surface patch in the sonar image. A generative sensor model must also describe how the imaged surfaces produce the corresponding pixel *intensities*.

Ideally, the pixel intensity is influenced only by the interaction of the sound waves with the imaged surfaces, although in reality there are multiple confounding factors. Assuming isotropic sound emission by the sensor, this ideal model can be expressed generally as

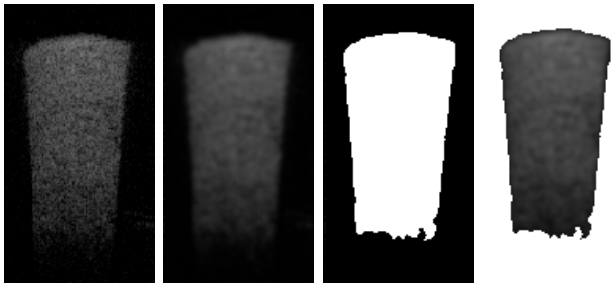
$$I(r, \theta) = \int_{\phi=\phi_{\min}}^{\phi=\phi_{\max}} \mathbf{1}(r, \theta, \phi)\, \Omega(r, \theta, \phi)\, d\phi \tag{4}$$

where $\mathbf{1}(r, \theta, \phi)$ is an indicator function denoting the existence of an imaged surface at the 3D location and $\Omega(r, \theta, \phi)$ encodes how the sound is reflected by the surface and propagated back to the sonar [16]. Note that this model disregards multipath returns, in which the sound reflects off of multiple surfaces before returning to the sensor.

While a variety of reflection models have been used that consider specular and / or diffuse scattering, the specular component often may appear to be negligible due to the generally rough surfaces of underwater objects and the grazing incident angles often used with sonar sensors [6, 25, 27]. In this work we adopt a simple diffuse reflection model for all imaged surfaces, assuming each pixel images a single surface patch:

$$I(r, \theta) = k \cos^m(\alpha) \tag{5}$$

where k is a normalization constant, $1 \leq m \leq 2$, and α is the angle of incidence between the incoming acoustic beam and the surface normal of the patch. We assume that a time / range varying gain (TVG / RVG) has been applied to the raw image to correct for the spatial spreading of the sound waves. It is important to note that our proposed algorithm may utilize *any* reflection model, not just the particular one we have selected for our experiments.



(a) (b) (c) (d)

Fig. 3: The stages of our frontend image processing pipeline, demonstrated on our test tank piling dataset. (a) The raw polar coordinate sonar image, (b) denoising using anisotropic diffusion, (c) the surface segmentation using MSER and (d) the binary surface mask applied to the denoised image.

IV. Frontend - Image processing

The frontend of our system operates on each input sonar image individually. The two goals of this module are: (1) to denoise the sonar image and (2) to identify the pixels that correspond to surface measurements. Upon completing these steps, the denoised sonar image and the binary image mask corresponding to object surfaces may be passed to the backend for surface reconstruction.

A. Denoising

Sonar images suffer from significantly higher speckle noise than optical images. Previous attempts to denoise sonar images include averaging multiple images taken from the same viewpoint [32]. Since our algorithm is targeted for robotic underwater mapping, in which the vehicle and sensor poses may not be precisely known or controlled, we seek a method of denoising each image individually. To this end, we adopt the procedure of anisotropic diffusion [37]. This step blurs the image as in a standard Gaussian blurring process, but preserves distinct edges in the image by scaling the diffusion in proportion to the inverse of the image gradient. This has been previously used with success as a denoising step before detecting point features for sparse SLAM systems [39, 44]. An example of the denoising process applied to a sonar image of a mock-up piling (a rectangular prism shape) is shown in Figs. 3a and 3b.

B. Surface segmentation

Convolutional neural networks (CNNs) have rapidly become the de facto approach to image segmentation in the field of computer vision [26]. Their emergence has been made possible in part due to very large amounts of training data available. Recent years have seen CNNs successfully applied to sonar images for various tasks, including crosstalk removal [40], object detection [23], and global context perception [11, 12]. However, collecting a sufficient quantity of sonar images for training is a significant challenge to the application of these methods for underwater sonar perception. While we perceive the future of surface segmentation to lie in the field of machine learning, we leave this approach to future work.

