Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

# Imaging-sonar-based Underwater Object Recognition Utilizing Object's Yaw Angle Estimation with Deep Learning ⋆

**Minsung Sung** * **Meungsuk Lee** * **Byeongjin Kim** *
**Son-Cheol Yu** *

* *Department of IT Engineering, Pohang University of Science and Technology, Pohang, Republic of Korea*
*(e-mail: {ms.sung, meungsuklee, kbj0607, sncyu}@postech.ac.kr)*

**Abstract:** This paper proposes a method to recognize underwater target objects and estimate their yaw angle using an imaging sonar. First, a light sonar simulator generated template images of the target objects from various viewing angles. Next, a generative adversarial network predicted a semantic map by segmenting the real sonar image for reliable recognition. Then, matching the template images and semantic map identifies the target object and its yaw angle. We verified the proposed method by installing objects in the indoor water tank. The proposed method can provide relative pose information of sensing platforms which is useful for pose control and navigation.

*Keywords:* GAN, Object Detection, Segmentation, Sonar Image, Sonar Simulator

## 1. INTRODUCTION

Underwater object recognition is essential to automate various underwater missions (Kim et al. 2018b, Sualeh and Kim 2019, and Kim et al. 2016). Autonomous underwater vehicle (AUV) and sonar sensors are widely used platform for the underwater object recognition (Joe et al. 2019, Pyo et al. 2017, Cao et al. 2018, and Maki et al. 2019). The AUV can cover a large area and cope with harsh underwater environments. Sonar sensors have a wide scanning range and tolerance in a turbid stream.

When using these platforms, estimating a yaw angle of underwater objects is a crucial problem for reliable recognition (Yu 2008). In a sonar image, shapes of the objects change significantly according to the viewing angles. The AUV can measure the roll, pitch, and range to the object using an inertial measurement unit and depth sensor, however, the angle that the AUV encounters the object is unpredictable.

Therefore, the AUV can detect the underwater target objects more reliably after estimating the yaw angle of an object. Performing cross-validations for multiple angles can improve recognition accuracy (Kim et al. 2019). Moreover, the estimated relative angles between the AUV and the underwater landmarks can be applied for various AUV operations such as recovery, pose control, and navigation(Ribas et al. 2008, Muñoz-Vázquez et al. 2017, and Pyo and Yu 2019).

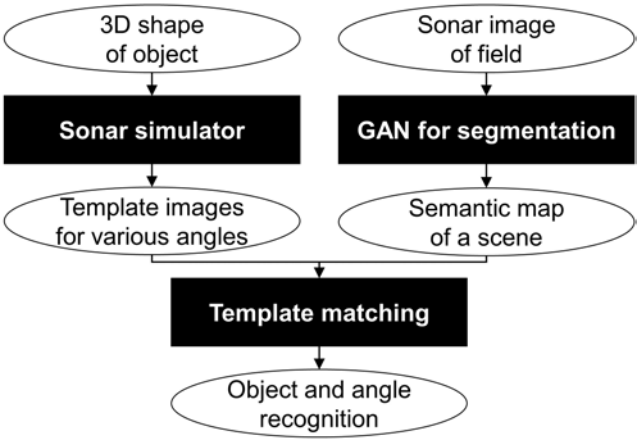



Fig. 1. Pipeline of the proposed object and angle recognition method.

We herein propose a method to recognize the underwater known-shape target object and estimate its yaw angle using a sonar simulator and generative adversarial network (GAN), as shown in Fig. 1. The shape of objects in a sonar image is sensitive to the viewing angles. Moreover, sonar sensors have a low signal-to-noise ratio (SNR). Therefore, sonar-based object recognition is challenging. If we know all shapes according to the yaw angle in advance, identifying class and angle of objects is possible (Cho et al. 2015). We implemented a ray-tracing-based sonar simulator to synthesize template images of target objects for various angles rapidly. We then preprocessed the real sonar images using GAN-based segmentation to improve the robustness of the recognition. Finally, we can

⋆ This research was a part of the project titled 'Gyeongbuk Sea Grant', funded by the Ministry of Oceans and Fisheries, Korea. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2017R1A5A1014883). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICT Consilience Creative program(IITP-2019-2011-1-00783) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

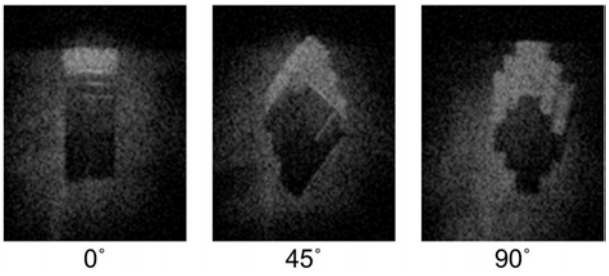



Fig. 2. Shapes of the same object changing according to the angles.

identify the target object and its yaw angle by matching the template image with the segmented sonar images.

