2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Fusing Concurrent Orthogonal Wide-aperture Sonar Images for Dense Underwater 3D Reconstruction

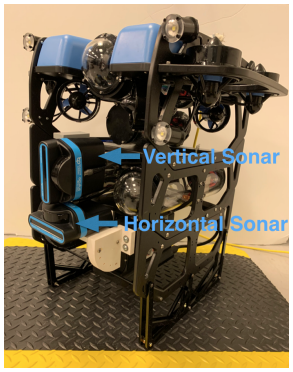
John McConnell, John D. Martin and Brendan Englot

***Abstract*— We propose a novel approach to handling the ambiguity in elevation angle associated with the observations of a forward looking multi-beam imaging sonar, and the challenges it poses for performing an accurate 3D reconstruction. We utilize a pair of sonars with orthogonal axes of uncertainty to independently observe the same points in the environment from two different perspectives, and associate these observations. Using these concurrent observations, we can create a dense, fully defined point cloud at every time-step to aid in reconstructing the 3D geometry of underwater scenes. We will evaluate our method in the context of the current state of the art, for which strong assumptions on object geometry limit applicability to generalized 3D scenes. We will discuss results from laboratory tests that quantitatively benchmark our algorithm's reconstruction capabilities, and results from a real-world, tidal river basin which qualitatively demonstrate our ability to reconstruct a cluttered field of underwater objects.**

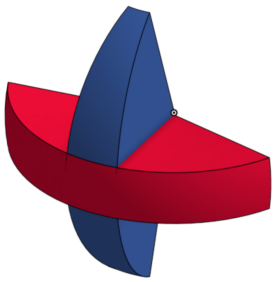
I. Introduction

Over the past several years, autonomous underwater vehicles (AUVs) have seen increased usage, addressing needs ranging from ship inspection to the assessment of offshore oil and gas assets in deep water [1]. To perform these tasks, AUVs can be equipped with a variety of sensors to localize and perform asset inspection. However, when operating in turbid, dark conditions, cameras lose their viability, making sonars the perceptual sensor of choice. When it comes to sonar, three primary modalities have proven useful in a wide variety of applications. Firstly, side-scan sonar has achieved great utility and ubiquity in seafloor mapping applications. Secondly, profiling and bathymetry sonars provide a narrow beam that is highly accurate, but requires many samples to achieve coverage. Thirdly, wide-aperture, forward-looking multi-beam imaging sonars provide an expansive field of view that may be flexibly tasked to gather imagery from a variety of perspectives at a fraction of the cost of its narrow beam competitors. However, imaging sonar is characterized by a high signal to noise ratio and under-constrained measurements, providing flattened 2D imagery of an observed 3D volume. Thus, a subsequent challenge is performing accurate 3D reconstruction of observed objects using its measurements, which lack an elevation angle.

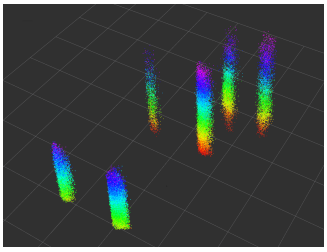
In this work, we will focus on the capabilities of AUVs operating in cluttered environments, such as subsea oil fields and nearshore piers where profiling sonars cannot provide the requisite situational awareness. While profiling sonars, paired with suitable AUV state estimation, can provide highly accurate 3D reconstructions, a profiling sonar can only observe a narrow slice of the environment at a time, making the problem of navigating in three-dimensional clutter potentially intractable. In this work we are motivated by the problem of providing an AUV with the best-possible situational awareness to safely navigate in clutter, rather than computing the most accurate reconstruction possible.



(a) BlueROV2 system overview



(b) Sonar overlapping fields of view



(c) Reconstructed Pier Pilings

Fig. 1: **System Overview.** Using two multi-beam sonars, correspondences between their observations are computed to extract 3D point clouds. Sonar fields of view corresponding to the hardware arrangement in (a) are shown in (b) - the red swath is from the horizontal sonar and the blue is from the vertical sonar, shown at a range of 10m. Fig 1(c) shows a reconstruction of pier pilings in the Hudson River. Data was collected at a fixed depth.

We propose to address an imaging sonar's lack of elevation angle by using an array of two orthogonally oriented sonars (Fig. 1). Our goal, for AUVs to operate reliably in cluttered 3D environments, requires dense volumetric data to be extracted from the perceptual system at every time step, supporting localization and collision avoidance. Accordingly, we propose a novel method to compute the correspondences between two overlapping sonar images of the same scene collected from different vantage points. The core contribution of this paper is a methodology for identifying suitable features and their correspondences across pairs of orthogonal sonar images, and hence measuring the

J. McConnell, J.D. Martin and B. Englot are with the Department of Mechanical Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA {jmcconn1, jmarti3, benglot}@stevens.edu

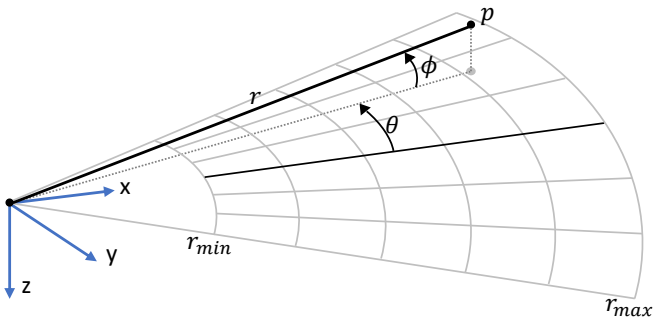


Fig. 2: **Forward looking imaging sonar model.** The point p can be represented by $[r, \theta, \phi]^T$ in a spherical coordinate frame. The range r and the bearing angle θ of p are measured, while the elevation angle ϕ is not captured in the resulting 2D sonar image.

features' locations in 3D Euclidean space. The outcome is the ability to perform dense 3D reconstruction using wide-aperture multi-beam imaging sonar, while making no restrictive assumptions about the scenes in view. Moreover, we do not rely on re-observing features at a later time step to resolve the ambiguity, representing progress towards being able to reconstruct complex 3D cluttered, dynamic scenes with wide-aperture imaging sonar.