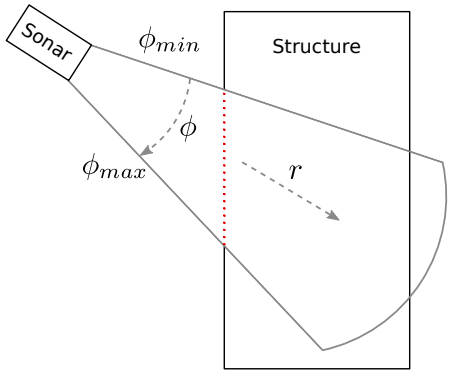


Fig. 4: A side view of a sonar imaging a piling. A 2D cross-section of the viewable frustum is depicted, which corresponds to a single bearing angle. The imaged area of the piling is shown in dotted red. This depicts how the elevation angle of the imaged surface increases or decreases monotonically with increasing range.

We take a simpler approach to surface segmentation by finding maximally stable extremal regions (MSER) [29] on the denoised image. This is a blob detection algorithm that we use to find large connected components with gradual changes in pixel intensity. Each segmented component corresponds to distinct, continuous surface imaged by the sensor. An example of the MSER algorithm applied to a denoised sonar image is shown in Fig. 3b - 3d. We denote the resulting binary image surface mask as $M(r, \theta)$, where $M(r_i, \theta_j) = 1$ denotes a detected surface.

V. Backend - Surface Reconstruction

In order to generate a 3D reconstruction from a sonar image and corresponding surface mask, several assumptions must be made. We assume that the sonar and scene geometry are configured such that each pixel (elevation arc) images a single surface patch. For simply shaped objects, this assumption holds true when the sonar is positioned at a grazing angle. We also assume that for a continuous surface, the elevation angle along a particular bearing angle either increases or decreases monotonically as the range increases. A violation of this assumption would cause a self-occlusion, and the corresponding pixels would presumably not be classified as surface pixels by the frontend of our algorithm.

Our approach is inspired by [4], which uses the 3D locations of the leading and trailing object edges as initialization to iteratively refine the 3D object reconstruction and update the generative model normalization parameter. However, if the full generative model is known a priori, a 3D reconstruction can be obtained using just one object edge as initialization.

A. Edge initialization

In this work, we focus our experiments on reconstructing underwater piling structures, which are long columns that support structures such as bridges or piers. We take advantage of the fact that a piling spans the entire elevation field of view of the sonar sensor, which is depicted in Fig. 4. As long as the sonar is tilted at least $\phi_{\max}$ degrees from perpendicular to the piling, each pixel's elevation arc will image only one surface patch. Furthermore, the closest detected surface patch in each image column (discrete bearing angle bin), may be easily determined to lie at elevation ϕ_{min}. The same principle may be applied to determine that the 3D position of the trailing edge of the surface is at $\phi_{\max}$. However, for larger tilt angles, the structure may not span the bottom edge of the elevation frustum. For the purposes of this work, we utilize the less restrictive single edge initialization, which may be applied to a variety of settings apart from piling inspection.

B. Dense 3D reconstruction

With the leading object edge located at ϕ_{min}, the remainder of the surface points in the image can be filled in using constraints from the generative model. We follow the general procedure described by Aykin et al. [4] but can reconstruct the points in a single pass through the image, without iteratively updating the normalization parameter.

The single pass through the image $I(r, \theta)$ is performed row-by-row, beginning with the first row r_0. We use the shorthand $I_{i,j} \coloneqq I(r_i, \theta_j)$ and $M_{i,j} \coloneqq M(r_i, \theta_j)$, and use $\mathbf{p}_{i,j}$ to denote the 3D point in the sensor frame corresponding to the pixel at $I_{i,j}$. Pixels in row r_i are stepped through column-by-column. If $M_{i,j}$, $M_{i+1,j}$, and either of $M_{i,j-1}$ or $M_{i,j+1}$ are identified as surface pixels, then we can use constraints from the generative sensor model to approximately compute the elevation angle of point $I_{i+1,j}$.