This paper is organized as follows: Section 2 explains the proposed sonar simulator, GAN, and template matching for the object and yaw angle recognition. In Section 3, we describe the experiments to develop the proposed method. Section 4 presents the recognition results. The paper ends with the conclusion in Section 5.

## 2. PROPOSED METHOD

Two causes can make imaging-sonar-based underwater object recognition challenging. First, the shape of the object changes significantly depending on the viewing angle in the sonar image like Fig. 2, but in what direction the AUV approaches the underwater object is hard to predict. Next, the sonar image has low SNR.

We proposed a method to recognize the target object accurately and further identifies its yaw angle by simulating all the shapes of the target object according to angles and matching the template images with the sonar images of an underwater scene. We also removed the degradation effects and improved the matching reliability by segmenting the sonar images using GAN. This section describes three elements of the proposed method; sonar simulator to synthesize template images, GAN to preprocess the sonar images, and template matching.

### *2.1 Ray-tracing-based Template Image Simulation*

The proposed method requires prior information about the shape of the target object in the sonar images according to the angles. Therefore, we developed a sonar simulator which can generate template images of target objects rapidly, instead of taking sonar images manually.

The sonar simulator emulated the imaging mechanism of the sonar sensor based on the ray tracing. It first calculated the reflection point at which the acoustic beam collided with an object as follows:

$$\overrightarrow{p_\theta} = \frac{\overrightarrow{N} \cdot \overrightarrow{p_1}}{\overrightarrow{N} \cdot \overrightarrow{v_\theta}} \overrightarrow{v_\theta}, \tag{1}$$

where $\overrightarrow{v_\theta}$ is a unit direction vector of the acoustic wave transmitted to azimuth angle $\theta$, $\overrightarrow{N}$ is the normal vector of the object surface, $\overrightarrow{p_1}$ is a position vector of the object surface.

We then calculated the intensity of the echo. In the real world, many parameters, such as salinity, temperature, and beam pattern, affect the intensity of acoustic beams. However, in the proposed method, fast processing speed is essential to generate a large number of template images for various angles. Moreover, the proposed method using template matching does not require photo-realistic images. To reduce the computational complexity, we modeled only transmission loss and calculated the semantic information such as shape and position of highlights and shadows, as follows (Etter 1995 and Kim et al. 2018a):

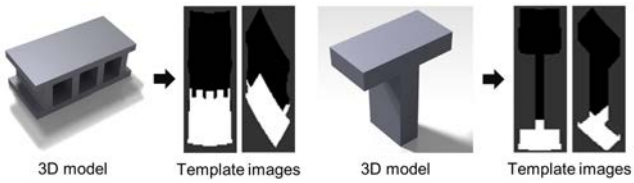



Fig. 3. Simulated template images given 3D model of object.

$$I = k \cdot \frac{I_0}{\left|\overrightarrow{p_\theta}\right|^2}, \tag{2}$$

where $k$ is a unit conversion constant, and $I_0$ is the initial intensity of the acoustic wave.

Finally, by mapping the intensity of the echo to the corresponding pixel $(|p_\theta|, \theta)$ and cropping the object region, the sonar simulator synthesized the template images for various angles like Fig. 3, given a three-dimensional (3D) model of the objects. The proposed simulator can accurately calculate the shape of the highlight and shadow of an object, which is essential information for template matching, in a short time.

### *2.2 GAN-based Sonar Image Segmentation*

We proposed the segmentation of a real sonar image to recognize the target object more reliably. The real sonar image has various degradation effects. On the other hand, the sonar simulator calculates the semantic information such as the shape of highlights and shadows. Therefore, segmentation removing degradation effects and generating semantic maps can improve the accuracy of the target recognition in the sonar image.

For the segmentation of the sonar image, we employed the pix2pix model (Isola et al. 2017) based on deep learning. Pix2pix is a generative model that can translate an input image into a target domain. We can segment sonar images by translating the images into a domain which only has highlights, shadows, and backgrounds using this model. This approach can accurately segment the sonar image of a more general scene. Classical methods of segmentation, including thresholding and boundary detection, do not show good results due to the high noise and low resolution of a sonar image. A discriminative model such as a convolutional neural network (CNN) is another deep-learning-based segmentation approach by classifying each pixel, but it can fail to segment when the image contains objects not included in the training data.

One of the difficulties in using deep-learning-based methods is the lack of a training dataset. The training dataset for the segmentation requires much time because each

pixel in each image should be annotated. The implemented sonar simulator can calculate the semantic map labeling each pixel of the underwater scene. Therefore, the proposed method can utilize deep learning easily.