In the sections to follow, we will first discuss related work that we draw on for inspiration and benchmarking. Next, we will discuss the specific challenges associated with 3D reconstruction with sonar and precisely define the problem to be solved. Lastly, we will present three experiments. The first will show that our algorithm is comparable with the state of the art in simple cases. The second is a challenging case that violates the assumptions of previous work in this area. Lastly, a field demonstration that shows our algorithm works at scale when observing a cluttered field of objects.

II. Related Work

A. Estimating Elevation Angle

The challenges associated with wide-aperture multi-beam imaging sonars have inspired an impressive body of work to address the fundamental limitations of their under-constrained measurements. Firstly, work from Aykin [2], [3] estimates the elevation angle of sonar image pixels in scenes where objects are lying on the seafloor.

The recent work of Westman [4] extends the work of Aykin and shows excellent results in a constrained nearshore pier environment. However, these methods rely on several assumptions that may often be violated. Firstly, both [3] and [4] assume that all objects in view have their range returns monotonically increase or decrease with elevation angle. While this assumption may hold true for some objects, it hinders the application of their methods to arbitrary objects and scenes. Additionally, [3] requires the leading and trailing edge of an observed object; this is obtained by examining the shadow area behind a segmentation created by the sonar's downward grazing angle. In contrast, [4] only needs to identify the leading *or* trailing edge of the object; however, in the experiments shown, the leading edge is always the closest return because of the sonar's downward grazing

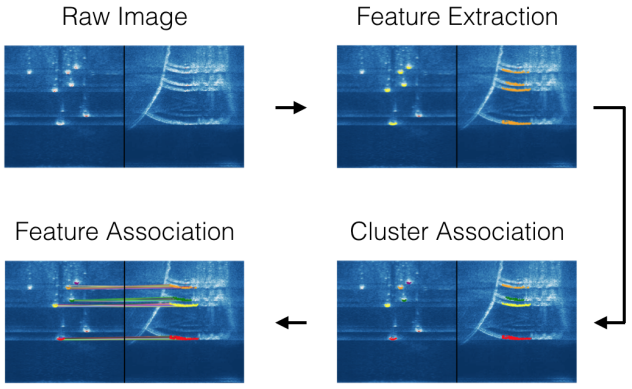


Fig. 3: **System architecture overview.** Raw image pairs are taken from the same time step, features are extracted, the features are clustered and then matched using cluster labels as constraints.

angle. Using a downward grazing angle makes the problem significantly simpler to solve, but comes at a price. By tilting the sonar downward, making the upper edge of the sonar beam parallel with the water plane, the AUV's situational awareness may be hampered. In cluttered environments, an AUV would be unable to see above it before transiting upward, and a safe navigation solution may not always be possible. Moreover, if the vehicle is perturbed in a way that violates this geometric assumption, the perception system could be driven to inaccuracy.

We also note that a recent technique has employed deep learning with convolutional neural networks to estimate the elevation angle associated with imaging sonar observations [5]. In our paper, we assume a vehicle may lack opportunities for prior training and exposure to the subsea objects and scenes it may encounter in a given mission.

B. Carving Out Low-Intensity Background Pixels

Another method proposed by Aykin [6] applies a space carving approach to produce surface models from an image's low-intensity background, which outer-bound the objects of interest. The min-filtering voxel grid modeling approach from Guerneve [7] similarly removes voxels from an object model based on observations of low-intensity pixels. These approaches require the objects of interest to be observed from multiple vantage points to achieve accurate reconstructions.

C. Acoustic Structure From Motion

A similar core issue can also be addressed from a simultaneous localization and mapping (SLAM) perspective. Rather than trying to estimate the elevation of pixels in a single frame, these works acquire features and use a series of views combined with a pose graph back-end [8] to determine 3D structure. This was proposed by Huang [9] in acoustic structure from motion (ASFM). This initial implementation has limitations, chief of which is the reliance on manually extracted features. This work was later built on by Wang [10], incorporating automated feature extraction and tracking. While these methods provide impressive results, they are focused on reconstructing the terrain under the vehicle, rather than providing adequate situational awareness around

the vehicle. The limitation of these methods for AUVs in clutter is the perception system requiring a series of frames to recover 3D information, rather than a single timestep.

D. Stereo Imagery and Feature Correspondences

Lastly, much has been accomplished in the realm of stereo vision, and of specific relevance, computing the correspondences between two vantage points of the same scene. This concept is widely examined in an extensive body of literature [11-14]. Further, many feature extraction methods have been developed over the years, including SIFT [15], SURF [16], ORB [17] and KAZE [18]. Recently AKAZE [19] has shown promise in computing correspondences between acoustic images from a multi-beam imaging sonar. Westman [20] shows the utility of AKAZE features in a SLAM solution using acoustic imagery. Further, Wang [10] utilizes these same features in terrain reconstruction with acoustic imagery.

We also note that the specific concept of using two imaging sonars in stereo for 3D perception has been employed previously, but with relatively small differences in position and orientation between the sonars, and for reconstructing *sparse* sets of point features. The concept, first proposed by Assalih [21], was implemented and further analyzed by Negahdaripour [22], and used to build 3D maps of both sparse features extracted from a planar grid, and from small seafloor objects. Beyond its output of sparse features, this work stands in contrast to ours as the sensors used have aligned axes of uncertainty, rather than the orthogonal axes of uncertainty utilized in our work. A notable example of *dense* 3D mapping is the Sparus AUV's two orthogonally oriented single-beam mechanically scanning sonars (one imager and one profiler), used for cave mapping [23], [24]. However, the imager and profiler were used independently to address SLAM and 3D mapping, respectively, in separate steps.

