Simulator-aided Edge-based Acoustic Camera Pose Estimation

Yusheng Wang
Department of Precision Engineering
The University of Tokyo
Tokyo, Japan
wang@robot.t.u-tokyo.ac.jp

Yonghoon Ji
Graduate School of
Advanced Science and Technology
JAIST
Nomi, Japan
ji-y@jaist.ac.jp

Dingyu Liu
Department of Precision Engineering
The University of Tokyo
Tokyo, Japan
liu@robot.t.u-tokyo.ac.jp

Hiroshi Tsuchiya
Research Institute
Wakachiku Construction Co., Ltd.
Sodegaura, Japan
hiroshi.tsuchiya@wakachiku.co.jp

Atsushi Yamashita
Department of Precision Engineering
The University of Tokyo
Tokyo, Japan
yamashita@robot.t.u-tokyo.ac.jp

Hajime Asama
Department of Precision Engineering
The University of Tokyo
Tokyo, Japan
asama@robot.t.u-tokyo.ac.jp

***Abstract*—In this work, a method to estimate accurate 6 degrees of freedom acoustic camera pose is proposed. Acoustic cameras, also known as 2D forward-looking sonar can generate high resolution 2D images in all water bodies. However, due to the unique imaging principle, techniques such as vision-based localization with acoustic images are still at the early stage. Even we have full 3D information of the scene, it is still difficult to localize the sensor. In this work, we deal with the problem of estimating the accurate relative pose between the camera and a 3D known target which can be applied to tasks like extrinsic calibration. Previous works mainly estimate the camera pose with planar targets; however, the methods are not robust to noise in the image. We propose an accurate pose estimation method with the help of an acoustic image simulator which can deal with 3D targets. We use edge features to match the real images and the synthetic images. A coarse-to-fine strategy is used for global localization and pose refinement. Experiments prove the effectiveness of the method.**

***Index Terms*—forward looking sonar, acoustic camera, simulation, localization, extrinsic calibration**

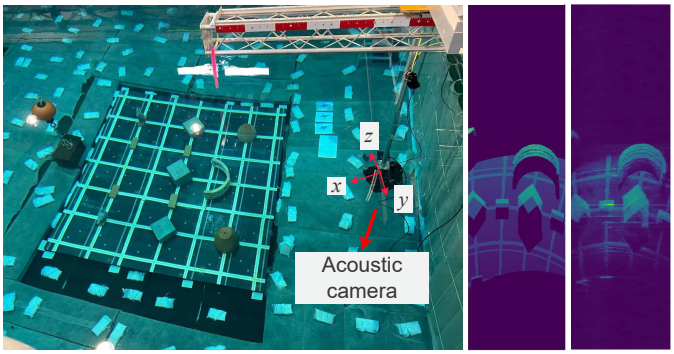

Fig. 1. Pose estimation by matching real and synthetic images. (a) Experimental environment. (b) Real acoustic image. (c) Closest synthetic acoustic image. The accurate pose of the acoustic camera can be estimated by the registration of real and synthetic acoustic images.

I. Introduction

Acoustic cameras also known as 2D forward looking sonars can generate high resolution 2D images even in turbid water [1]. They are gradually mounted on remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs) because of their high performances and compact sizes. They have been already applied to tasks such as 3D mapping and robot navigation [2]–[4]. However, due to the unique imaging principle, there are still many open problems to be solved. It is hard to estimate six degree-of-freedom (6DoF) pose even in a known scene which is important for robot navigation. In order to localize the sensor, techniques such as ultra-short baseline (USBL) perform poorly in shallow water and closed water tanks. Other methods such as combination of doppler velocity log (DVL) and inertial measurement unit (IMU) are limited by drift problems. Although the aforementioned methods can acquire the movement of the sensor, it is still necessary to find the relative pose between the camera and the scene which is also known as extrinsic calibration. This is extremely important for tasks such as dataset collection and underwater monitoring. Currently, the poses are usually manually measured by divers by tools like rulers. However, since the lens-based acoustic camera imaging principle is complex, it is difficult to locate sensor origin.

Previous methods used co-planar points to estimate the extrinsic parameters of the acoustic camera [5], [6]. However, such methods are not robust to noise in the acoustic images. For example, dislocation and ghosting may exist due to multi-path reflection and other ultrasound phenomenon. For many cases, it is even hard to find the 2D-2D correspondence manually. With inaccurate positions of feature points, for z-axis in local acoustic camera coordinate, there will be a

This work was supported by JSPS KAKENHI Grant Number 21J12362.