Assuming the elevation angle (and therefore 3D location) of $\mathbf{p}_{i+1,j}$ is known, we can compute the surface normal of the patch at $I_{i,j}$ using the cross product of the neighboring 3D points:

$$\mathbf{v}_{ij} = \mathbf{d}_{ij} \times \mathbf{e}_{ij} \tag{6}$$

$$\hat{\mathbf{n}}_{ij} = \frac{\mathbf{v}_{ij}}{\|\mathbf{v}_{ij}\|_2}. \tag{7}$$

Here, $\mathbf{d}_{ij} = \mathbf{p}_{i+1,j} - \mathbf{p}_{i,j}$ and $\mathbf{e}_{ij} = \mathbf{p}_{i,j-1} - \mathbf{p}_{i,j}$ or $\mathbf{e}_{ij} = \mathbf{p}_{i,j+1} - \mathbf{p}_{i,j}$, depending on which pixel in neighboring columns is identified as a surface pixel. Then using the vector corresponding to the ray of incident sound from the sensor $\hat{\mathbf{p}}_{ij} = \mathbf{p}_{ij} / \|\mathbf{p}_{ij}\|_2$, we can compute the angle of incidence as:

$$\alpha = \operatorname{acos}\left(|\hat{\mathbf{n}}_{ij} \cdot \hat{\mathbf{p}}_{ij}|\right). \tag{8}$$

We can then use the generative model in Equation 5 to compute the model-predicted image intensity for the given elevation angle.

We perform a search of discrete elevation angles taken at uniform intervals from the range of feasible elevation angles: $[\phi_{i,j}, \min(\phi_{i,j} + \Delta\phi_{\max}, \phi_{\max})]$, where the maximum change in elevation angle from pixel to pixel $\Delta\phi_{\max}$ may be manually tuned. We set $\phi_{i+1,j}$ to the elevation angle with the smallest absolute error between the actual image measurement and model-predicted intensity. If there are not sufficient neighboring surface pixels to solve for $\mathbf{p}_{i+1,j}$, we assume that $\phi_{i+1,j} = \phi_{i,j}$. This procedure proves to work quite well for continuous surfaces, but may fail for images with more complex, disjointly segmented shapes.

	AADE (m)			RMSE (m)		
Dataset	SC	OGM	Ours	SC	OGM	Ours
Tank piling	0.033	0.038	**0.0176**	0.035	0.047	**0.022**
Field piling	0.136	0.152	**0.039**	0.168	0.207	**0.047**

TABLE I: Quantitative evaluation of three-dimensional object reconstructions from the test tank experiment. The two metrics we use are average absolute distance error (AADE) and root mean square error (RMSE). Our surface reconstruction method results in considerably more accurate surface models than the baseline methods, according to these two metrics.

If the trailing object edge is known as well as the leading edge, then the parameters of the generative model k and m may be iteratively refined until the trailing object edge determined by integrating the generative model aligns with the known trailing edge, as in [4]. We leave this to future work, however, as the trailing edges of the pilings in our experiments are difficult to consistently localize.

C. TSDF integration

Given the high levels of noise in the sonar image that remain after denoising and various unmodeled effects, the 3D point cloud generated by a single image may be quite noisy and inaccurate, even for simple structures such as pilings.

A truncated signed distance field (TSDF) is a volumetric map representation that has been used to generate high quality surface reconstructions from multiple noisy 3D scans generated by RGB-D cameras [35] and laser scanners [10]. Since point measurements typically correspond to rays of light or a laser beam, voxels are updated by stepping along the ray from the sensor to the point measurement. Each voxel tracks a value that is updated with a weighted, signed distance of the voxel from the surface along the line of sight. The zero crossings denote the estimated surface and a point cloud or triangle mesh may be generated from the TSDF.

While the TSDF is a quite intuitive choice of map representation for RGB-D sensors, in which each pixel corresponds to a ray-based surface observation, it is not so obvious a choice for the imaging sonar, where pixels correspond to elevation arcs. However, it is a good fit for our framework since we generate dense 3D surface measurements for each pixel. Furthermore, each surface measurement is made by acoustic waves propagating along the ray between the sensor and surface patch. This allows us to use the standard TSDF ray casting updates to fuse multiple surface measurements into a more accurate global model.

