Underwater Localization using an Optic and Acoustic Stereo Imaging System for Autonomous Intervention Robots

Jisung Park[1] and Jinwhan Kim[1]
[1]*Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea*
Email: jinwhan@kaist.ac.kr

Optical and acoustic stereo imaging has great potential for the precise and consistent localization of intervention underwater robots; however, it is still being explored due to its sensing limitations and various technical challenges. This study presents a novel localization method by combining an inertial navigation system and an optical and acoustic stereo imaging system. As a strategy for localization correction relative to underwater structures, the robot's pose is estimated based on a single acoustic image using a sonar simulator for mid-range localization, and a robust visual tracking using a 3-D wireframe model is employed for high-precision localization near the target structures. The performance of the proposed technique was demonstrated through experimental validation using real data obtained from a test tank.

Introduction: The localization technique is essential for automating underwater intervention robots that are widely used for various intervention tasks such as maintenance and repair operations on subsea platforms. Optical cameras and imaging sonars have been explored as important sensors for localization correction of intervention robots because each provides rich visual information for close-range environments and range information over a wide image area, even in extreme underwater environments, respectively. For example, the 2-D geometric transformation between two acoustic images enables an effective localization correction, and it can be obtained via feature point matching [1] and topology graph matching [2]. However, they are applicable only to images with minor differences in imaging position and abundant acoustic textures. Meanwhile, optical images enable 6-DOF pose estimation for close-range subsea structures. Pose estimation has been primarily based on template matching [3] and edge-based matching [4]. However, it is hard to expect they would behave robustly against distant objects because of light attenuation and turbidity in the underwater environment. Localization correction using an optic-acoustic stereo system allows the intervention robots to have a more practical localization system suitable for intervention tasks than using a single sensor [5]. However, it is still regarded as a challenging issue because both sensors have different measurement characteristics and have sensing limitations in the underwater environment. Therefore, efficient strategies to use both sensors for localization correction need to be further explored.

This study proposes a novel localization technique using an optic-acoustic stereo system for underwater intervention robots. This technique is based on an inertial navigation system that uses optic-acoustic images for localization correction to enable near-far range localization from subsea platforms. The merits of the proposed technique were verified experimentally using real data obtained from a test tank.

Localization strategy: The proposed localization system aims to estimate the 3-D positional relationship between an intervention robot and subsea structures. It is based on an inertial navigation system that uses an extended Kalman filter. In addition, It uses relative poses to the structures obtained using an optic-acoustic stereo system for localization correction. Pose estimation is performed sequentially for each sensor, and each estimation result is selectively fused to the inertial navigation through a reliability evaluation. The subsea structures are assumed to be designed and deployed by humans, and their shapes and layouts are entirely known. Consequently, it enables pose estimation based on image alignment using an sonar simulator for a single-shot acoustic image. In addition, it enables robust visual tracking and pose estimation using 3-D wireframe models of the structures for an optical image.

Pose estimation with acoustic image alignment: Pose estimation for acoustic images consists of image-pose pair group generation, image-pose pair nearest-neighbor (NN) search, and image alignment. First, the group generation builds a database composed of pose hypotheses and their simulated images over a bounded localization area using an sonar simulator. Next, the NN search finds an image-pose pair in the database most similar to the input image and then returns its pose label. Finally, the image alignment matches the input image and the image generated in the continuous pose space of the simulator to determine the position where the input image was actually taken.

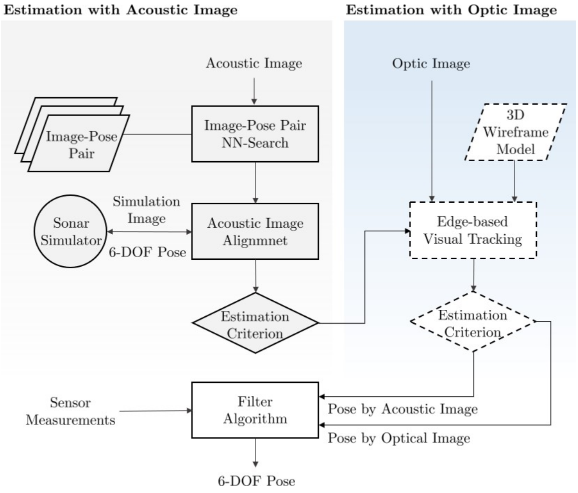


Fig 1 *Schematic diagram of the proposed localization system: Gray solid line boxes indicate components related to the pose estimation for the acoustic image, and dashed line boxes indicate the pose estimation for the optical image.*