III. Problem Description

We consider the problem of reconstructing 3D geometry with data gathered by an imaging sonar. Environments are represented as a collection of points $\mathbf{p} \in \mathbb{R}^3$, which define the location of its surface relative to a robot. We express these using coordinates from the robot's local frame $\mathcal{R}$:

$$\mathbf{p}^{(\mathcal{R})} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = R \begin{pmatrix} \cos\phi\cos\theta \\ \cos\phi\sin\theta \\ \sin\phi \end{pmatrix}. \tag{1}$$

Here, X, Y, and Z are the Cartesian coordinates corresponding to the range $R \in \mathbb{R}_+$, bearing $\theta \in \Theta$, and elevation $\phi \in \Phi$, with $\Theta, \Phi \subseteq [-\pi, \pi)$, illustrated in Fig. 2. An imaging sonar measures points in spherical coordinates by emitting acoustic pulses and measuring the associated intensity $\gamma \in \mathbb{R}_+$ from their returns. This information is organized into an *intensity image*, which we view as a set of range-angle-intensity vectors: $\mathbf{z} \in \mathbb{R}_+ \times [-\pi, \pi) \times \mathbb{R}_+$. While intensity image source data comes from three-dimensional observations, images only contain one angle: either bearing or elevation.[1] Therefore, the robot's goal is to reconstruct the underlying 3D coordinates, $\mathbf{p}^{(\mathcal{R})}$, by associating data from multiple images where both θ and ϕ are present.

[1]The associated angle is bearing when the angle sweeps over the x-y plane and elevation when sweeping over x-z.

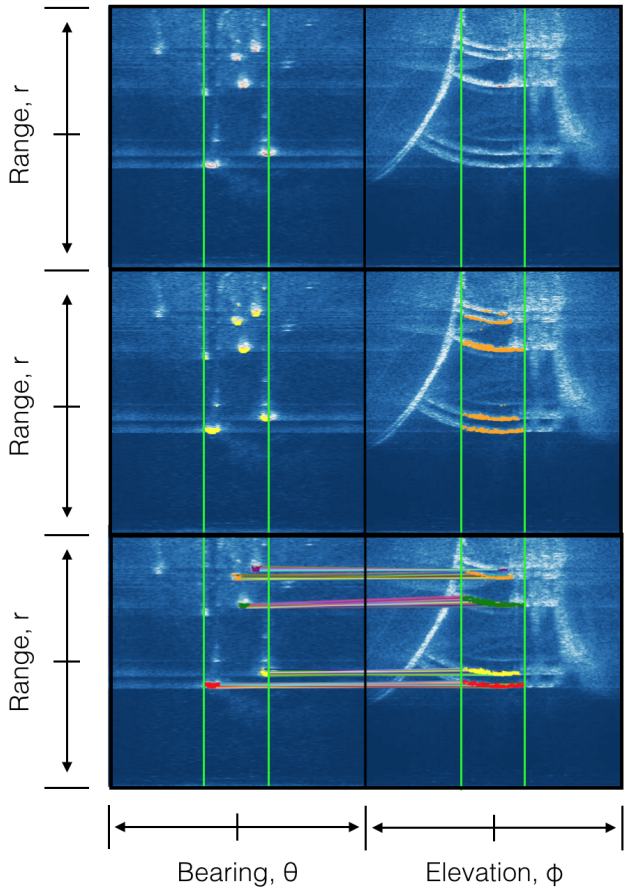


Fig. 4: **Architecture overview with sample images.** The left column shows the horizontal sonar with the right showing the vertical sonar. The top row shows raw images, with green lines denoting the overlapping area pictured in Fig. 1(b). Features are shown in the middle row. The bottom row shows matched clusters in color and lines drawn between matched features.

A. Data Association

This paper studies methods for data association. We assume that the robot is equipped with two forward-looking acoustic sensors. The sensors are mounted such that their fields of view overlap and permit θ and ϕ to be simultaneously observed. This implies that with a proper calibration, two points from each image correspond to the same object location $\mathbf{p}^{(\mathcal{R})}$. We denote these points as

$$\mathbf{z}^{(h)} = (R^{(h)}, \theta, \gamma^{(h)})^\top, \quad \mathbf{z}^{(v)} = (R^{(v)}, \phi, \gamma^{(v)})^\top. \tag{2}$$

The horizontal sensor, h, compresses measurements in the x-y plane of $\mathcal{R}$, whereas the vertical sensor, v, compresses points in x-z. Their associated images sets are denoted

$$\mathcal{Z}^{(h)} = \{\mathbf{z}_1^{(h)}, \cdots, \mathbf{z}_N^{(h)}\}, \quad \mathcal{Z}^{(v)} = \{\mathbf{z}_1^{(v)}, \cdots, \mathbf{z}_N^{(v)}\},$$

where $N \in \mathbb{N}$ represents the number of observations from each sensor.