VI. Experimental Results

To evaluate the proposed system, we quantitatively and qualitatively compare our 3D reconstructions from real world test tank and field datasets to the 3D reconstructions resulting from two baseline methods: SC and OGM. SC and OGM are considered the leading state-of-the-art algorithms for real-time 3D imaging sonar reconstruction. We compare our proposed method to our own implementations of SC and OGM, which use fixed-size voxel grids for mapping. Our implementation of SC actually uses an OGM-like tracking of the probability of occupancy, rather than min-filtering. This allows for using a tunable threshold to acquire object points and to make the algorithm more robust to errors in the sensor pose and FORM image segmentation. For both baseline methods, the threshold to distinguish occupied from free voxels was tuned to generate the best reconstruction. Our proposed framework uses Voxblox [36] which implements spatially-hashed voxels [24, 35] for memory-efficient TSDF integration. For our proposed reconstruction method, we discard surface measurements from the outer 20% of columns on either side of the image, as the image intensity does not adhere to the generative model well due to anisotropic emission by the sonar.

The imaged targets in these experiments are a mock-up and a real-world piling. While these objects consist of approximately planar segments, our proposed method does not make any planarity assumptions.

Both the test tank and field datasets were recorded using a SoundMetrics DIDSON imaging sonar [2] mounted on a Bluefin Hovering Autonomous Underwater Vehicle (HAUV). Actuators allow us to tilt the DIDSON through a 90° range of motion. A spreader lens is used to increase the elevation aperture $\phi_{\max} - \phi_{\min}$ from 14° to 28°. Poses are acquired from the proprietary vehicle navigation and the actuator encoders. The vehicle odometry is highly accurate for short-term localization but inevitably drifts over time. For this reason, we only image two faces of each piling – the drift that accumulates from circumnavigating the entire piling is too great for mapping with known poses.

A. Test tank experiments

We imaged a mock-up piling of dimensions approximately 0.61 m x 0.61 m x1.83 m. We image two faces of the piling, tilting the sonar between approximately 20° and 50° from the horizontal, which allows the sonar to cover the entirety of the piling, except the small portion that passes through the top of the viewing frustum. Voxel grids for all reconstruction methods use a voxel size of 2.5 cm, including the TSDF, to produce reconstructions in real time at 5-10 frames per second. The generative model parameters $k = 0.37$ and $m = 1$ were used to model the acoustic reflection properties of the mock-up piling. Upon generating the 3D reconstruction, the surface points from each model are extracted and aligned to a ground truth model with ICP. The ground truth model was obtained using a FARO Focus3D survey laser scanner.

Fig. 5 shows top-down and isometric views for the three evaluated reconstruction methods. The point clouds are colored according to the point-to-plane error evaluated during ICP alignment, with the same color scale used across all three models. The top-down views show how SC and OGM fail to "carve out" space in front of each piling face. This causes the reconstructed surface to bulge out towards the bottom of the piling. On the other hand, our proposed method fuses the estimated 3D point clouds from each input image to generate a rather accurate estimate of the surface. While some inaccuracies in the resulting surface exist, there is no prominent bulge or spreading of the surface towards the bottom of the piling - both faces of the reconstructed piling

