Robust Inertial-aided Underwater Localization and Navigation based on Imaging Sonar Keyframes

Yang Xu, Ronghao Zheng, Meiqin Liu, and Senlin Zhang

Abstract—Imaging sonars have shown better flexibility than optical cameras in underwater localization and navigation for autonomous underwater vehicles (AUVs). However, the sparsity of underwater acoustic features and the loss of elevation angle in sonar frames have imposed degeneracy cases, namely under-constrained or unobservable cases according to optimization-based or EKF-based simultaneous localization and mapping (SLAM). In these cases, the relative ambiguous sensor poses and landmarks cannot be triangulated. To handle this, this paper proposes a robust imaging sonar SLAM approach based on sonar keyframes (KFs) and an elastic sliding window. The degeneracy cases are further analyzed and the triangulation property of 2D landmarks in arbitrary motion has been proved. Comparative experiments validate the effectiveness of the proposed method and its robustness to outliers from the wrong data association, even without loop closure.

Index Terms—Autonomous underwater vehicle, Localization and navigation, Imaging sonar, Keyframe, SLAM

I. INTRODUCTION

Underwater autonomous localization and navigation with no prior map become more and more essential for autonomous underwater vehicles (AUVs) to complete various kinds of missions. However, the electromagnetic signals from global positioning system attenuate quickly underwater, and the external underwater acoustic positioning systems such as ultra-short baseline, short baseline, long baseline are restricted by limited and expensive beacons and nodes [1]–[3], thus the internal aided navigation (IAN) systems are more preferred.

Among the IAN systems, the widely used inertial navigation system (INS) accumulates error rapidly over time, and the visually augmented navigation (VAN) systems always subject to close sensing range, visibility, and illumination conditions [4]. Therefore, acoustic sensors such as side-scan sonars, multibeam sonars, and imaging sonars are more appropriate in the undersea scene, among which the imaging sonars perform better in underwater object detection, collision avoidance, and acoustic visual navigation [5].

Several practices only use imaging sonar image sequences to estimate sonar poses, such as the 2D optical flow-based method in [6] and 3D motion estimation in [7], [8]. Assuming a planar seafloor, Negahdaripour utilized multiple view bundle adjustment (BA) to estimate the 3D sonar motion just from sonar image flow containing 3D objects and their shadows on the seafloor [9]. However, the feature sparsity and low resolution in noisy sonar images may result in shorter distance and less reliability compared with multi-sensor estimation in underwater application.

To better locate and navigate the AUVs using imaging sonar and other navigation sensors, Extended Kalman Filter (EKF)-based estimation methods have already been applied to fuse multi-sensor measurements in semi-structured or structured underwater environments such as harbors, marinas, and ship hull inspection [10]–[13]. Mallios et al. utilized an EKF to estimate the local pose increment to correct the acoustic image distortions produced by vehicle motion, and an augmented state EKF to estimate and keep the registered scans’ poses [14]. Chen et al. proposed a Rao-Blackwellized particle filter SLAM algorithm for an AUV equipped with a slow mechanically scanning imaging sonar [15].

Nevertheless, the main defects of EKF-based underwater SLAM approaches involve the increasing scale of state vector and the covariance matrix and a resulting higher computational complexity, especially in a large underwater scene.

In contrast, the acoustic BA-based method can also incorporate the multi-source measurements, but in an optimization way, which outperforms in precision and robustness than the EKF-based method when processing potential loop closure and front-end failure [7]. Shin et al. proposed a two-view acoustic BA framework for an AUV mapping the seafloor [16]. More recently, Westman et al. studied further based on the two-view acoustic BA and presented the pose-graph imaging sonar SLAM framework [17], [18]. By introducing terrain factors connecting landmark nodes, Wang et al. modelled the subsea terrain as a Gaussian Process random field defined on a Chow–Liu tree [19].

II. RELATED WORKS

Notably, the special acoustic imaging principle and the missing vertical bearing angle are found to bring about the spurious motion, i.e., the inherent ambiguities of 3D motion and scene structure interpretation, as Negahdaripour explored and analyzed in [20] from 2D forward imaging sonar image sequences using multi-view geometry. More recently, Huang and Kaess observed several landmark degeneracy cases when sonar moves in special directions [21], [22]. In other words, the sparsity of underwater acoustic features and the loss of
elevation angle in 2D sonar images, degeneracy cases occur in SLAM back-end when it comes to wrong data association and insufficient sonar measurements, namely under-constrained cases in the back-end optimization [17], [21], such as relative pose ambiguity and landmark cannot be triangulated. This is also called state unobservable in EKF-based SLAM [23], [24].

To solve this, Huang and Kaess presented the acoustic structure from motion (ASFM) method to optimize both sonar poses and 3D feature positions in this circumstance. They also used on-board navigation sensors such as IMU and DVL to mitigate the degeneracy phenomenon [21], [22]. Further, Westman et al. analyzed the inherent causes attributed to the degeneracy cases in 3D sonar pose and landmark position, from the point of view of matrix singularity [17]. With the learning-based loop closure mechanism, Li et al. proposed a pose-graph imaging sonar SLAM scheme based on ASFM to handle the degeneracy cases and improve accuracy in ship hull inspection [25]. However, successful loop closure mainly depends on rich acoustic features, which may fail in the unstructured scene and bring even larger error. The useful constraints in previous sonar frames have also been neglected.

So far, how to extract and utilize constraints from past sonar images to enhance the SLAM accuracy and robustness to outliers, even without loop closure, then achieve the long-term underwater localization and navigation is still an open problem.