Given the intensity images $\mathcal{Z}^{(h)}, \mathcal{Z}^{(v)}$, we formalize the data association problem as vertex matching in a bipartite

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertices $\mathcal{V} = \mathcal{Z}^{(h)} \cup \mathcal{Z}^{(v)}$ contain all observed intensity points, and the sets

$$\mathcal{E}_i = \{(\mathbf{z}_i^{(h)}, \mathbf{z}_1^{(v)}), \cdots, (\mathbf{z}_i^{(h)}, \mathbf{z}_N^{(v)})\},$$

define all realizable associations between points in the horizontal image and points in the vertical image. Their union defines the total edge set $\mathcal{E} = \cup_{i=1}^{N} \mathcal{E}_i$. Solutions are obtained by finding a set $\mathcal{S} \subset \mathcal{E}$ such that

$$\mathcal{S} = \bigcup_{\mathcal{E}_i \in \mathcal{E}} \underset{(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(v)}) \in \mathcal{E}_i}{\arg\min} \mathcal{L}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(v)}), \tag{3}$$

where $\mathcal{L}(\mathbf{z}_i, \mathbf{z}_j)$ denotes the loss between features. Additionally we require the association to be bijective, in that for any two edges $(\mathbf{z}_i^{(h)}, \mathbf{z}_i^{(v)})$, $(\mathbf{z}_j^{(h)}, \mathbf{z}_j^{(v)}) \in \mathcal{S}$ it must be that $\mathbf{z}_i^{(v)} \neq \mathbf{z}_j^{(v)}$. We estimate $\mathbf{p}^{(\mathcal{R})}$ from Eq (1), using the fused spherical coordinates

$$\hat{\mathbf{p}}^{(\mathcal{R})} = \left(\frac{R^{(h)} + R^{(v)}}{2}, \theta^{(h)}, \phi^{(v)} \right)^{\top}. \tag{4}$$

Here we use the empirical mean of ranges and the angular values from the resulting association.

B. 3D Reconstruction

Given a set of 3D points $\widehat{\mathcal{P}}^{(\mathcal{R})} = \{\hat{\mathbf{p}}_i^{(\mathcal{R})}\}_{i=1}^{M}$, we can complete the reconstruction by mapping these into a fixed inertial frame $\mathcal{I}$. This is accomplished with the linear transformation $\mathbf{T} \in \mathbb{R}^{3\times3}$. When applied to the set of all points, the result is a point cloud which we call *the map*:

$$\widehat{\mathcal{P}}^{(\mathcal{I})} = \{\hat{\mathbf{p}}^{(\mathcal{I})} | \hat{\mathbf{p}}^{(\mathcal{I})} = \mathbf{T}\hat{\mathbf{p}}^{(\mathcal{R})} \ \forall \ \hat{\mathbf{p}}^{(\mathcal{R})} \in \widehat{\mathcal{P}}^{(\mathcal{R})}\}. \tag{5}$$

IV. Proposed Algorithm

Here we describe our proposed methodology for identifying feature correspondences across concurrent orthogonal sonar images (summarized in Figures 3 and 4). The fundamental goal of this pipeline is to associate range measurements from orthogonal, overlapping vantage points that are each lacking a single dimension in their respective spherical coordinate frames. With a set of matched features, the algorithm output (per Eq. (4)) is a set of fully defined points in 3D Euclidean space.

A. Feature Extraction

In our acoustic imagery we do not process every pixel, as not all pixels represent meaningful returns. We must first identify which pixels belong to surfaces in the scene, and to do this, we apply feature extraction to each acoustic image.

To extract features we use the constant false alarm rate (CFAR) technique [25]. This class of algorithm has shown utility in processing radar images [25], [26] as well as side-scan sonar images [27], which are similarly noisy sensing modalities. We have found it to be the most effective feature detector in practice for reliably and consistently eliminating the second returns that regularly appear in sonar imagery.

CFAR uses a simple threshold to determine if a pixel in a given image is a contact or not a contact. However, it produces a *dynamic* threshold by computing a noise estimate

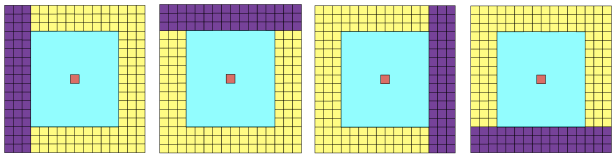


Fig. 5: **SOCA-CFAR overview.** Purple cells show the training cells, blue the guard cells and red the cell under test.

for the area around the cell under test via cell averaging. The technique is sensitive to multiple targets in the image, especially when the noise estimate includes other positive contacts. It is for this reason we utilize a variant known as "smallest of cell averages" (SOCA-CFAR) [26]. SOCA-CFAR computes four noise estimates and utilizes the smallest of the four. This estimate is expressed in Eq. (6), with $\mathbf{x_m}$ as a training cell, $\mathbf{N}$ as the number of training cells in that quadrant and μ as the estimate of noise. This process is shown in Fig. 5. We note that when averages are computed (purple cells), a layer of guard cells (blue) is wrapped around the cell under test to prevent portions of the signal from leaking into the noise estimate. Next, the detection constant is computed in Eq. (7), with α as the detection constant. $\mathbf{N}$ is once again the number of cells in the quadrant and $\mathbf{P_{fa}}$ is the specified false alarm rate. The threshold, β can be computed using Eq. (8), with $\mu_{\mathbf{min}}$, the minimum of the computed quadrant averages. The result of this step in the algorithm is two sets of contacts, $\mathcal{S}_t^{(h)}$ from the horizontal sonar and $\mathcal{S}_t^{(v)}$ from the vertical. Note that at each time step, there are two sonar images; these images are independently analyzed for features. A sample pair of sonar images with CFAR points is shown in the 2nd row of Fig. 4.

$$\mu = \frac{1}{N} \Sigma_{m=1}^{N} x_m \tag{6}$$

$$\alpha = N(P_{fa}^{-1/N} - 1) \tag{7}$$

$$\beta = \mu_{min}\alpha \tag{8}$$

B. Clustering and Cluster Association

Once features are identified in an image pair, these features require an additional constraint for robust association. To minimize the number of extraneous matches, we can take advantage of the fact that each cluster represents a surface in view. By first clustering the extracted features and then comparing features between matched clusters only, we can create a significantly more robust pipeline.