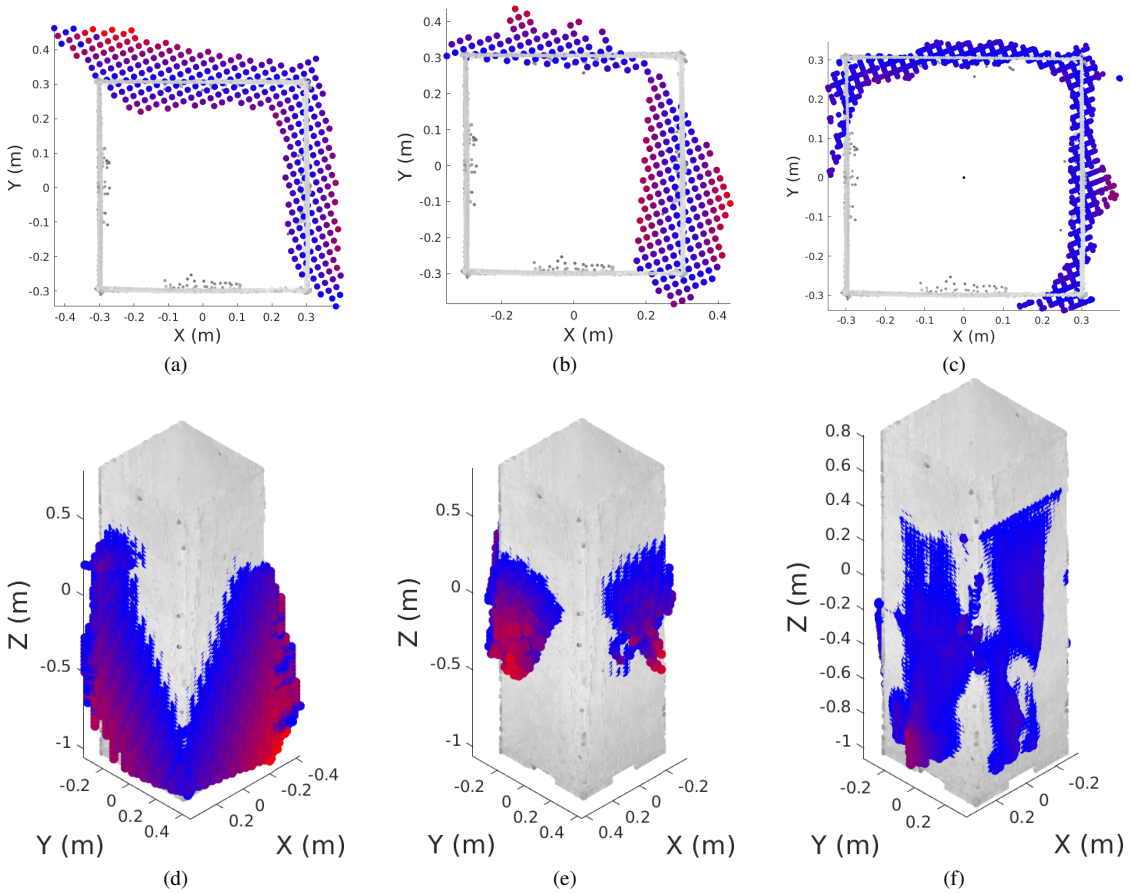


Fig. 5: 3D reconstructions of the mock-up piling from our test tank experiment. The gray cloud depicts the ground-truth model generated by a survey laser scanner. Colored points are the sonar reconstruction, with blue encoding low point-to-plane alignment error and red encoding high error. (a) - (c) Top-down views of the reconstructed point clouds of the **SC**, **OGM**, and **Proposed** algorithms, respectively, compared to the ground truth model. (d) - (f) Isometric views of the same reconstructions.

are quite close to vertical and planar.

Furthermore, we quantitatively evaluate the error of the resulting models using average absolute distance error (AADE) and root mean square error (RMSE) of the point-to-plane error metric. Table I shows that our method significantly increases the accuracy of the surface estimate compared to SC and OGM.

B. Field experiments

As the ultimate goal of this work is to enable robotic mapping and inspection of real world environments, we conducted field tests to reconstruct a pier piling in a harbor environment. Since the piling is larger than the one used in our test tank, voxel grids for all algorithms use a voxel size of 10 cm to maintain real-time reconstruction. The generative model parameters $k = 0.28$ and $m = 2$ were used to model the acoustic reflection properties of the piling. A photo of the piling and the mesh reconstruction generated by our algorithm are shown in Fig. 1. As a ground-truth model is not available for such a piling, we manually measured the width of the piling underwater (69 cm), and assume a purely rectangular prism shape. The shape of the piling is somewhat distorted by biofouling, as is visible in the photo, but the rectangular prism model remains a rather accurate estimate. Similar to the tank piling, we imaged two faces of the piling, as the vehicle state estimate drifted too much for a full circumnavigation.

Fig. 6 shows the same top-down and isometric views as for the tank piling dataset. SC and OGM clearly cannot accurately reconstruct the piling surface below a depth of 2.5 m, while our algorithm reconstructs a rather planar surface all the way to a depth of 5 m. Table I demonstrates the quantitative improvement in our reconstruction's accuracy, as evaluated against the ideal piling model.

We note that while SC and OGM may theoretically be able to generate a surface estimate of these simple structures with accuracy comparable to our method, this would require obtaining a much wider variety of viewpoints. For real-world experiments, this would mean longer mission times and potentially higher state estimate uncertainty.