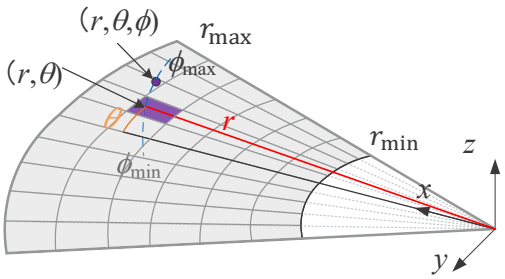

Fig. 2. Projection model. A 3D point (r, θ, ϕ) is projected to (r, θ) in polar coordinate.

decimeter level error for the localization task in the water tank [6], [7]. To better locate the camera, the detection of the illuminated area [7], [8] is usually necessary. However, it is difficult to extract the area when there are objects at the boundaries.

In this paper, a method to estimate accurate extrinsic parameters based on known 3D targets is proposed. As shown in Fig. 1, by finding the closest synthetic acoustic image in the database, it is possible to locate the camera. The initial pose can be firstly estimated online by querying a pre-generated large synthetic image database. Then, the result is refined by generating a local database offline around the estimated initial pose. In this work, we use edge features to mitigate the domain gap between the real and synthetic images. We find the most similar synthetic image compared to the real image. The viewpoint of the synthetic image is considered to be the optimized camera pose. Experiments prove the feasibility of the method.

The rest of the paper is organized as follows. In Section II, preliminaries such as sonar projection model and image simulation are introduced. Section III explains the proposed method. Experiments and evaluations are presented in Section IV. Finally, conclusions and future works are presented in Section V.

II. Preliminaries

A. Projection Model

A 3D point is usually represented by polar coordinate as (r, θ, ϕ) in sonar coordinate. As a multiple beam sonar, it emits N fan-shaped beams along azimuth angle direction and records the backscattered intensity and time of flight. During image generation, the elevation angle ϕ is missing so that the 2D pixel is represented as (r, θ). The scope of the elevation angle is $[\phi_{\min}, \phi_{\max}]$. Enlarging $|\phi_{\max} - \phi_{\min}|$ may generate more optical-like image. On the other hand, if $|\phi_{\max} - \phi_{\min}|$ is close to zero, it is similar to a profiling sonar. For ARIS EXPLORER 3000, the scope of elevation angle is around $14°$ and the image in Fig. 1(b) is an example.

B. Simulator

In our previous works, we built a sonar simulator in Blender[1]. By setting the attenuation of the ray strength based on the inverse square law and assuming the reflection model as Lambertian, it is possible to generate an image of backscattered intensity $\mathbf{I}_f$ using perspective camera model with the same aperture angle as acoustic camera based on ray tracing. Denoting the corresponding depth image as $\mathbf{D}_f$, an acoustic image $\mathbf{I}_a$ can be formed from $\mathbf{I}_f$ and $\mathbf{D}_f$ by projecting the information to $\phi = 0$ plane in camera coordinate [9], [10]. The intensities of the 3D points with the same (r, θ) are linearly integrated here.

III. Approaches

A. Simulator-aided Pose Estimation

We estimate the 6DoF pose with the help of the acoustic image simulator. The basic idea is to find the virtual viewpoint in the simulator closest to the viewpoint in the real world for pose estimation. The acoustic image simulator is developed to generate the synthetic image $\mathbf{I}_s$ from camera pose $\mathbf{w}$ and 3D model $\mathcal{M}$. The process can be described as

$$\mathbf{I}_s = g(\mathbf{w}, \mathcal{M}). \tag{1}$$

Here $g(.)$ refers to the image simulation process. Since range information can be directly detected from the sensor, scale problem does not exist in the acoustic camera. Theoretically, if the relative pose between the target and the camera is the same in real and virtual space, regardless of the domain gap, the same images can be captured in real world and the simulator. It is worth mentioning that the parameters between the virtual camera and the real one have to be the same, such as $r_{\min}$, $r_{\max}$ and the resolution in range direction. If we define the real image as $\mathbf{I}_r$ and the metrics to measure the difference between two images as $d(.)$, the optimization process can be written as follows.

$$\tilde{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, d(\mathbf{I}_r, \mathbf{I}_s). \tag{2}$$

In practical situation, the real images $\mathbf{I}_r$ and the synthetic images $\mathbf{I}_s$ may have a large domain gap. In this paper, we assume the geometry information in $\mathbf{I}_r$ and $\mathbf{I}_s$ is consistent. We use edge information to measure the difference between image in different domains. Another problem is the search space. Searching 6DoF pose in meter-level space require large computational cost. To improve the performance, a coarse-to-fine strategy is used.