The group generation samples a certain number of pose hypotheses over a 3-D pose space in a bounded localization area and generates simulated acoustic images for each pose hypothesis. For example, N sampling for each degree of freedom of a 3-D pose produces 6^N pairs. In this paper, we considered 3^N pairs, assuming that the navigation sensors provide the roll, pitch, and altitude information of the camera. The image-pose pair NN search finds a pair in the group that is most similar to the input image I_t at time t. The NN search consists of two steps: pair clustering and pair selection. Pair clustering is an offline process that clusters pairs using image hashing, minhashing which is a data dimensionality reduction method, and locality-sensitive hashing (LSH) which is a hash-based clustering and search technique. First, it creates a $n \times m$ hash table containing m poses and their n dimension binary hashes of images. Subsequently, minhashing uses random permutations to transform the hash table into a $r \times m$ signature matrix that maintains the Jaccard similarity between hashes. Then, LSH divides the signature matrix into b bands and assigns similar pose labels (or keys) to the same bucket. Pair selection is an online process that assigns an input image to a bucket and chooses a pair with the smallest image distance in the bucket. Similar to pair clustering, it sequentially applies image hashing, minhashing, and LSH to the input image to search a bucket. The most similar pair is then found in the bucket using the correlation ratio, which is a similarity measure suitable for multimodality data comparison. The pose label of the pair was used as an initial for image alignment.

The initial poses of the NN search are in discrete pose spaces and may differ from the actual poses. Therefore, it needs a procedure to find the actual pose from the initial pose in a continuous pose space, assuming that the actual pose is near the initial pose. Hence, we employed an image alignment technique that aligns the input and the simulation image to find more accurate pose in continuous pose spaces. The image alignment problem can be formulated as follows:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}}\, d\left(S(\mathbf{x}), I_t\right). \tag{1}$$

The equation determines the 6-DOF pose $\hat{\mathbf{x}}$ at which the sonar simulator produces a simulated image $S = [S_r, S_\psi]$ that minimizes the image distance $d(\cdot)$ with respect to the input acoustic image $I^t = [I_r, I_\psi]$, where the subscripts r and ψ denote two acoustic images each with different imaging geometries: x-y and r-azimuth. The number of parame-

Algorithm 1: Acoustic Image Alignment

```
Input: $\tilde{\mathbf{x}}, I^t = [I_r, I_\psi]$
Result: $\hat{\mathbf{x}}$
Param: $K, N, \zeta, \mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3]$
$\mathbf{x}_1 \longleftarrow \tilde{\mathbf{x}}, F_I \longleftarrow$ FeatureExtraction $(I^t)$
for $k \leftarrow 1$ to $K$ do
    for $n \leftarrow 1$ to $N$ do
        $\lambda_n \longleftarrow$ BidirectionalLineSearch $(\mathbf{x}_k, \mathbf{d}_n, I^t)$
        $\mathbf{x} \longleftarrow \mathbf{x}_k + \lambda_n \mathbf{d}_n$
    end
    $\mathbf{d}_k \longleftarrow$ SetSearchDirection $(\mathbf{x}_k, \mathbf{x})$
    $\lambda_k \longleftarrow$ BidirectionalLineSearch $(\mathbf{x}_k, \mathbf{d}_k, I^t, F_I)$
    $\mathbf{x}_{k+1} \longleftarrow \mathbf{x}_k + \lambda_k \mathbf{d}_k$, $\hat{\mathbf{x}} \longleftarrow \mathbf{x}_{k+1}$
    $\mathcal{E} \longleftarrow |\mathbf{x}_k - \mathbf{x}_{k+1}|$
    if $\mathcal{E} < \zeta$ then
        Return$(\hat{\mathbf{x}})$
    end
end
```