Clustering is performed using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [28] algorithm on both sets of features $\mathcal{S}_t^{(h)}$ and $\mathcal{S}_t^{(v)}$. DBSCAN is selected because, in each image, the number of clusters is unknown, and this approach does not require knowledge of the scene a priori. DBSCAN works by iterating through the data, catalogging all core points with more than *minSamples* neighbors lying within a radius ϵ, which, along with their neighbors, form our initial clusters. All other unassigned points lying within ϵ of a cluster are assigned to that cluster. The result of this process is two sets of clusters: $\mathcal{C}_t^{(h)}$ from the horizontal sonar and $\mathcal{C}_t^{(v)}$ from the vertical sonar.

Next, to match clusters across orthogonal sonars, we compute four descriptors and then minimize a cost function to associate clusters. Each cluster is defined by its mean range μ, variance in range $\sigma^{\mathbf{2}}$, and min and max in range, shown in Eq. (9). Each cluster $\mathbf{c}_t^{(h)}$ is assigned to the cluster in $\mathcal{C}_t^{(v)}$ that minimizes cost function (10):

$$\mathbf{c}_t = \begin{bmatrix} \mu & \sigma^2 & r_{min} & r_{max} \end{bmatrix}^\top, \tag{9}$$

$$\mathcal{L}(\mathbf{c}_t^{(h)}, \mathbf{c}_t^{(h)}) = ||\mathbf{c}_t^{(h)} - \mathbf{c}_t^{(v)}||_2. \tag{10}$$

C. Feature Association

Following the clustering algorithm output, the next stage is to match individual features within our matched clusters. A feature can only be matched to another feature if they belong to the same cluster, which greatly reduces the potential for extraneous feature matches.

To match features, we once again adopt the descriptor and cost function paradigm. Each feature is defined by range, $\mathbf{r}$ intensity γ, and a mean intensity kernel. We consider two mean terms for each feature, which are broken into two axes. $\mu_{\mathbf{x}}$ is the mean intensity of $\mathbf{i}$ points right and $\mathbf{i}$ points left of the feature. $\mu_{\mathbf{y}}$ is the mean intensity of $\mathbf{i}$ points above and $\mathbf{i}$ points below the feature, in image coordinates.

The reasoning for this is straightforward; the body of work addressing elevation angle estimation in sonar imagery relies on the relationship between incident angle and intensity. Since our goal is to match similar measurements, we leverage this relationship by hypothesizing that not only should similar measurements have similar range, but also intensities, because of their similar incident angles. However, due to the noise in acoustic images, we adopt not only range and intensity as feature descriptors, but also local averages of intensity. Note that feature descriptors in Eqs. (11) and (12) have mismatched axes in their μ terms, which is due to the orthogonality of the images:

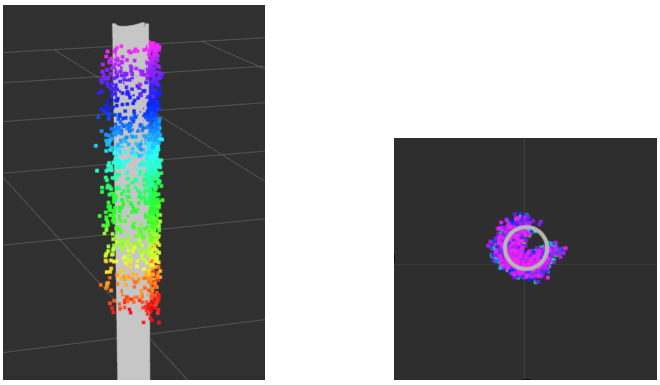
$$\mathbf{z}_i^{(h)} = \begin{bmatrix} r & \gamma & \mu_x & \mu_y \end{bmatrix}^\top \tag{11}$$

$$\mathbf{z}_j^{(v)} = \begin{bmatrix} r & \gamma & \mu_y & \mu_x \end{bmatrix}^\top \tag{12}$$

$$\mathcal{L}(\mathbf{z}_i^{(h)}, \mathbf{z}_j^{(v)}) = ||\mathbf{z}_i^{(h)} - \mathbf{z}_j^{(v)}||_2. \tag{13}$$

We next compute the minimum-loss association among the features residing in previously associated clusters. When carrying out this process, we only allow any feature to be matched with a single other feature. Moreover, before evaluating cost we *normalize* all components by subtracting from each feature the min image intensity and dividing by the difference between min and max. These min and max values are computed for each image, at every timestep.

In our implementation, we have developed two versions of this process; the first is a "brute force" method in which all possible correspondences for a given cluster are tested for each feature, and the lowest cost is used provided it is below a designated threshold. Alternatively, we take a fixed number of random samples from a cluster and again use the lowest cost, provided it is below our threshold.



(a) Isometric view (b) Top down view

Fig. 6: **Results from pier piling mockup.** A model of the structure's true dimensions is shown in gray (outer diameter is 9cm), with multi-colored points depicting algorithm output. Sonars were operated at 5m range, and results from our algorithm's "fast, clustering" configuration are shown. Colors indicate height.

These two versions of our proposed algorithm will later be quantitatively compared.

V. Experiments and Results

In this section, we will examine three experiments and provide an analysis of their outcomes. We will discuss the specifics of our experimental setup and the hardware used to validate the proposed perceptual framework of Fig. 3.

A. Hardware Overview

In order to collect data for evaluation, our heavy-configuration BlueROV2 underwater robot was deployed with a customized sensor payload that includes a camera, a Rowe SeaPilot Doppler velocity log (DVL), a VectorNav VN100 inertial measurement unit (IMU), and a Bar30 pressure sensor. In addition, two forward-looking multi-beam imaging sonars were used (selected based on our available resources for these experiments); Oculus M750d and Oculus M1200d, which were operated in their low-frequency, wide-aperture modes at 750kHz and 1.2Mhz, respectively. In this mode, both sensors have a vertical aperture of 20° and a horizontal aperture of 130°. The M750d was mounted in the horizontal configuration, shown in red in Fig. 1(b), and the M1200d was mounted in the vertical configuration, shown in blue in Fig. 1(b). These sonars return slightly different intensity magnitudes due to their frequencies, however at each time step all feature intensity values extracted from an image are normalized, as described in Section IV.C.