Motivated by these related works [16], [17], [21], in this paper, we propose a new approach to enhance the accuracy and robustness of imaging sonar SLAM based on the two-view acoustic BA. Firstly, as an extension of our previous work [26], we further analyze the degeneracy cases in feature-based 2D imaging sonar SLAM, including the relative pose ambiguity and landmark triangulation failure, and prove the triangulation property of 2D landmarks in arbitrary motion. Secondly, we define and identify the well-constrained sonar keyframes (KFs) corresponding to the non-degeneracy cases via the effective constraint information volume. Finally, we fuse the inertial navigation increment and utilize these KFs to constrain the potential ill-conditioned back-end sliding window optimization, the window size can be adjusted based on multiple indicators related to the KFs so the past effective constraints can be mostly utilized, so the SLAM optimization scale and computational complexity can be ensured. In addition, the proposed scheme is more robust to the outliers from the front-end. More extra comprehensive simulations are carried in a combined path using a simulated remotely operated vehicle (ROV) equipped with a forward-looking sonar, which also verifies the effectiveness and robustness to sparse features and wrong associations. The experiment of the marina dataset also validates this.

The rest of this paper is organized as follows. We present the inertial-aided 2D imaging sonar SLAM in Sect. III and analyze the degeneracy cases in Sect. IV. In Sect. V we define the sonar keyframe and propose a keyframe-based elastic windowed SLAM optimization scheme. Comparative experiment results and discussions are given in Sect. VI. Concluded remarks and future work are provided in Sect. VII.

### III. INERTIAL NAVIGATION-AIDED IMAGING SONAR SLAM

In this paper, we consider a feature-based SLAM problem using imaging sonar and low-cost inertial navigation sensors, where the inertial measurement is acquired from an on-board IMU and the absolute depth measurement is from a depth meter. We briefly introduce the imaging sonar model and the inertial kinematic model within the factor-graph SLAM framework as follows, which will be the basis for our proposed method.

#### A. Imaging sonar model

A typical 2D forward-looking imaging sonar model is shown in Fig. 1. Here we use the sonar spherical coordinates to parameterize a feature point \( l = [\psi, r, \theta]^T \), thus the coordinates transformation can be described as follows:

\[
\pi(p^l) = \begin{bmatrix}
\psi \\
\frac{r}{\|p_l^T\|^2}
\end{bmatrix} = \begin{bmatrix}
\arctan2(p^l_{\psi'j}, p^l_{\psi'z}) \\
\frac{r \cos \psi \cos \theta}{r \sin \psi \cos \theta}
\end{bmatrix},
\]

where \( p_l^T \) is the sonar Cartesian coordinates of point \( l \), \( r \in [r_{min}, r_{max}] \) is the measured range, and \( \psi \) is the horizontal bearing angle. \( \pi(\cdot) \) is the projection function which projects \( l \) to a single pixel point at the 2D sonar image. Note that the sonar images lost the elevation angle \( \theta \) of acoustic features in the 3D-2D projection. The sonar vertical aperture and the horizontal field of view (FOV) meets \( \theta \in [\theta_{min}, \theta_{max}] \) and \( \psi \in [\psi_{min}, \psi_{max}] \), respectively.

#### B. Underwater Inertial Navigation

Consider an IMU state vector given as follows:

\[
x_t = \begin{bmatrix}
p_W^T \\
v_W^T \\
R_{WB}^T b_g^T b_a^T
\end{bmatrix}^T,
\]

where \( p_W \) and \( v_W \) are the position and velocity in the world frame \( W \), respectively. \( R_{WB} \) is the rotation matrix from world frame to body frame \( B \), \( b_g \) and \( b_a \) are the biases of gyroscope and accelerometer from IMU.
The corresponding IMU kinematic model is [27], [28]:

\[ \dot{b}_g(t) = \eta_g(t), \]  
\[ \dot{b}_a(t) = \eta_a(t), \]  
\[ \dot{\Delta} R_{WB}(t) = R_{WB}(t)\Omega \left( \tilde{\omega}^m_{WB}(t) - \dot{b}_g(t) - \eta_g(t) \right), \]  
\[ \dot{v}_W(t) = R_{WB}(t)\dot{\tilde{\omega}}^m_{WB}(t) - \dot{b}_a(t) - \eta_a(t) + g_W, \]  
\[ \dot{p}_W(t) = v_W(t), \]

where \( \Omega(\omega) \) is the skew-symmetric matrix of vector \( \omega \in \mathbb{R}^{3 \times 1}. \) \( \tilde{\omega}^m_{WB} \) and \( \dot{\tilde{\omega}}^m_{WB} \) are the angular velocity and linear acceleration measured directly by IMU. \( g_W \) is the gravity acceleration constant vector in the world frame. \( \eta_g \) and \( \eta_a \) are the Gaussian noises in the IMU measurements.

As illustrated in Fig. 2, the difference in update rates between IMU and sonar may result in significant accumulated error on pose prediction of underwater inertial odometry. Therefore, in this paper, we only concern about the pose increment \( \Delta p_W \) and \( \Delta R_{WB} \) between two consecutive sonar frames. In time interval \([t, t + \Delta t]\), the increment can be expressed by:

\[ \Delta p_W = R_{WB}(t)s(t) + g_W \Delta t, \]
\[ \Delta R_{WB} = v_W(t)\Delta t + R_{WB}(t)y(t) + \frac{1}{2}g_W \Delta t^2, \]
\[ R_{WB}(t + \Delta t) = R_{WB}(t)\Omega(u(t)), \]

where

\[ s = \int_t^{t + \Delta t} R_{B_i, B_j}(\tilde{\omega}^m_{WB}(\tau) - \dot{b}_g(\tau) - \eta_\tau(\tau))d\tau, \]
\[ y = \int_t^{t + \Delta t} \int_t^s R_{B_i, B_j}(\tilde{\omega}^m_{WB}(\tau) - \dot{b}_g(\tau) - \eta_\tau(\tau))d\tau ds, \]
\[ u = \int_t^{t + \Delta t} (\tilde{\omega}^m_{WB}(\tau) - \dot{b}_g(\tau) - \eta_\tau(\tau))d\tau. \]

This increment can be evaluated by discrete numerical integration methods such as Runge-Kutta and Simpson Formula.