VII. Conclusion

In this paper we have presented an algorithm for mapping with known poses using imaging sonar and a generative sensor model. Using very general prior information about the environment, the 3D location of the leading object edge may be accurately determined. Using this edge as initialization, the generative model may be used to fill-in the rest of the

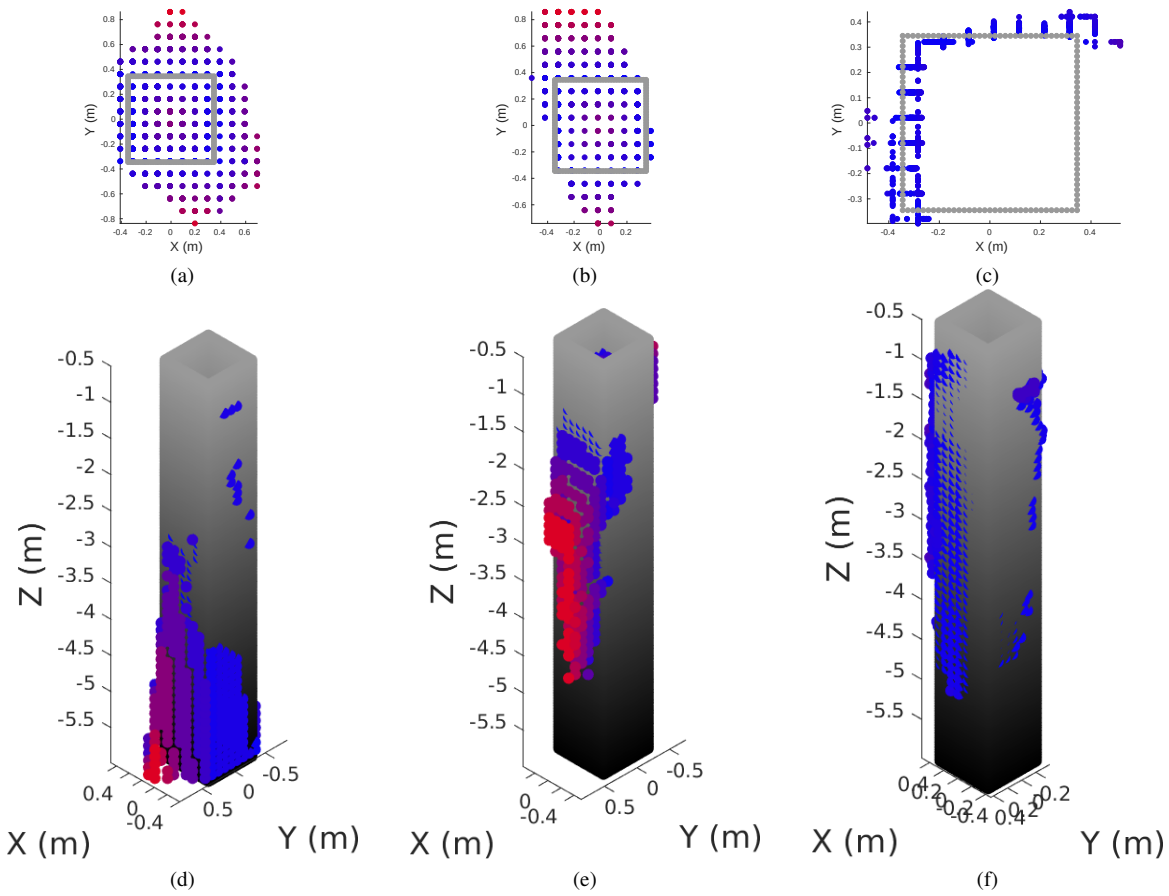


Fig. 6: 3D reconstructions of the real-world piling in the field. The gray-scale cloud depicts the ideal model according to our measurements of the piling. Colored points are the sonar reconstruction, with blue denoting low point-to-plane alignment error and red denoting high error. (a) - (c) Top-down views of the reconstructed point clouds of the **SC**, **OGM**, and **Proposed** algorithms, respectively, compared to the ground truth model. (d) - (f) Isometric views of the same reconstructions.

object surface. Using known sensor poses, the point clouds resulting from each input image are fused in a global model using a TSDF to smooth the surface estimate. We have demonstrated experimentally that our proposed method can outperform the existing state-of-the-art algorithms in terms of accuracy and that it requires fewer viewpoints and images to generate a surface model.