Figure 3 shows the optimization process of acquiring the pose of the real image. First, due to the heavy noise in real acoustic image, we use BM3D [11] to eliminate the noise in the acoustic image. Then, Canny edge detector is used to find the edge of the objects in the scene. In order to find the proper pose, we first generate a large number of synthetic images with corresponding poses all around the target in the acoustic simulator. We compute the difference of $\mathbf{I}_s$ and $\mathbf{I}_r$ in edge

[1]https://github.com/sollynoay/Sonar-simulator-blender

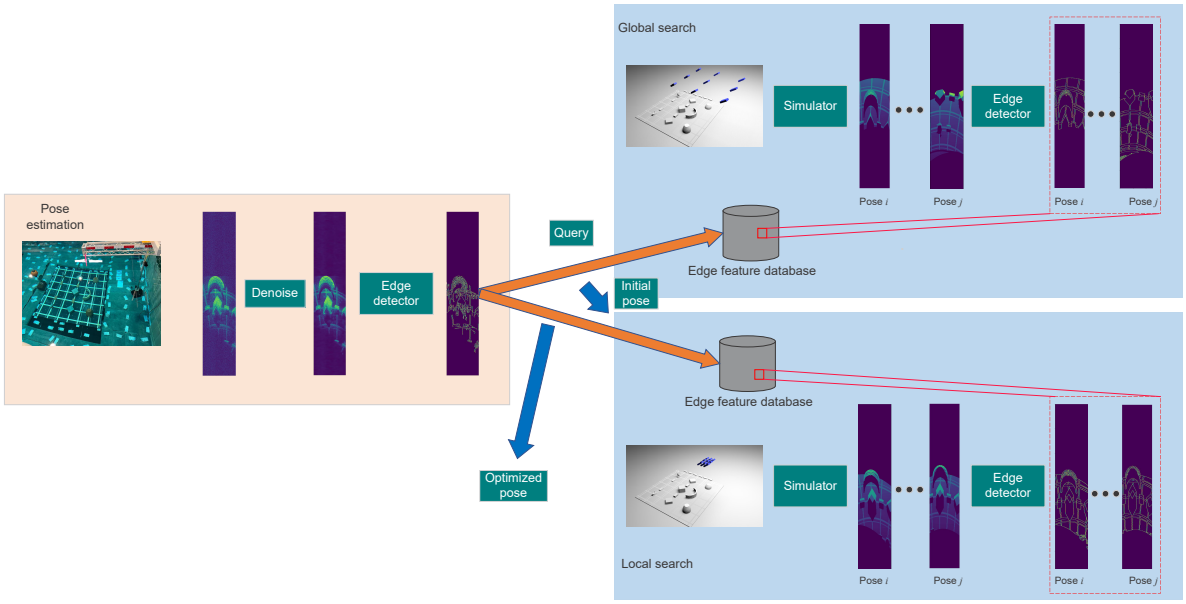

Fig. 3. Overview of the proposed method. A global database is generated beforehand containing virtual images from a large variety of viewpoints. The edge image captured from the real acoustic camera is used to search for the closest edge image in the database. A rough estimation of pose can be acquired from the process. For calibration tasks, a more accurate pose is necessary. We generate a local database with poses around the initial pose and repeat the searching process again for the final result.

domain. The position of edge points in the image are denoted as $\{(x, y)|(x, y) \in S\}$. Then, we use Chamfer distance to measure the difference between the two images in edge space as follows.

$$d(\mathbf{I}_1, \mathbf{I}_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} ||x-y||_2^2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} ||x-y||_2^2. \tag{3}$$

By looking for the minimum distance, it is possible to acquire an initial pose $\hat{\mathbf{w}}$. To acquire a more accurate result, we search for a smaller space with finer steps around the initial pose and repeat the aforementioned process. Finally we can acquire the optimized pose $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{I}}_s$.