C. Feature-based two-view imaging sonar SLAM

Assuming Gaussian noises in the odometry and sonar measurements, we can get the following measurement functions:

\[ z^i_{odom} = f(x_{i-1}, x_i) + N(0, \Lambda_i), \]
\[ z^k_{sonar} = h(x_{ik}, l_{jk}) + N(0, \Sigma_k), \]

where \( x_i = [\psi_{x_i}, \varphi_{x_i}, \phi_{x_i}, t_{x_i}, t_{y_i}, t_{z_i}]^T \) \( (i = 1, 2, 3, \ldots) \) represent sonar pose and \( l_{jk} \) denotes the \( j \)-th sonar feature. The sonar measurement associated with pairwise pose \( x_{ik} \) and feature \( l_{jk} \) is expressed as Eq. (1), then gets the predicted 2D polar coordinates \( \hat{z}^k_{sonar} \).

It is worth noting that in two-view SLAM, the previous pose \( x_{i-1} \) denoted by \( x_A \) is set to be constant zero as a reference, and the current pose \( x_i \) denoted by \( x_B \) represents the relative pose between these two consecutive sonar frames. The prior factor is omitted here. Therefore, for each feature point \( l_j \) measured at \( x_A \), the predicted sonar measurement at pose \( x_B \) is as follows:

\[ \hat{z}^k_{sonar} = h(x_{ik}, l_{jk}) = \pi(q_{jk}), \]
\[ q_{jk} = T_{x_B}^{-1} \begin{bmatrix} \frac{1}{2} p_{jk}^T \ t_{x_B} \end{bmatrix} = \begin{bmatrix} R_{x_B} \ t_{x_B} \end{bmatrix}, \]

where \( p_{jk}^T \) is the feature’s Cartesian coordinates at sonar pose \( x_A \) and \( q_{jk} \) is the corresponding reprojected coordinates at pose \( x_B \). \( T_{x_B} \in \mathbb{R}^{4 \times 4} \) is the homogeneous transformation matrix, \( R_{x_B} = \Omega_{xy}(\psi_{x_B}, \varphi_{x_B}, \phi_{x_B}) \) is the relative 3D rotation matrix, and \( t_{x_B} = [t_x^B, t_y^B, t_z^B]^T \) is the relative translational vector.

Therefore, the SLAM problem converts into a Maximum a Posteriori (MAP) problem: Given all the observation \( z = \{ z_{sonar}^1, \ldots, z_{sonar}^M \} \) at \( x_A \) and \( x_B \), respectively, with no prior knowledge, find a maximum posterior probability set of poses and landmarks \( \Theta = \{ x_B, l_1, \ldots, l_M \} \) according to the Bayesian Rule:

\[ \Theta^* = \arg \max_{\Theta} p(\Theta | z) \propto \arg \max_{\Theta} p(z | \Theta) \]
\[ = \arg \max_{\Theta} \prod_{k=1}^N p(z^k_{sonar}, z^i_{odom} | \Theta). \]

As the left sub-figure in Fig. 3, we introduce the factor graph optimization [29] to model the above MAP problem. Since the conditional probability density function of each measurement can be expressed as negative exponential form:

\[ p(z^k_{sonar} | x_{ik}, l_{jk}) \propto \exp \left\{ - \frac{1}{2} \frac{|| \hat{z}^k_{sonar} - z^k_{sonar} ||^2}{\Sigma_k} \right\}, \]
\[ p(z^i_{odom} | x_{i-1}, x_i) \propto \exp \left\{ - \frac{1}{2} \frac{|| \hat{z}^i_{odom} - z^i_{odom} ||^2}{\Lambda_i} \right\}. \]
then we can get a nonlinear least-squares (NLS) problem by
taking the natural logarithm of Eq. (17):
\[
\Theta^* = \arg\min_\Theta \sum_{k=1}^N \left\| h(x_{ik}, l_j) - z_{sonar}^k \right\|^2 + \sum_{i=1}^N \left\| f(x_{i-1}, x_i) - z_{odom}^i \right\|^2_\Lambda.
\]

Note that the odometry term can be incorporated in the relative
pose \( x_B \) to be estimated.

Applying first-order Taylor expansion for Eq. (19) at the
linearized point \( \Theta^0 = \{x_0, l_0^j\} \) and transforming the Mahalanobis norm to Euclidean norm, we can get a linear least-
squares (LLS) problem as follows:
\[
\Delta^* \approx \arg\min_\Delta \sum_{k=1}^N \left\| h(\Theta^0) + H_k \Delta - z_{sonar}^k \right\|^2 = \arg\min_\Delta \left\| A \Delta - b \right\|^2,
\]
where
\[
A_k = \Sigma_k^{-1/2} H_k,
\]
\[
b_k = \Sigma_k^{-1/2} (z_{sonar}^k - h(\Theta^0))
\]
are the whitened Jacobian matrix and the error vector, respectively. \( \Sigma_k^{-1/2} \) is the square-root information matrix and
\[
\Delta^* = \Theta - \Theta^0
\]
is the update vector to be solved.