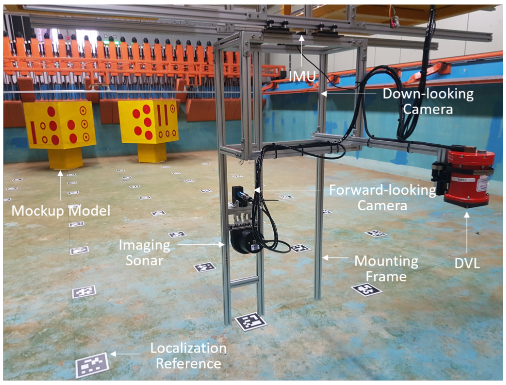


Fig 2 *Experimental setup for data acquisition: Sensors, mounting frame, two mock-up models, and markers for localization reference were deployed in the water tank.*

ters to be estimated can be varied according to the sensor configuration and localization strategy. We only considered the 3-DOF pose, ignoring the camera's roll, pitch, and vertical motion. In addition, this study only considered image alignment for acoustic images with multimodal relationship, sparse acoustic texture, and no significant difference in imaging position. To this end, we consider an image distance $d(S_*(\mathbf{x}), I_*) = C(S_*(\mathbf{x}), I_*) + \beta \cdot M(S_*(\mathbf{x}), I_*)$, which is the weighted sum of the correlation ratio M, which can measure pixel dependence for multimodality imaging data, and the Chamfer distance C, which indirectly measures the geometric difference even for images with sparse textures. The Chamfer distance is calculated for feature points clustered with a positive gradient on the range axis of the acoustic image, which is known as a representative feature point representing the outline of an object in the acoustic image. Assuming a static environment, we can directly estimate the camera heading motion by measuring the pixel motions on the azimuth axis of the r-azimuth acoustic image. Therefore, the resulting image distance is the weighted sum of the distances between the two images, $d(S(\mathbf{x}), I_t) = \alpha \cdot d(S_r(\mathbf{x}), I_r) + (1 - \alpha) \cdot d(S_\psi(\mathbf{x}), I_\psi)$.

Algorithm 1 summarizes the procedure of the parameter estimation of image alignment. This takes an initial $\tilde{\mathbf{x}}$ and an input image I^t and then returns an estimated pose $\hat{\mathbf{x}}$. K and N represent the number of iterations for the parameter estimation and line search, respectively. The latter corresponds to the degree of freedom of the pose that we want to estimate, and it was set to three here. The vector **d** is a unit vector that represents each search direction of the pose to be estimated. It is used for a line search based on the golden section search in *BidirectionalLineSearch*, and the line search exists in the inner and main loops. The first for the inner loop determines the direction of acceleration of the parameter updates, and the second for the main loop determines the step size of update acceleration λ. A criterion is used to determine whether pose estimation is to be used for filter updates. It is calculated as the ratio of the two correlation ratios $C(S_r(\mathbf{x}), I_r^t)/C(S_r(\mathbf{x}), G(S_r(\mathbf{x})))$, where $C(S_r(\mathbf{x}), G(S_r(\mathbf{x})))$ is the correlation ratio between a simulated image and the simulated image with Gaussian smoothing $G(\cdot)$. $C(S_r(\mathbf{x}), I_r^t)$ is the correlation ratio between the simulation and input images. The estimation is excluded from the filter update if the criterion is less than the threshold value.