B. Coarse-to-fine Strategy

In this work, we assume we have a prior knowledge of the camera pose. The camera is mounted on a moving device and the moving range is limited to a certain scope. The orientation of the camera is controlled by a rotator and towards the ground. Such problem setting can narrow the search space at the beginning. However, in a large-scale water tank environment, the search space is still quite large. The pre-generated synthetic image database cannot cover the whole search space which can only offer a rough estimation. After acquiring an initial pose $\hat{\mathbf{w}}$, we then carry out a local optimization in a small but finer search space. In the simulator, we generate random acoustic images from the following method. Random poses are generated around the initial pose as follows.

$$\mathbf{w} = \hat{\mathbf{w}} + \boldsymbol{\epsilon}, \ \epsilon_k \sim \mathcal{U}(-\beta_k, \beta_k), \tag{4}$$

where ϵ_k and β_k are the k-th element in vector $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$, and $\mathcal{U}$ refers to a uniform distribution. By implementing simulator-aided pose estimation in local space, an accurate pose can be acquired. In this work, we denote pose $\mathbf{w}$ as follows.

$$\mathbf{w} = [x, y, z, \varphi_x, \varphi_y, \varphi_z]^\top, \tag{5}$$

where x, y, x refers to the position in world coordinate and $\varphi_x, \varphi_y, \varphi_z$ refers to $z-y-x$ Euler angles.

IV. Experiments

We carried out experiments in a water tank. For global localization, we generated a database with 21,870 images. We searched the camera poses in a 1.8 m×1.8 m×0.4 m space with a variety of orientations at each position. For local refinement, we set $\boldsymbol{\beta} = (0.05, 0.05, 0.2, 0.017, 0.044, 0.017)$. The uncertainty in z-axis is relatively larger empirically. We generated a database of 10,000 images as the refinement database each time. Examples of the results are shown in Fig. 4. The images for each column represents real images after BM3D, initial synthetic images, final synthetic image, the edge maps of real images, the edge maps of final synthetic images, and the overlap representation of the edge maps of real images (red) and the edge maps of final synthetic images (green), respectively. We successfully found viewpoints close to the real experiment in simulation environment. Although there are ghosting and dislocation effects in real acoustic images, the algorithm can still find the proper solution. The camera pose in global coordinate is shown in Table I.

It was measured by ruler that the z value is around 2 m, and φ_x was close to zero, φ_y was around 0.87 rad, and φ_z was around 1.57 rad. The estimated results satisfied our prior information. For further evaluation, we evaluate the chamfer distance (CD) and the reprojection error (RE) of the red points