The closed-form \( A \) is more appropriate in back-end optimization for its computational efficiency, rather than numerical
computation iteratively, i.e. \( (k \) is omitted here):
\[
\frac{\partial h(x_{ik}, l_j)}{\partial x_i} = \frac{\partial h(x_{ik}, l_j)}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial x_i},
\]
\[
\frac{\partial h(x_{ik}, l_j)}{\partial l_j} = \frac{\partial h(x_{ik}, l_j)}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial l_j},
\]
where, specifically, the sub-matrices composing a whole Jacobian matrix are as follows:
\[
\frac{\partial h(x_{ik}, l_j)}{\partial x_i} = 0_{2\times6}, \quad \frac{\partial h(x_{ik}, l_j)}{\partial l_j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},
\]
\[
\frac{\partial h(x_{ik}, l_j)}{\partial q_{ij}} = \begin{bmatrix} \partial^T_{r \theta} \end{bmatrix} = \begin{bmatrix} \partial^T_{r \theta} \end{bmatrix},
\]
where \( \Omega(q_j) \) is a skew-symmetric matrix as:
\[
\Omega(q_j) = \begin{bmatrix} 0 & -q_{jz} & q_{jy} \\ q_{jz} & 0 & -q_{jx} \\ -q_{jy} & q_{jx} & 0 \end{bmatrix},
\]
\[
q_j = \begin{bmatrix} q_jx \\ q_jy \\ q_jz \end{bmatrix},
\]
\[
\Omega(q_j) = \begin{bmatrix} 0 & -q_{jz} & q_{jy} \\ q_{jz} & 0 & -q_{jx} \\ -q_{jy} & q_{jx} & 0 \end{bmatrix},
\]
\[
\partial^T_{r \theta} = \begin{bmatrix} -r_j \cos \psi_j \sin \theta_j & r_j \cos \theta_j \cos \psi_j \sin \theta_j & -r_j \cos \theta_j \sin \psi_j \\ r_j \cos \psi_j \sin \theta_j & r_j \cos \theta_j \cos \psi_j \sin \theta_j & -r_j \sin \psi_j \\ \sin \theta_j & \cos \theta_j & 0 \end{bmatrix}.
\]

IV. DEGENERACY CASES ANALYSIS IN IMAGING SONAR SLAM

**A. Relative pose ambiguity**

We first consider the degeneracy case of relative pose ambiguity between two viewpoints with multiple DOFs. Especially,
the sparse insufficient observations and wrong feature matches may directly lead to this. We present the necessary condition for the number of needed landmark observations between two sonar frames from the existence on the solutions of LLS problem mentioned above.

Generally, in the two-view SLAM optimization, we assume all $M$ landmarks can be observed from two poses so that the maximum number of all known observation equations is $4M$. There are $6 + 3M$ unknown variables containing 6 from pose $x_B$ and $3M$ from the position of $M$ landmarks. The necessary condition to ensure the LLS can be normally solved is as follows:

$$4M \geq 6 + 3M,$$

i.e., $M \geq 6$, which means that if the number of features is less than 6, the back-end LLS problem, i.e. Eq. (20), cannot be fully constrained. This exactly implies that too sparse features may lead to underdetermined LLS and pose degeneracy. Fortunately, we can introduce more absolute sensors to reduce the number of unknown variables so that the required features can be fewer.

Besides, an alternative approach referred to the numerical optimization theory can be used to solve the underdetermined LLS problem, requiring the normal equation:

$$A^T A \Delta^* = A^T b,$$

where $A \in \mathbb{R}^{p \times q}$ ($p \geq q$) is a full column rank Jacobian matrix, and $A^T A$ is the positive definite Hessian matrix. A conventional solution to this equation is to introduce the pseudo inverse $(A^T A)^{-1} A^T b$. In addition, the QR and Cholesky decomposition methods are more preferred in the sparse optimization problem such as SLAM [30].

However, the above-mentioned methods cannot handle the degeneracy caused by the inevitable front-end perturbations or even wrong feature matches even using modern front-end algorithms.

Intuitively, consider one wrong association between two consecutive sonar frames. As the right figure in Fig. 3 $x_i (i = A, B)$ and $l_j (j = 1, 2, 3)$ are neighboring poses and landmarks marked with white and blue circles respectively, and $m_k (k = 1, \ldots, M, M = 5)$ are sonar measurements marked with green squares. The dashed line between $l_1$ and $x_B$ depicts that $m_3$ is expected to associate with them, but actually, it associates with $l_2$ and $x_B$ due to the wrong feature match. The red square marker represents the actual wrong association.

More specifically, the resulting effects on the back-end optimizer of wrong association are evidently embodied in the underlying Jacobian matrix. As in Fig. 4 the green and orange rectangles represent the Jacobian submatrices with respect to measurement $m_k$ and landmark $l_j$, and the white ones are zero submatrices. The immediate adverse effect is that the rows of whole Jacobian matrix related to $m_3$ and $m_4$ become all same, i.e., the actual Jacobian matrix structure in Fig. 4(a) becomes rank-deficient in Fig. 4(b) at the optimizer. Consequently, the Hessian matrix also gets no longer positive definite, leading to unstable numerical computation and larger error.

B. Landmark degeneracy

The landmark degeneracy cases (or called feature triangulation failure) are mainly caused by specific sonar motion [21]. Yang et al. had theoretically proved that in the 3D environment, any pure translational motion on $x, y$ directions or pure rotation around the $z-$axis, or any combination of the above motion would result in feature triangulation failure [23]. Moreover, The features within other basic motion such as $x$ rotation, $y$ rotation, and $z$ translation required more sonar measurements to be triangulated.