Edge-based pose estimation using particle filtering: Pose estimation for optical images is based on visual tracking technique via geometric particle filtering on the Lie group with a 3-D wireframe model [6]. State $\mathbf{x}_t$ at time t is represented as $\mathbf{x}_t = \mathbf{x}_{t-1} \cdot \exp(\mathbf{A}_{t-1} + \mathbf{dW}_t \sqrt{\Delta t})$, where state $\mathbf{x}_t \in SE(3)$ is in the Lie group. $\mathbf{A}_{t-1} = \lambda_a \log(\mathbf{x}_{t-2}^{-1} \mathbf{x}_{t-1})$ is the first-order autoregressive (AR) state dynamics. $\mathbf{dW}_t$ is the Wiener process noise in $\mathfrak{se}(3)$ with covariance $\mathbf{\Sigma}_w \in \Re^{6\times6}$. The particles are evaluated using the measurement likelihood. A 3-D wireframe CAD model for each pose particle is projected onto the image plane. The moving edge (ME) algorithm then finds the edge pixels of the object in the image [7]. After the ME search, the residual vector $\mathbf{r} = [r_1, \cdots, r_{N_z}]^T$ is determined as the Euclidean distances between the sample and candidate points searched by the ME algorithm. The resulting measurement likelihood is defined with the residuals and the ratio of matched points as follows:

$$p(\mathbf{z}_t \mid \mathbf{x}_t) \propto \exp\left(-\lambda_v \frac{N_p - N_i}{N_p}\right) \exp\left(-\lambda_r \bar{\mathbf{r}}\right) \qquad (2)$$

where N_p is the number of sample points in the CAD model and N_i is the number of matched points in the ME algorithm. λ_r and λ_v are tuning parameters that control the sensitivity of each term. The followings are the estimation criterion $\lambda_v \frac{N_p - N_i}{N_p} < \epsilon_1$ and $\lambda_r \bar{\mathbf{r}} < \epsilon_2$.

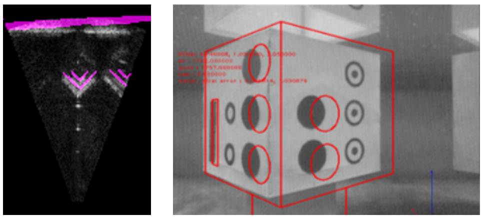
Filter system: The filter system uses a 6-DOF kinematic model of a vehicle using inertial sensor measurements [8]. $\mathbf{x}_v = [x\, y\, z\, \phi\, \theta\, \psi\, u\, v\, w]^T$ is the state vector, where x, y, and z are the positions; ϕ, θ, and ψ are the Euler angles in the global frame; and u, v, and w are the linear velocities in the vehicle-fixed frame. The vehicle pose in the global frame is obtained by $\mathbf{z}_{imu} = [z_{\dot{u}}\, z_{\dot{v}}\, z_{\dot{w}}\, z_{\dot{p}}\, z_{\dot{q}}\, z_{\dot{r}}]^T$ from the IMU, where $z_{\dot{u}}$, $z_{\dot{v}}$, and $z_{\dot{w}}$ are the linear accelerations, and z_p, z_q, and z_r are the angular velocities. The resulting state vector $\mathbf{X} = [\mathbf{x}_v\ \mathbf{x}_m]^T$ is composed of the state of the subsea structure $\mathbf{x}_m = [x_m\, y_m\, z_m\, \phi_m\, \theta_m\, \psi_m]^T$ and vehicle state $\mathbf{x}_v$. The system dynamics have motion models of the vehicle and subsea structure, which can be represented as

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}_v \\ \dot{\mathbf{x}}_m \end{bmatrix} = \begin{bmatrix} f(\mathbf{x}_v, \mathbf{z}_{\text{imu}}) \\ 0 \end{bmatrix} + \mathbf{w} \qquad (3)$$

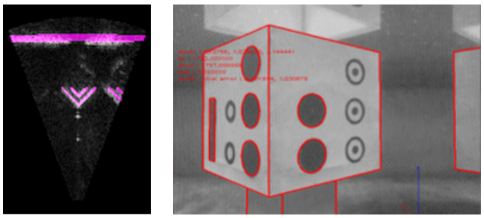
where $f(\mathbf{x}_v, \mathbf{z}_{\text{imu}})$ denotes the motion model [8]. Here, $\mathbf{w}$ denotes the uncertainty assumed to follow a zero-mean Gaussian distribution. Because the subsea structure is assumed to be stationary, the time derivative $\dot{\mathbf{x}}_m$ of the state vector is set to zero.