The Robot Operating System (ROS) [29] was used to operate the vehicle, and to log data from a topside computer.

B. Overlapping Area Considerations

When extracting features, we take a highly conservative approach to ensure that features are only extracted from the region of overlap between the sonars. We only extract features from the area inside the vertical aperture of each sonar's orthogonal companion, shown in green in Fig. 4. This conservative approach ensures no features are extracted outside of the overlapping area depicted in Fig. 1(b).

C. Compensating for Sonar Misalignment

Before any data association can take place, we must transform our vertical sonar features $\mathcal{S}_t^{(v)}$ to the horizontal sonar frame. In our experimental setup, there is a 10cm vertical offset between the sonar coordinate frames, shown in Fig. 1(a). This is accounted for by applying Eq. (1) to transform our features in $\mathcal{S}_t^{(v)}$ to Cartesian coordinates, translate them a few centimeters downward to the horizontal sonar coordinate frame, and then apply the inverse of Eq. (1) to transform the features back into spherical coordinates for association. To apply Eq. (1), an elevation angle of zero is assumed for all points in the vertical sonar image (this assumption is used to facilitate data association only).

D. Error Metrics

To compare different parameterizations of our method against each other and a benchmark, we utilize two error metrics. Firstly we compute mean absolute error (MAE); this is calculated by comparing the generated point cloud to a CAD model of the object. Secondly, we compute the root mean squared error (RMSE); once again, this is generated by comparing the point cloud to the CAD model. For our tank experiments, to capture the transformation between the ROV frame and the CAD model with high accuracy, the objects in our sonar imagery are hand segmented to identify both the object coordinate frame origin (located at the center of the pipe comprising all structures), and its rotation in yaw relative to the ROV. The remaining angles in the transformation are obtained from the ROV's IMU.

E. Simple Object Reconstruction

In this first experiment, a cylindrical piling mock-up is submerged in our test tank, and a sequence of data is collected with our sonar pair oriented at a grazing angle of 20° below horizontal. This grazing angle is used in order to facilitate benchmarking, as the current state of the art [4] assumes this problem structure. When collecting data, the ROV is piloted in a circular-segment orbit pattern around the structure. During this flight pattern, the structure is held at a similar range and bearing while the ROV traverses to port or starboard. Data collection occurs at a fixed depth.

The purpose of this experiment is to evaluate the performance of the proposed algorithm on a simple structure. Additionally, this experiment allows us to evaluate our algorithm against the current state of the art, as [4] shows several experiments reconstructing similar objects. During the comparison, it is essential to note that we use SOCA-CFAR feature extraction in conjunction with our implementation of [4] to ensure that all systems run on the same inputs.

While the goal of our algorithm is to move toward the reconstruction of arbitrary objects with wide-aperture multibeam sonar, we are acutely aware that the algorithm we are proposing requires a second sonar. This experiment serves to show that performance does not degrade in the case of objects that can be successfully reconstructed with a single sonar. We compare four variations of our algorithm against [4], as it achieves the best performance in our tank among the suitable algorithms in the literature. Additionally, we posit that [4] fairly represents the foundational works of [2], [3], building upon them and offering broader applicability.

Algorithm	MAE (cm)	RMSE (cm)
Westman and Kaess [4]	2.20	2.64
Brute Force Without Clustering	2.16	2.53
Fast Without Clustering	2.34	2.76
Brute Force With Clustering	2.27	2.70
Fast With Clustering	2.35	2.83

TABLE I: A summary of reconstruction performance corresponding to the reconstruction of a single piling (pictured in Fig. 6).

We compare four configurations of our algorithm to show performance gains and trade-offs for different versions quantitatively. Firstly we evaluate the introduction of clusters as feature correspondence constraints. Secondly, we compare the "brute force" approach in which all feature combinations are checked via (13), and a version in which ten random samples from a feature's corresponding cluster are checked, with the best adopted if below a designated threshold. This second version runs in real-time over all data gathered, while the brute force method runs significantly slower. We also note that [4] runs in real-time in these experiments. The feature correspondence threshold is the same for all methods and experiments provided in this paper; it is set to 0.1.

Experimental results are shown in Table I and Fig. 6; our proposed algorithm performs comparably to the current state of the art, and moreover, the results from our implementation of [4] are in line with those shown in the original paper. Any variation in performance relative to [4] can be attributed to the fact that the sonar used in our implementation has half the angular resolution as in [4], and we analyze raw point clouds rather than filtered surface meshes.

When analyzing the results of different configurations of our proposed algorithm, it is not surprising that in this simple case, with only a single cylindrical piling in view, the cluster constraints provide little added value. Additionally, the trade-off between fast and brute force methods is evident, with a slight loss of accuracy in exchange for real-time viability.

F. Complex Object Reconstruction

Recall that to reconstruct objects using a single wide-aperture imaging sonar, the framework proposed in [4] must make two critical assumptions; the first being that the range to an object increases or decreases monotonically with elevation angle. The consequence of an object violating this assumption is the inability to reconstruct the geometry accurately. Moreover, [4] states that "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." The second key assumption is that the sonar is oriented at a downward grazing angle, which enables identification of an object's leading edge. In our proposed framework, we require no assumptions about the geometry of the objects in view, nor do we require the sensor at a grazing angle.