Further, we give a complementary proposition and theoretical analysis based on these previous work, which will also be applied in this paper.
Proposition 1. Consider a 2D imaging sonar with a narrow vertical aperture, the elevation angles of landmarks can be estimated using the planar assumption and method in [9], which can eliminate the elevation ambiguity. Therefore, given the absolute (or reliable) on-board observations about z translation measured by depth meter and pitch/roll angles by the accurate inertial navigation system, the landmark can be triangulated during pure x/y motion or pure z rotation (yaw), or the combinations of these motion, i. e., the triangulation failure/degeneracy in the three remaining motion can be eliminated.

Proof. Intuitively, we raise an elementary triangulation example of a feature/landmark. As the local factor-graph shown in Fig. 5, given two known poses \( x_A \) and \( x_B = [\psi_x, t_x^0, t_y^0]^T \), and the corresponding measurement factors \( m_1 \) and \( m_2 \), the landmark \( l_j \) is to be triangulated. Thus, the sonar prediction function can be linearized to:

\[
h(l_j) = h(l_j^0 + \Delta l_j) \approx h(l_j^0) + \mathbf{H} \Delta l_j, \tag{29}\]

where \( l_j^0 \) is the linearization point taken as the feature coordinates from previous sonar observation. The \( x_B \) is omitted in \( h(\cdot) \) for brevity. It is worth noting that the unknowns in feature position \( l_j = [\psi_j, r_j]^T \) to be estimated reduce to 2 and can be solved by the update vector \( \Delta l_j \), and the related Jacobian matrix formed by one pose and one feature is \( \mathbf{A}_l = \partial h(l_j)/\partial l_j \in \mathbb{R}^{2 \times 2} \).

1) Pure x motion: In this case, considering a minor increment in \( x_B \) to represent the pure x motion in one time step yields:

\[
\mathbf{t}_{x_B} = \begin{bmatrix} \delta x & 0 \end{bmatrix}^T, \quad \mathbf{R}_{x_B} = \mathbf{I}_{2 \times 2}, \quad \mathbf{q}_j = \mathbf{R}_{x_B}(\mathbf{p}_j - \mathbf{t}_{x_B}) = \begin{bmatrix} r_j \cos \psi_j - \delta x \\ r_j \sin \psi_j \end{bmatrix}. \tag{30}\]

Therefore, based on Eq. (24), the current Jacobian is:

\[
\mathbf{A}_l = \frac{\partial h(l_j)}{\partial l_j} = \begin{bmatrix} -\frac{\partial q_y}{\partial l_j} & \frac{\partial q_x}{\partial l_j} \\ \frac{\partial q_x}{\partial l_j} & -\frac{\partial q_y}{\partial l_j} \end{bmatrix} \partial \mathbf{q}_j \tag{31}
\]

\[
= \begin{bmatrix} \frac{q_x^2 - r_j \delta x \cos \psi_j}{\sqrt{q_x^2 + q_y^2}} & -\delta x \sin \psi_j \\ \frac{r_j \delta x \sin \psi_j}{\sqrt{q_x^2 + q_y^2}} & \frac{q_y^2 + r_j \delta x \cos \psi_j}{\sqrt{q_x^2 + q_y^2}} \end{bmatrix}, \quad \text{rank}(\mathbf{A}_l) = 2. \tag{32}\]

Hence, the Jacobian matrix \( \mathbf{A}_l \) and the corresponding Hessian matrix are both full-rank, which means that Eq. (28) will be solvable uniquely and the landmark can be triangulated with an initial state \( l_j^0 = [\psi_j^0, r_j^0]^T \).

2) Pure y motion: This case is similar to the pure x motion, so we give a brief result:

\[
\mathbf{t}_{x_B} = \begin{bmatrix} 0 \\ \delta y \end{bmatrix}, \quad \mathbf{R}_{x_B} = \mathbf{I}_{2 \times 2}, \quad \mathbf{q}_j = \begin{bmatrix} r_j \cos \psi_j \\ r_j \sin \psi_j - \delta y \end{bmatrix}, \tag{33}\]

\[
\frac{\partial h(l_j)}{\partial l_j} = \begin{bmatrix} \frac{r_j^2 - r_j \delta y \sin \psi_j}{\sqrt{q_x^2 + q_y^2}} & \delta y \cos \psi_j \\ -\frac{r_j \delta y \cos \psi_j}{\sqrt{q_x^2 + q_y^2}} & \frac{q_y^2 + r_j \delta y \sin \psi_j}{\sqrt{q_x^2 + q_y^2}} \end{bmatrix}, \quad \text{rank}(\mathbf{A}_l) = 2. \tag{34}\]

In this case, a full-rank Jacobian matrix \( \mathbf{A}_l \) means the landmark can be triangulated uniquely too.

3) Pure z rotation: In this case, the translational vector \( \mathbf{t}_{x_B} \) is zero, the rotation matrix and \( \mathbf{q}_j \) are:

\[
\mathbf{R}_{x_B} = \begin{bmatrix} \cos \delta \psi & \sin \delta \psi \\ -\sin \delta \psi & \cos \delta \psi \end{bmatrix}, \quad \mathbf{q}_j = \begin{bmatrix} r_j \cos \psi_j \\ r_j \sin \psi_j \end{bmatrix}. \tag{35}\]

This yields:

\[
\frac{\partial h(l_j)}{\partial l_j} = \begin{bmatrix} -\frac{\partial q_y}{\partial \psi_j} & \frac{\partial q_x}{\partial \psi_j} \\ \frac{\partial q_x}{\partial \psi_j} & -\frac{\partial q_y}{\partial \psi_j} \end{bmatrix} \partial \mathbf{q}_j \tag{36}
\]

\[
= \begin{bmatrix} \cos \delta \psi & \sin \delta \psi \\ -\delta \psi \sin \psi_j & \cos \delta \psi \end{bmatrix}. \tag{37}\]

In general, the incremental yaw angle \( \delta \psi \) is quite small in a time step, i. e., \( \sin \delta \psi \approx 0 \) and \( \cos \delta \psi \approx 1 \), then we can get \( \text{rank}(\mathbf{A}_l) = 2 \).