The measurement model employs two measurements for filter update. The first is the motion measurement, $\mathbf{z}_v = [u_d\ v_d\ w_d\ d_d]^T$, composed of the linear velocity (u_d, v_d, w_d) and altitude d_d obtained from the DVL. In addition, it contains the relative pose between the subsea structure and the vehicle: $\mathbf{z}_m = [\ z_x\ \ z_y\ \ z_\psi\ ]^T$ from a pose estimation for the acoustic image, or $\mathbf{z}_m = [\ z_x\ \ z_y\ \ z_z\ \ z_\phi\ \ z_\theta\ \ z_\psi\ ]^T$ from the pose estimation for the optical image. Motion measurement can be represented as $\mathbf{z} = [\ \mathbf{z}_v\ \ \mathbf{z}_m\ ]^T + \mathbf{v}$. $\mathbf{v} = [\ \mathbf{v}_v\ \ \mathbf{v}_m\ ]^T$ is for the measurement noise, where $\mathbf{v}_v$ is the motion measurement, which is assumed to follow a zero-mean Gaussian distribution with covariance $\mathbf{R}_v$, and $\mathbf{v}_m$ is for the pose estimation, which is assumed to follow a zero-mean Gaussian distribution with covariance $\mathbf{R}_m$. $\mathbf{v}_m$ and $\mathbf{R}_m$ are modeled differently with different scale factors according to what a pose estimation is carried out.

Experimental Results: The proposed localization techniques were validated under a scenario in which an intervention system maneuvers near a cluster of subsea platforms for their tasks. The validation used real data consisting of optical images, acoustic images and sensor measurements (DVL and IMU) obtained by moving a mounting frame along a rectangular path in the test tank. The imaging plane of the imaging sonar was



(a) NN search *(b) Model projection*



(c) Alignment *(d) Edge matching*

Fig 3 *Pose estimation results: Image alignment results are shown in (a) and (c), respectively. The 3-D wireframe model is projected onto the image plane using the initial pose of image alignment (b), and the model is precisely aligned to the object through particle filtering (d).*

Table 1. Performance of two pose estimations: RMSE, success rate, and convergence time in mean. N, N_a, and N_o are the number of input acoustic images, success pose estimation for input acoustic images, and success pose estimation for optical images, respectively.

	x (m)	y (m)	yaw (°)	Success rate	Time (sec)
Acoustic	0.271	0.212	3.38	0.15 (N_a/N)	4.21
Optical	0.192	0.217	2.93	0.82 (N_o/N_a)	1.19

fixed to be parallel to the bottom of the tank. Two scaled mock-up models of the control panel of a wellhead, an offshore platform for oil or gas production, were built and placed on the bottom of the test tank.

Fig. 3(a) shows superimposed images between the input and the simulation images founded by the NN search. The success criterion for NN search is when the retrieved pose has a position error of less than 0.4 m and an azimuth error of less than 5 degrees. The performance of NN searches was evaluated along with this criterion. The NN search shows higher accuracy when the camera and the mock-up model are close. The pose accuracy was increased as the sampling size increased. Fig. 3(c) presents superimposed images by the image alignment. The image alignment was evaluated with the pose error and convergence time as shown in Table 1. Likewise, the imaging distortion increased the pose error for images representing distant objects. However, the overall performance was acceptable to be used for the localization correction of the intervention robots. Besides, the image alignment converges slower than the image update rate of the actual imaging sonar, but it could be improved by adjusting the trade-off between accuracy and speed parameters. Fig. 3(b) and (d) show the results of the edge-based visual tracking. Since optical images are subject to light attenuation and distortion, pose estimation was mostly done at the close range of the mock-up model. Since the pose estimation directly estimates the 6-DOF pose for the mock-up model, it has a clear advantage to the operation of the intervention robots that require accurate 3-D positional information for their tasks. Table 1 shows the performance of two pose estimations with their RMSE, success rate, and convergence time.