In this experiment, we test our proposed algorithm on a mock-up of a critical piece of subsea infrastructure, the

blow out preventer (BOP) - approximated by a rectangular object mounted on a cylinder. BOPs sit on the seafloor at the wellhead during offshore drilling and are increasingly subject to regulatory scrutiny and industry monitoring requirements. Critically though, like many other subsea assets, the vertical cross-section of this object does not conform to the aforementioned geometric assumptions.

Once again, data is collected by piloting the ROV in a circular-segment orbit around a portion of the structure (low clearances in our tank prevent full circumnavigation of the structure). Recall that this flight pattern keeps the structure at a similar range and bearing while the ROV traverses to port or starboard. In this experiment, however, the sonar is not at a grazing angle, configured instead for maximum situational awareness. Unlike in the other experiments in this paper, data is collected at a variety of depths.

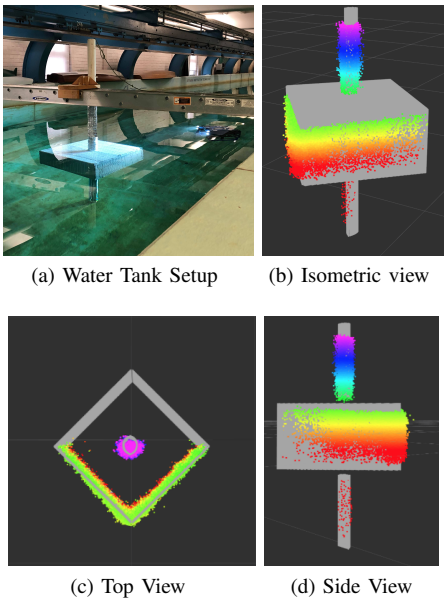
A summary of the results is provided in Table II and Fig. 7. Not only is our algorithm able to provide a realistic reconstruction of the object, but it can do so within 6cm of its true geometry. In analyzing the results of this experiment, the benefits of clustering as a constraint in feature association are evident; clustering dramatically improves both MAE and RMSE. Again the trade-off between brute force and fast feature-matching is evident; a small decrease in performance is required in order to run in real-time. This trend is not evident without clustering, and the reason for this is not known for certain, but the random guessing associated with the fast version may be acting as a filter. The reconstruction's appearance in Fig. 7 is realistic, with no evident outliers, particularly in the vertical axis. While it is unfortunate our algorithm is unable to reconstruct the top surface of the rectangular object, this was not surprising given the object's sharp angles, and the occlusions they present. From the vantage points examined, the top rectangular surface of the structure was imaged at much lower intensity than its front edges. Video playback of our algorithm's mapping of the BOP structure is provided in our video attachment. We have omitted a direct comparison with [4] for this structure in an effort not to misrepresent their work, in a modality they explicitly state their algorithm is not intended for.

Algorithm	MAE (cm)	RMSE (cm)
Brute Force Without Clustering	62.29	69.39
Fast Without Clustering	30.24	45.91
Brute Force With Clustering	5.31	10.06
Fast With Clustering	5.74	12.75

TABLE II: A summary of model reconstruction performance corresponding to the reconstruction of the BOP mockup.

G. Field Based Object Reconstruction

The final experiment to be presented is a demonstration of the proposed algorithm operating in the field. For this work, our ROV was deployed in the Hudson River at Hoboken, NJ. A short inspection flight was flown at the junction of two piers with six pilings supporting the structure close to each other. This exercise takes place at a fixed depth and the sonars without any grazing angle. The version of the algorithm used for this demonstration is fast, with clustering.



(a) Water Tank Setup (b) Isometric view

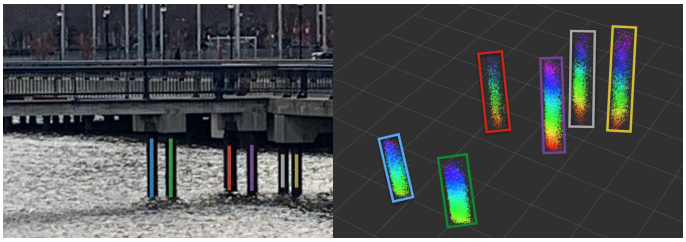
(c) Top View (d) Side View

Fig. 7: **Results from blowout preventer mockup.** Gray shows the true structure geometry (a 9cm OD pipe, with an 82 x 82 x 45cm box, whose center is 1.2m from the bottom of pipe), and in (c), the top of the structure is not shown for ease of visualization. Sonars were operated at 5m range, and results from our algorithm's "fast, clustering" configuration are shown. Color indicates height.

The results, shown in Figs. 1(c) and 8 (and whose sonar images also appear in illustrative Figs. 3 and 4), demonstrate that a robot employing the proposed framework can achieve significant situational awareness, even while operating at a fixed depth. We can recover dense 3D point clouds at every timestep, which are registered here using only iterative closest point (ICP). For the same inspection coverage to be achieved with a profiling sonar, changes in depth would be required. Furthermore, reconstructing these pilings with a single wide-aperture sonar [4] would require a grazing angle that would limit situational awareness. Moreover, if the vehicle is perturbed, an event of significant likelihood given the wakes and currents encountered in this tidal river basin, the requirements of prior algorithms may be violated. There is another crucial trade-off here: to extract 3D information from the sensors, a reduction in the horizontal field of view to the overlapping area between sonars is required. We believe this is a reasonable trade, though, as Figs. 3 and 4 demonstrate that reasonable coverage can be achieved, even with the reduced-size portions of sonar images that overlap, at the sensing range of 10m explored in this experiment.