VIII. Future Work

This line of work may be extended to increase both the *accuracy* of the resulting surface maps and the *complexity* of surfaces that are capable of being reconstructed. Our proposed method assumes that the leading object edge in the sonar image may be triangulated accurately, but it remains unclear how to accurately localize such edges when the object does not extend through the sonar's elevation field of view. Additionally, our method relies upon the sonar images strictly adhering to the provided generative sensor model, when there are multiple factors that cause real-world sonar images to deviate from the theoretical model. Future work ought to investigate procedures for accurately calibrating the sonar sensor to characterize these effects and produce a more accurate generative sensor model. This future work also ought to consider non-diffusely reflecting materials and structures.

Of particular interest is the general framework of *optimizing* the surface model (and possibly sensor poses) using a generative model and multiple image measurements rather than *fusing* the surface models from each individual frame. This approach has been previously investigated but using only a single, simulated image to optimize a rough initial surface estimate [34]. Furthermore, the procedure assumes that a sufficiently accurate initial surface estimate is provided using space carving. However, serious doubts remain regarding the effectiveness of space carving in generating an accurate initial surface model for real-world applications. Other initialization methods, such as utilizing motion cues or a search of the feasible space of solutions, should be investigated in future work. A generalized edge initialization procedure would also enhance the versatility of our proposed algorithm.

IX. Acknowledgments

We would like to acknowledge P. Sodhi and M. Hsiao for insightful discussions as well as A. Hinduja, S. Suresh, T. Angert, and C. Whittaker for help devising and carrying out the experiments.

REFERENCES

[1] "SoundMetrics ARIS," http://www.soundmetrics.com/Products/ARIS-Sonars.

[2] "SoundMetrics DIDSON 300," http://www.soundmetrics.com/Products/DIDSON-Sonars/DIDSON-300m.

[3] M. D. Aykin and S. Negahdaripour, "On feature extraction and region matching for forward scan sonar imaging," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2012, pp. 1–9.

[4] ——, "Forward-look 2-D sonar image formation and 3-D reconstruction," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2013, pp. 1–10.

[5] ——, "On 3-D target reconstruction from multiple 2-D forward-scan sonar views," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, May 2015, pp. 1949–1958.

[6] ——, "Modeling 2-D lens-based forward-scan sonar imagery for targets with diffuse reflectance," *Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 569–582, 2016.

[7] ——, "Three-dimensional target reconstruction from multiple 2-d forward-scan sonar views by space carving," *Journal of Oceanic Engineering*, vol. 42, no. 3, pp. 574–589, 2016.

[8] H. Cho, B. Kim, and S.-C. Yu, "Auv-based underwater 3-D point cloud generation using acoustic lens-based multibeam sonar," *Journal of Oceanic Engineering*, 2017.

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

[10] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," 1996.

[11] R. DeBortoli, A. Nicolai, F. Li, and G. A. Hollinger, "Assessing perception quality in sonar images using global context," in *Proc. IEEE Conference on Intelligent Robots and Systems Workshop on Introspective Methods for Reliable Autonomy*, 2017.

[12] ——, "Real-time underwater 3D reconstruction using global context and active labeling," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 1–8.

[13] M. Fallon, M. Kaess, H. Johannsson, and J. Leonard, "Efficient AUV navigation fusing acoustic ranging and side-scan sonar," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, May 2011, pp. 2398–2405.

[14] F. Ferreira, V. Djapic, and M. Caccia, "Real-time mosaicing of large scale areas with forward looking sonar," *IFAC-PapersOnLine*, vol. 48, no. 2, pp. 32–37, 2015.

[15] T. Guerneve and Y. Petillot, "Underwater 3d reconstruction using blueview imaging sonar," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2015, pp. 1–7.

[16] T. Guerneve, K. Subr, and Y. Petillot, "Three-dimensional reconstruction of underwater objects using wide-aperture imaging sonar," *J. of Field Robotics*, 2018.

[17] N. O. Handegard and K. Williams, "Automated tracking of fish in trawls using the DIDSON (Dual frequency IDentification SONar)," *ICES Journal of Marine Science*, vol. 65, no. 4, pp. 636–644, 2008.