The proposed localization system was compared with the integrated IMU-DVL system and ground-truth system, each of which employs different information for localization corrections with the same system model. The integrated IMU-DVL system uses DVL measurements, and ground truth uses DVL measurements and the relative pose to markers placed at the bottom of the test tank. Fig 4 shows the paths generated by each localization system. Red lines indicate the paths generated by DVL-IMU and the proposed localization system. The blue line represents the path obtained using the ground truth system. The two squares represent the mockup model, whose shape and layout were the same as those of the experimental setup. Meanwhile, the circle marker is the position where pose estimation performed localization correction. Because acoustic and optical images are subject to distortion for distant objects, most pose estimation was performed near the mockup model. The pose estimation for distant objects has slightly more errors, but it will provide useful information for mid-range localization. Table 2 summarizes RMSE of tow localization systems. This shows that a reliable localization system for intervention robots has been achieved.

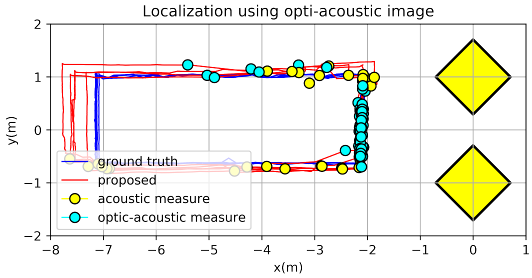


Fig 4 *Localization results of proposed localization system*

Table 2. RMSE of two localization systems: The results are evaluated for the entire path and a specific localization area within a distance of 3 meters from the mock-up model, respectively

		Entire path			**within 3 meters**		
	Length (m)	x (m)	y (m)	yaw (°)	x (m)	y (m)	yaw (°)
IMU+DVL	52.28	0.41	0.32	4.22	0.38	0.4	3.71
Proposed	51.15	0.29	0.24	2.16	0.06	0.04	1.83

Conclusion: In this study, an underwater localization algorithm was developed using an optic-acoustic stereo system for underwater intervention robots. The feasibility and usability were verified using real data obtained from a test tank. This study implemented global localization for a single-shot acoustic image and expanded its usability with edge-based tracking for high-precision localization near subsea structures. In the experiment, the proposed localization system showed reliable accuracy for the intervention system to maneuver near offshore platforms and perform autonomous intervention tasks without the need for external beacons for positioning, significantly improving the overall efficiency and reliability of underwater operations.

References

1. Shin, Y., et al.: Bundle adjustment from sonar images and slam application for seafloor mapping. In: OCEANS 2015 - MTS/IEEE Washington, pp. 1–6. (2015)
2. Santos, M.M., et al.: Underwater place recognition using forward-looking sonar images: A topological approach. Journal of Field Robotics 36(2), 355–369 (2019)
3. Palomeras, N., et al.: Toward persistent autonomous intervention in a subsea panel. *Autonomous Robots* 40(7), 1279–1306 (2016)
4. Park, J., Kim, T., Kim, J.: Model-referenced pose estimation using monocular vision for autonomous intervention tasks. Autonomous Robots 44(2), 205–216 (2020)
5. Marani, G., Choi, S.K., Yuh, J.: Underwater autonomous manipulation for intervention missions AUVs. *Ocean Engineering* 36(1), 15–23 (2009)
6. Choi, C., Christensen, H.I.: Robust 3D visual tracking using particle filtering on the special euclidean group: A combined approach of keypoint and edge features. *The International Journal of Robotics Research* 31(4), 498–519 (2018)
7. Bouthemy, P.: A maximum likelihood framework for determining moving edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(5), 499–511 (1989)
8. Kim, T., Kim, J.: Nonlinear filtering for terrain-referenced underwater navigation with an acoustic altimeter. In: *MTS/IEEE OCEANS*, pp. 1–6. (2014)