4) Combination motion: The landmark can also be triangulated within the combinations of the above primitive motion since the linear property. The detail is omitted here.

\[\square\]

Remark 1. Note that the statement in Proposition 1 is equivalent to the reduced 2D case approximately.

For these degeneracy cases discussed above, we are prone to apply the singular-value decomposition (SVD) approach to solve this class of nonlinear state estimation problems since the good sensitivity of singular values to perturbations [30, 31].

Specifically, for a LLS problem, a minor singular value represents a less constrained update vector, i. e., a degeneracy case. This motivates us to utilize the singular values to discriminate the degeneracy cases and well-constrained cases in the following section.

V. SONAR KEYFRAME-BASED SLAM OPTIMIZATION

A. Overview and initialization

Fig. 4 shows the proposed sonar keyframe-based elastic windowed optimization scheme. The initialization process mainly involves IMU calibration, IMU-Sonar extrinsic parameters estimation, and other sensors’ initialization, which are omitted here. The front-end generates sonar features and feature matches between successive frames using A-KAZE and FLANN methods. The detected features in each frame are ordered with unique id and saved to the feature database with pixel coordinates unless already in it or matched to existed features. The inertial navigation data between two consecutive sonar poses is completed by the method in Sec. III-B.

B. Identifying degeneracy cases and sonar keyframes

As presented in Proposition 1, given the reliable measurements on z, pitch, and roll motion, the degeneracy cases will only relate to pose ambiguity. A sonar frame would be regarded as an under-constrained one if:
Initialization & Front-end Data Processing

Back-end Sliding Window Optimization

New sonar frame matches

Identifying degeneracy cases and keyframes

Compute Jacobian matrix and its singular values

Under-constrained frame?

Yes

Reject this frame

No

Sonar KF?

Yes

Save to KFs Database

No

Sonar KF?

Update by Inertial Navigation

Elastic Windowed Optimization based on Two-view Acoustic BA

Exceed max window Size?

Yes

Remove lower rank KFs

No

Elastic Window Optimization

Update global pose & landmark position

Fig. 6. Block diagram of the proposed sonar KF-based elastic windowed optimization. As the main block, the back-end optimization includes under-constrained case detection, keyframe selection, and window size regulation.

1) the number of detected and matched features \( N_F \) between two consecutive sonar frames is less than a threshold \( N_{F_{th}} \);
2) the minimum singular value \( \sigma_{min} \) of the Jacobian matrix \( A \) formed by poses and features in the two frames meets \( \sigma_{min} < \sigma_{low} \).

The under-constrained frames will be rejected outside of the sliding window optimization and the corresponding poses and landmarks will be updated only by inertial navigation. On the contrary, the current frame will be the keyframe if \( \sigma_{min} \) is larger than a threshold \( \sigma_{high} \) chosen empirically. The selected KFs will be saved into the KFs database as the candidate frames for the elastic window. Thus we have already defined the saliency for sonar images to identify keyframe.

C. Adjusting the Elastic Window Size and Optimization

Here we choose the qualified KFs from the database to add in the sliding window based on the following rules: (i)
1) the number of same feature ids between the candidate KF and current frame need to be larger than a threshold \( N_{coview} \geq N_{th-cov} \);
2) On this basis, we rank the KFs based on the magnitude of minimum singular value in a descending order, the KF with a greater \( \sigma_{min} \) than the average minimum singular value will be added into the window, i.e.:

\[
\sigma_{KF_{min}} \geq \sigma_{aver} = \frac{1}{N_S} \sum_s \sigma_{s_{min}}.
\]

(35)

Considering maintain an appropriate SLAM optimization scale, the window size \( N_S \) need to be adjusted automatically in a range of \([2, N_{S_{max}}]\) based on two-view acoustic BA. If the window size exceeds the boundary \( N_{S_{max}} \), we only reserve the KFs which rank top \( N_{S_{max}} - 2 \), other KFs are removed from sliding window. \( N_{S_{max}} \) and \( N_{th-cov} \) are chosen depending on the specific mission.

Once the sonar frames in the sliding window are determined, the NLS problem will be well-conditioned approximately, and can also be solved by, e.g., Gauss-Newton or Levenberg-Marquardt method in GTSAM [32].

Note that the Jacobian matrix is computed in an analytical form and we get the singular values by SVD: \( A = USV^T \), where \( U \in \mathbb{R}^{p \times p} \) and \( V \in \mathbb{R}^{q \times q} \) are orthogonal matrices, \( S = [\sigma_1 \; 0_{p-q}]^T \in \mathbb{R}^{p \times q} \) contains singular values \( \sigma_1 \geq \cdots \geq \sigma_p, \sigma_{low} \) and \( \sigma_{high} \) are computed by the closed-form Jacobian matrix and selected empirically in simulations and experiments. \( N_{F_{th}} \) derives from the existence of solutions for the LLS problem mentioned in Sec. IV-A, that is, we need \( M \geq N_{F_{th}} = 6 \) in 3D scene and \( M \geq N_{F_{th}} = 2 \) in 2D scene to ensure the LLS equations to be well-constrained, or so-called non-underdetermined.
VI. RESULTS AND DISCUSSIONS

To validate the proposed method, we first conduct two underwater simulations on the high-fidelity physical model-based simulation platform named Gazebo UUV Simulator [33], and then an field experiment using an open dataset.