[18] B. T. Henson and Y. V. Zakharov, "Attitude-trajectory estimation for forward-looking multibeam sonar based on acoustic image registration," *Journal of Oceanic Engineering*, no. 99, pp. 1–14, 2018.

[19] 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.

[20] 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.

[21] N. Hurtós, D. Ribas, X. Cufí, Y. Petillot, and J. Salvi, "Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments," *J. of Field Robotics*, vol. 32, no. 1, pp. 123–151, 2014.

[22] B. Kim, H. Cho, H. Joe, and S.-C. Yu, "Optimal strategy for seabed 3d mapping of auv based on imaging sonar," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2018, pp. 1–5.

[23] J. Kim and S.-C. Yu, "Convolutional neural network-based real-time ROV detection using forward-looking sonar image," in *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*, 2016, pp. 396–400.

[24] M. Klingensmith, I. Dryanovski, S. Srinivasa, and J. Xiao, "Chisel: Real time large scale 3D reconstruction onboard a mobile device using spatially hashed signed distance fields," in *Robotics: Science and Systems (RSS)*, vol. 4, 2015, p. 1.

[25] G. Lamarche, X. Lurton, A.-L. Verdier, and J.-M. Augustin, "Quantitative characterisation of seafloor substrate and bedforms using advanced processing of multibeam backscatter - application to cook strait new zealand," *Continental Shelf Research*, vol. 31, no. 2, pp. S93–S109, 2011.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[27] N. T. Mai, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and H. Asama, "Acoustic image simulator based on active sonar model in underwater environment," in *2018 15th International Conference on Ubiquitous Robots (UR)*, 2018, pp. 775–780.

[28] G. Marani, S. K. Choi, and J. Yuh, "Underwater autonomous manipulation for intervention missions AUVs," *Ocean Engineering*, vol. 36, no. 1, pp. 15–23, 2009.

[29] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[30] S. Negahdaripour, "Application of forward-scan sonar stereo for 3-D scene reconstruction," *Journal of Oceanic Engineering*, 2018.

[31] ——, "Analyzing epipolar geometry of 2-d forward-scan sonar stereo for matching and 3-d reconstruction," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2018, pp. 1–10.

[32] S. Negahdaripour, P. Firoozfam, and P. Sabzmeydani, "On processing and registration of forward-scan acoustic video imagery," in *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, 2005, pp. 452–459.

[33] S. Negahdaripour, M. D. Aykin, and S. Sinnarajah, "Dynamic scene analysis and mosaicing of benthic habitats by FS sonar imaging-issues and complexities," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2011, pp. 1–7.

[34] S. Negahdaripour, V. M. Milenkovic, N. Salarieh, and M. Mirzargar, "Refining 3-D object models constructed from multiple FS sonar images by space carving," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2017, pp. 1–9.

[35] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3D reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 6, p. 169, 2013.

[36] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3D euclidean signed distance fields for on-board MAV planning," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 1366–1373.

[37] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

[38] Y. Petillot, S. Reed, and J. Bell, "Real time AUV pipeline detection and tracking using side scan sonar and multi-beam echo-sounder," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2002, pp. 217–222.

[39] Y. S. Shin, Y. Lee, H. T. Choi, and A. Kim, "Bundle adjustment from sonar images and SLAM application for seafloor mapping," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, Oct. 2015, pp. 1–6.

[40] M. Sung, H. Cho, H. Joe, B. Kim, and S.-C. Yu, "Crosstalk noise detection and removal in multi-beam sonar images using convolutional neural network," in *Proc. of the IEEE/MTS OCEANS Conf. and Exhibition*, 2018, pp. 1–6.

[41] P. Teixeira, M. Kaess, F. Hover, and J. Leonard, "Underwater inspection using sonar-based volumetric submaps," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, Oct. 2016, pp. 4288–4295.

[42] Y. Wang, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and H. Asama, "Three-dimensional underwater environment reconstruction with graph optimization using acoustic camera."

[43] Y. Wang, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and A. Hajime, "3D occupancy mapping framework based on acoustic camera in underwater environment," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 324–330, 2018.

[44] E. Westman, A. Hinduja, and M. Kaess, "Feature-based SLAM for imaging sonar with under-constrained landmarks," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2018.