VI. Conclusions

In this paper, we have presented a new framework for achieving dense 3D reconstruction of arbitrary scenes using a pair of orthogonally oriented, wide-aperture multi-beam imaging sonars. We have provided a detailed description of a pipeline for robust feature extraction, clustering, and descriptors that facilitate accurate matching across concurrent sonar frames. Further, we have shown quantitatively that our algorithm performs competitively with the state of the art in reconstruction using a single imaging sonar, in simple



(a) Outdoor Pier Pilings (b) Piling Reconstruction

Fig. 8: Hudson River pier pilings at left with reconstruction at right (shown also in Fig. 1(c)). Data was collected at a fixed depth, with sonars operated at 10m range, and results from our algorithm's "fast, clustering" configuration are shown. Point colors indicate height.

cases where a comparison is possible. Most importantly, this methodology introduces a new dense sonar mapping capability for complex scenes, and unlike previous work it requires few if any assumptions about the environment, advancing progress toward reconstructing arbitrary geometries and cluttered scenes with wide-aperture multi-beam sonar.

Future work will focus on using this methodology as a basis for mapping beyond the area of explicit overlap between sonars, and incorporating this framework into a process for robust, underwater 3D active SLAM.

ACKNOWLEDGEMENTS

This research was supported by a grant from Schlumberger Technology Corporation. We thank Timothy Osedach, Stephane Vannuffelen and Arnaud Croux for constructive comments that have improved the quality of this manuscript.

REFERENCES

[1] J. Vincent, S. Vannuffelen, S. Ossia, A. Speck, G. Strunk, A. Croux, A Jarrot, G. Choi, T.P. Osedach, A. Gelman, S. Grall, G. Eudeline, "Supervised Multi-Agent Autonomy for Cost-Effective Subsea Operations," *Offshore Technology Conference*, 2020.

[2] M.D. Aykin and S. Negahdaripour, "On Feature Matching and Image Registration for Two-dimensional Forward-scan Sonar Imaging," *Journal of Field Robotics*, vol. 30(4), pp. 602-623, 2013.

[3] M.D. Aykin and S. Negahdaripour, "Forward-look 2-D sonar image formation and 3-D reconstruction," *Proceedings of the IEEE/MTS OCEANS Conference*, 2013.

[4] E. Westman and M. Kaess, "Wide Aperture Imaging Sonar Reconstruction using Generative Models," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8067-8074, 2019.

[5] R. DeBortoli, F. Li, and G. Hollinger, "ElevateNet: A Convolutional Neural Network for Estimating the Mission Dimension in 2D Underwater Sonar Images," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8040-8047, 2019.

[6] M.D. Aykin and S. Negahdaripour, "Three-Dimensional Target Reconstruction From Multiple 2-D Forward-Scan Sonar Views by Space Carving," *IEEE Journal of Oceanic Engineering*, vol. 42(3), pp. 574-589, 2017.

[7] T. Guerneve, K. Subr, and Y. Petillot, "Three-dimensional reconstruction of underwater objects using wide-aperture imaging SONAR," *Journal of Field Robotics*, vol. 35(6), pp. 890-905, 2018.

[8] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Transactions on Robotics*, vol. 24(6), pp. 1365–1378, 2008.

[9] T.A. Huang and M. Kaess, "Incremental data association for acoustic structure from motion," *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1334-1341, 2016.

[10] J. Wang, T. Shan, and B. Englot, "Underwater Terrain Reconstruction from Forward-Looking Sonar Imagery," *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3471-3477, 2019.

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47(1–3), pp. 7–42, 2002.

[12] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25(8), pp. 993–1008, 2003.

[13] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: from software to hardware," *International Journal of Optomechatronics*, vol. 2(4), pp. 435–462, 2008.

[14] B. Tippetts, D. J. Lee, K. Lillywhite, and J. Archibald, "Review of stereo vision algorithms and their suitability for resource-limited systems," *Journal of Real-Time Image Processing*, vol. 11(1), pp. 1–21, 2013.

[15] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[16] H. Bay, A. Ess, T. Tuytelaars, L. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110(3), pp. 346-359, 2007.

[17] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564-2571, 2011.

[18] P. Alcantarilla, A.Bartoli and A. Davison, "KAZE Features," *Proceedings of the European Conference on Computer Vision*, pp. 214-227, 2012.

[19] P. Alcantarilla, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," *Proceedings of the British Machine Vision Conference*, 2013.

[20] E. Westman, A. Hinduja, and M. Kaess, "Feature-Based SLAM for Imaging Sonar with Under-Constrained Landmarks," *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3629-3636, 2018.

[21] H. Assalih, Y. Petillot, and J. Bell, "Acoustic Stereo Imaging (ASI) System," *Proceedings of the IEEE/MTS OCEANS Conference*, 2009.

[22] S. Negahdaripour, "Application of Forward-Scan Sonar Stereo for 3-D Scene Reconstruction," *IEEE Journal of Oceanic Engineering*, vol. 45(2), pp. 547-562, 2020.

[23] A. Mallios, P. Ridao, D. Ribas, M. Carreras, and R. Camilli, "Toward autonomous exploration in confined underwater environments," *Journal of Field Robotics*, vol. 33(7), pp. 994-1012, 2016.

[24] A. Mallios, E. Vidal, R. Campos, and M. Carreras, "Underwater caves sonar data set," *The International Journal of Robotics Research*, vol. 36(12), pp. 1247-1251, 2017.

[25] M. Richards, *Fundamentals of Radar Signal Processing*, McGraw Hill, 2005.

[26] K. El-Darymli, P. McGuire, D. Power and C. Moloney, "Target detection in synthetic aperture radar imagery: a state-of-the-art survey," *Journal of Applied Remote Sensing*, vol. 7(1), pp. 6014-6058, 2013.

[27] G. Acosta and S. Villar, "Accumulated CA–CFAR Process in 2-D for Online Object Detection From Sidescan Sonar Data," *IEEE Journal of Oceanic Engineering*, vol. 40(3), pp. 558-569, 2015.

[28] M. Ester, H. Kriegel, J. Sander, X, Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, 1996.

[29] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "Robot Operating System" *ICRA Workshop on Open Source Software*, 2009.