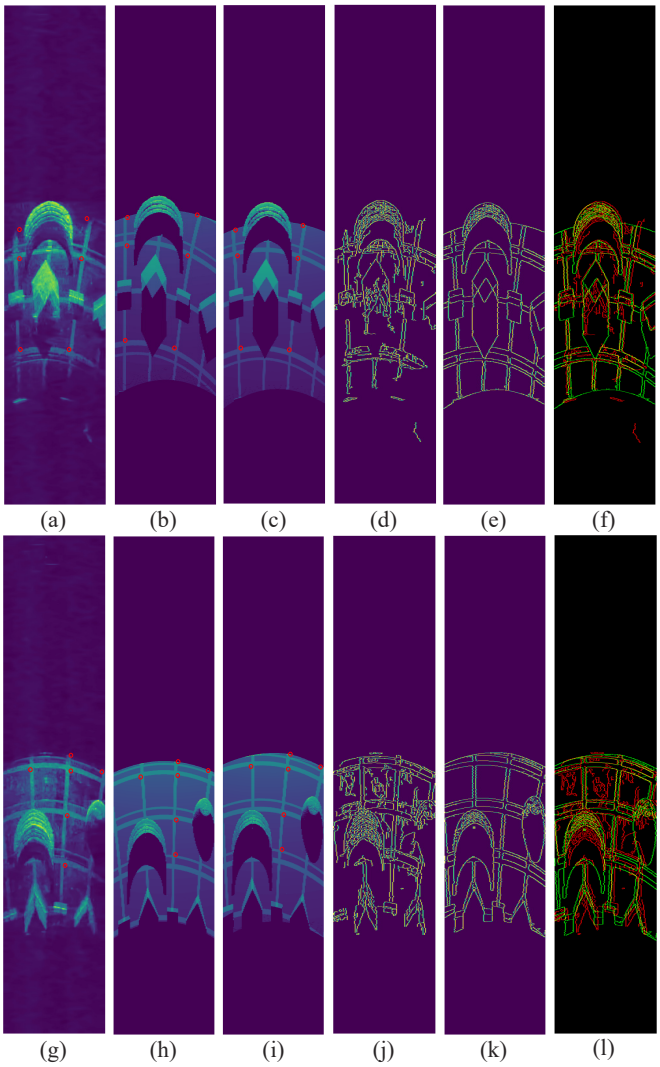

Fig. 4. Experiment results. (a) and (g) show the real images after denoising. (b) and (h) are the optimal synthetic images initial estimation. (c) and (i) are the optimal synthetic images after refinement. (d) and (i) indicate edge images from (a) and (g). (e) and (k) are the edge images from (c) and (i). We overlay (d) (e) and (j) (k) using red and green colors.

TABLE I

	x [m]	y [m]	z [m]	φ_x [rad]	φ_y [rad]	φ_z [rad]
Image 1	-1.54	0.15	2.01	0.02	0.88	1.48
Image 2	-1.18	0.36	2.04	0.01	0.87	1.57

in Fig. 4 before and after refinement between the real and synthetic images. The calculation of reprojection error of each points is as follows.

$$error = |\Delta r| \times \sin(|\Delta\theta|), \tag{6}$$

where Δr and $\Delta\theta$ are the difference of (r, θ) in two images. Results are shown in Table II. Both CD and RE are the smaller the better. It is clear that the error are smaller after local search.

We used a PC with Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz and an NVIDIA GeForce RTX 3090 GPU. For each acoustic image simulation, it took around 4 seconds per virtual image. In other words, it may took 11 hours to build a refinement database. Our purpose is to find the accurate solution as extrinsic calibration. If a rough estimation is enough, it is not necessary for the time-consuming refinement process. Other process, such as BM3D may took 3 seconds for a 128×1343 image on CPU. Computing edges for 10,000 images took 28 seconds. Searching for the virtual image with minimum chamfer distance took 7 minutes. There was much space left to be improved. In the future, we are going to optimize the whole process for more flexible situation, such as real-time localization.

TABLE II

	CD initial	CD final	RE initial	RE final
Image 1	9.22	6.33	0.865	0.175
Image 2	7.31	7.14	3.072	0.076

V. Conclusions

We proposed a method to estimate the 6DoF pose of acoustic camera based on a known target. With the help of the acoustic image simulator, it is possible to find the accurate pose. Future work may include realizing real-time performance and achieving more accurate results considering sonar artifacts such as multi-path reflection.

References

[1] E. Belcher, W. Hanot, and J. Burch, "Dual-frequency identification sonar (didson)," *Proceedings of the 2002 IEEE International Symposium on Underwater Technology (UT2002)*, pp. 187–192, Apr. 2002.

[2] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Acoustic camera-based pose graph slam for dense 3-d mapping in underwater environments," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 3, pp. 829–847, 2021.

[3] S. Negahdaripour, "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, Aug. 2013.

[4] J. Li, M. Kaess, R. M. Eustice, and M. Johnson-Roberson, "Pose-graph slam using forward-looking sonar," *IEEE Robot. and Autom. Lett.*, vol. 3, no. 3, pp. 2330–2337, Jul. 2018.

[5] S. Negahdaripour, "Calibration of didson forward-scan acoustic video camera," *Proceedings of the MTS/IEEE Conference OCEANS 2005*, vol. 2, pp. 1287–1294, Sep. 2005.

[6] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Planar anp: A solution to acoustic-n-point problem on planar target," *Proceedings of the Global Oceans 2020*, pp. 1–6, 2020.

[7] Y. Wang, Y. Ji, D. Liu, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Acmarker: Acoustic camera-based fiducial marker system in underwater environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5018–5025, 2020.

[8] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, "Rotation estimation of acoustic camera based on illuminated area in acoustic image," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 163–168, Sep. 2019.

[9] R. Cerqueira, T. Trocoli, G. Neves, S. Joyeux, J. Albiez, and L. Oliveira, "A novel gpu-based sonar simulator for real-time applications," *Computers & Graphics*, vol. 68, pp. 66–76, 2017.

[10] Y. Wang, Y. Ji, D. Liu, H. Tsuchiya, A. Yamashita, and H. Asama, "Elevation angle estimation in 2d acoustic images using pseudo front view," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1535–1542, 2021.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.