In the first two simulations, as in Fig. 7, we use a ROV equipped with imaging sonar, DVL, IMU, depth meter and other on-board sensors to follow a predefined 10 m × 10 m squared path. Note that this path is the combination of pure x, y motion, and pure z rotation, the triangulation degeneracy will not exist given accurate depth, roll, and pitch measurements in this combined motion according to Proposition 1. The ROV keeps a relatively low speed and senses the fixed artificial landmarks along the path. Important parameters are listed in Table I. Note that since we aim to verify the open-loop performance, the loop closure mechanism is not used in the following experiments, though our proposed SLAM is compatible with it.

Since we only concern about the horizontal pose and landmark estimation, we set the imaging sonar at the same depth as all landmarks, so these small angles can be approximated to 0, in other words, this scene is similar to the planar 2D case. Therefore, the minimum number of tracked features is chosen as \( N_{Fthr} = 2 \). We also choose \( \sigma_{least} = 0.13 \) and \( \sigma_{th} = 0.8 \) empirically. Besides, the maximum window size \( N_{Smax} \) and the minimum number of co-viewed features \( N_{th} \) is set to 5 and 4, respectively. Note that this 2D simplification would introduce additional feature position error according to Eq. (1).

As in Fig. 8, the features in the consecutive simulated sonar images are detected and matched by A-KAZE and FLANN approaches in the front-end, the wrong matches has been initially removed by ratio test.

We compare our method with two-view acoustic BA used in [16], [17] in the following experiments.

A. First simulation: sparse features

To simulate the feature sparsity in underwater environment, we place 28 small landmarks along the path and only 25 landmarks are measured. The number of detected acoustic features at each sonar frame is shown in Fig. 9. It’s worth to notice that the number of features reduces to 1 at frames 88 - 89 and 164 - 167, less than the minimum requirement \( N_{Fthr} = 2 \) as analyzed in Sec. IV-B, which would potentially cause undesired results.

TABLE I

| Important Parameters | Values                          |
|---------------------|--------------------------------|
| Sonar range \([r_{min}, r_{max}]\) (m) | [0.5,9]                        |
| Sonar bearing FOV \([\psi_{min}, \psi_{max}]\) (rad) | [-\(\pi/4\), \(\pi/4\)]      |
| Sonar elevation FOV \([\theta_{min}, \theta_{max}]\) (rad) | [-\(\pi/18\), \(\pi/18\)]   |
| Sonar range noise \(\sigma\) (m) | 0.05                           |
| Sonar bearing noise \(\sigma\) (rad) | 0.02                           |
| Odometry translational noise \(\sigma\) (m) | 0.05                           |
| Odometry rotational noise \(\sigma\) (rad) | 0.02                           |
| Sonar position w.r.t ROV (m) | [1.4,0,-0.6]                  |
| Sonar azimuth w.r.t ROV (rad) | [0,0,0]                        |

Fig. 10 (a) compares the trajectories and landmarks estimated by our method and two-view BA with the ground truth. The DR trajectory is also illustrated as a reference. Fig. 10 (b) shows the corresponding trajectory root mean square errors (RMSEs) of different methods. Fig. 10 (c) gives the RMSEs and mean absolute errors (MAEs) of all estimated landmarks at \( x, y \) directions.

In the first one-third of the path, the proposed method outperforms than two-view BA since the well-constrained sonar keyframes are added into the elastic window, and these two are both much better than the INS DR. However, after
encountering the first feature sparsity case at about frame 88, the acoustic BA begins to accumulate larger errors both in pose and landmark estimation, then gets even worse and collapses when meeting the second sparsity case. According to the analysis in Sec. IV, this phenomenon mainly attributes to the insufficient features and the resulting under-determined least square equations. In contrast, the RMSE curve of our proposed method fluctuates when these two sparsity cases occur, but the error is always bounded within about 0.5 m even at the consecutive under-constrained frames 164 - 167 because of the presence of degeneracy identification. On the other hand, the continuous fluctuation in a period of RMSE after the sparsity cases occur exactly indicates the adverse effects of under-constrained frames on feature-based sonar SLAM.

Notably, our method limits the effects in a small range and shows the effectiveness on degeneracy cases. The average position RMSE and MAE of detected landmarks also show the estimation accuracy of the proposed method.

B. Second simulation: add wrong associations

To further verify the robustness of the proposed method, we add two wrong associations at frame 50 to simulate the wrong associations that occurred inevitably in the front-end. Specifically, we manually associate landmarks 6 and 7 to 1 and 8, respectively.

In this case, as shown in Fig. 11(a), the estimated trajectory of two-view acoustic BA starts to deviate the path seriously after adding the wrong feature matches, the large deviation does not get corrected in the remaining path for a long time. The sudden increase and quick accumulation of trajectory RMSE in Fig. 11(b) also evident this.

On the contrary, the dashed rectangle in Fig. 11(a) shows that the deviation in the trajectory of our method is corrected almost instantaneously and the pose estimation always keeps near consistent with the ground truth later. This mainly owes to the minimum singular value $\sigma_{\text{least}} = 0.13$ mentioned earlier identifying the degeneracy cases. The RMSE curve of our method is shown in Fig. 11(b). The overall average landmark position RMSE and MAE in Fig. 11(c) are larger than the ones in Fig. 10(c), but the amplitudes are still limited to 0.5 m with our method, while the two-view acoustic BA results in errors larger than 1 m in both experiments.

Hence, this experiment illustrates our method is robust and resilient to the outliers from the front-end. As evident from these experimental results, the proposed method achieves the expected performance.

C. Field experimental Results

Aiming to validate the practical benefit of our proposed method, we adopt the dataset collected using the Ictinea AUV by Ribas [10] in an abandoned marina (see Fig. 12 (a)). The vehicle was mounted with imaging sonar and onboard navigation sensors such as IMU and DVL. The AUV traveled along a 600 m path with a relatively low speed using more than 45 minutes. The omnidirectional mechanically scanned imaging sonar was set to a maximum range of 50 m, a resolution of 0.1 m, and a step angle of 1.8°. The differential GPS data was used as the ground truth and gathered by the GPS equipped on a buoy rigidly attached on the top of the AUV.

Fig. 12 (b) shows the resulting trajectories in this experiment, and the comparative position RMSEs of estimated trajectories are in Fig. 12 (c). The red markers “+” depict the sonar features extracted from sonar images along the path and the blue markers “*” represent the salient features, which can be chosen from the sonar features by a voting algorithm. Note that we use no loop closure detection in this experiment.
Similar to the former results, our proposed method performs better than the two-view BA and INS DR in pose estimation, which accords with the previous analyses and also shows the effectiveness in field applications.

VII. CONCLUSIONS AND FUTURE WORK

This paper mainly contributes a sonar keyframe-based elastic windowed optimization for long-term underwater autonomous localization and navigation with AUVs. We raise the concept of sonar keyframes and relevant selection criteria based on the effective constraint information volume and
then choose the well-constrained frames to add to the elastic sliding window using certain rules. The under-constrained sonar frames are identified by the singular values of their corresponding Jacobian matrix. Moreover, the potential ill-conditioned least-squares optimization in the state estimation problem has been improved by using these past useful constraints and fusing inertial measurements. The robustness to outliers is enhanced within this scheme either. Future work mainly involves conducting field experiments.

REFERENCES

[1] L. Paull, S. Saedee, M. Seto, and H. Li, “AUV navigation and localization: A review,” IEEE Journal of Oceanic Engineering, vol. 39, no. 1, pp. 131–149, 2013.

[2] Y. Wu, X. Ta, R. Xiao, Y. Wei, D. An, and D. Li, “Survey of underwater robot positioning navigation,” Applied Ocean Research, vol. 90, p. 101845, 2019.

[3] H.-P. Tan, R. Diamant, W. K. Seah, and M. Waldmeyer, “A survey of techniques and challenges in underwater localization,” Ocean Engineering, vol. 38, no. 14-15, pp. 1663–1676, 2011.

[4] R. M. Eustice, O. Pizarro, and H. Singh, “Visually augmented navigation for autonomous underwater vehicles,” IEEE Journal of Oceanic Engineering, vol. 33, no. 2, pp. 103–122, 2008.

[5] M. Teng, L. Ye, Z. Xuyin, Q. Zhang, Y. Jiang, C. Zheng, and T. Zhang, “Robust bathy metric slam algorithm considering invalid loop closures,” Applied Ocean Research, vol. 102, p. 102298, 2020.

[6] B. T. Henson and Y. V. Zakharov, “Attitude-trajectory estimation for forward-looking multibeam sonar based on acoustic image registration,” IEEE Journal of Oceanic Engineering, vol. 44, no. 3, pp. 753–766, 2018.

[7] H. Sekkat and S. Negahdarpour, “3-D motion estimation for positioning from 2-D acoustic video imagery,” in Iberian conference on Pattern Recognition and Image Analysis. Springer, 2007, pp. 80–88.

[8] S. Negahdarpour, “On 3-D motion estimation from feature tracks in 2-D fs sonar video,” IEEE Transactions on Robotics, vol. 29, no. 4, pp. 1016–1030, 2013.

[9] S. Negahdarpour and A. Taatian, “Bundle adjustment for 3-d motion and structure estimation from 2-d optical and sonar views,” in OCEANS 2008. IEEE, 2008, pp. 1–6.

[10] D. Ribas, P. Ridao, J. D. Tardós, and J. Neira, “Underwater SLAM in man-made structured environments,” Journal of Field Robotics, vol. 25, no. 11-12, pp. 898–921, 2008.

[11] A. Mallios, P. Ridao, D. Ribas, and E. Hernández, “Scan matching SLAM in underwater environments,” Autonomous Robots, vol. 36, no. 3, pp. 181–198, 2014.

[12] M. Walter, F. Hover, and J. Leonard, “SLAM for ship hull inspection using exactly sparse extended information filters,” in 2008 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2008, pp. 1463–1470.

[13] M. Franchi, A. Ridolfi, and M. Pagliai, “A forward-looking sonar and dynamic model-based auv navigation strategy: Preliminary validation with feelhippo auv,” Ocean Engineering, vol. 196, pp. 1–15, 2020.

[14] A. Mallios, P. Ridao, D. Ribas, and E. Hernandez, “Probabilistic sonar scan matching slam for underwater environment,” in OCEANS'10 IEEE SYDNEY. IEEE, 2010, pp. 1–8.

[15] L. Chen, A. Yang, H. Hu, and W. Naeem, “Rbpf-msis: Toward real blackwellized particle filter slam for autonomous underwater vehicle with slow mechanical scanning imaging sonar,” IEEE Systems Journal, vol. 14, no. 3, pp. 3301–3312, 2019.

[16] 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 Washington. IEEE, 2015, pp. 1–6.

[17] E. Westman and M. Kaess, “Degeneracy-aware imaging sonar simultaneous localization and mapping,” IEEE Journal of Oceanic Engineering, vol. 2019.

[18] E. Westman, A. Hinduja, and M. Kaess, “Feature-based SLAM for imaging sonar with under-constrained landmarks,” in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 1–9.

[19] J. Wang, T. Shan, and B. Englot, “Underwater terrain reconstruction from forward-looking sonar imagery,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 3471–3477.