The network consists of a generator and a discriminator. The generator is U-Net (Ronneberger et al. 2015) with 15 layers. U-Net has an encoder-decoder architecture, which can make an accurate semantic map of a scene preserving contextual information of input image. Moreover, the U-Net has skip-connections which copy the feature map of $n^{th}$ layers when the $(16-n)^{th}$ layers restore the image. Therefore, the network can localize the position of the features more accurately in the generated semantic map. Because an advantage of the sonar sensor is to provide accurate range and azimuth information of the scene, U-Net, which can localize features using the skip-connections, is proper for the segmentation of sonar images.

The discriminator checks whether an input image is a target-domain-like image synthesized by the generator or a real image of the target domain. Accurate classification makes the generator produce more target-domain-like images to deceive the discriminator. Therefore, we employed a CNN for the discriminator that has shown outstanding performance in classification. The discriminator has four convolutional layers, and it observes an input image in patch units, allowing the generator to represent the detail better.

We then designed loss function as below, so that the network can handle more general scene.

$$\begin{aligned} Loss_{GAN}(G,D) &= \mathbb{E}_{x,y}[\log D(x,y)] \\ &\quad + \mathbb{E}_{x,z}[1-\log D(x,G(x,z))] \\ &\quad + \lambda \mathbb{E}_{x,y,z}[||\mathcal{N}(y-x)-\mathcal{N}(G(x,z)-x)||_1], \end{aligned} \quad (3)$$

where $G$ is the generator, $D$ is the discriminator, $x$ is the given sonar image, $y$ is the label, $z$ is a random input vector, and $\mathcal{N}(x)$ is a normalize function that maps $x$ to $[-1,1]$. The last term of the loss function makes the network to focus on degradation effects added to the semantic information. As a result, it helps the network handle sonar images of a more general scene.

### *2.3 Template Matching for Object and Angle Recognition*

The proposed method finally identifies the target object and its angle by matching the template image with the segmented image. We employed the two-dimensional (2D) discrete cross-correlation for template matching. 2D discrete cross-correlation measures the similarity between two images based on the distribution of pixel values. We set the window of the same size as the template image in the segmented image, and calculate $R_\theta$, which indicates the similarity between a sonar image and the template image of angle $\theta$, as follows:

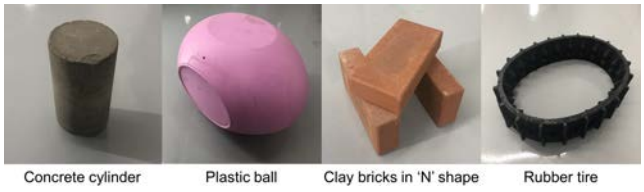



Fig. 4. Objects used to construct training data of GAN.

$$R_\theta = \max_{i,j} \left[ \frac{\sum_{m,n}[S(m+i,n+i)-\bar{S}][T_\theta(m,n)-\bar{T_\theta}]}{\sqrt{\sum_{m,n}[S(m+i,n+i)-\bar{S}]^2 \sum_{m,n}[T_\theta(m,n)-\bar{T_\theta}]^2}} \right], \quad (4)$$

for $0 \le i \le M-Mt$, $0 \le j \le N-Nt$, $1 \le m \le M_t$, $1 \le n \le N_t$, where $i$ and $j$ is pixel coordinate of the window, $M$ and $N$ are the size of the input sonar image, $M_t$ and $N_t$ are the size of the template image, $S$ is a given sonar image, $T_\theta$ is a template image of the target object for angle $\theta$, $\bar{S}$ and $\bar{T_\theta}$ are the mean of the pixel intensity of each image.

The $R_\theta$ shows the probability that an area similar to the target template image exists in the sonar image. So, we can recognize the target object checking how close this value to one.

We can also identify the yaw angle of the object by checking which angle template image has the highest correlation with the segmented image, as follow:

$$obj_\theta = \arg\max_\theta \{R_\theta\}, \quad (5)$$

when $\max_\theta\{R_\theta\}$ exceeds a predefined threshold value.

## 3. DEVELOPMENT OF THE NETWORK

### *3.1 Constructing Training Dataset*

The GAN to segment sonar images should be trained first to develop the proposed method. Training of the GAN requires image pairs composed of a real sonar image and their corresponding semantic map.

Constructing the training dataset with various types of sonar images helps make the GAN handle the sonar image of the more general scene. We captured sonar images in a small water tank focusing on two points. First, we used objects of various shapes and materials, such as concrete cylinder, plastic ball, and clay bricks, like Fig. 4. Next, we translated and rotated the object to various positions and angles. We also changed the tilt angle of the sonar sensor. For the sonar sensor to capture the images, we used the dual-frequency identification sonar (DIDSON) (Belcher et al. 2002). Tables 1 and 2 explains the capturing environment and the specifications of the DIDSON, respectively.

In order for the GAN to learn to segment a given sonar image, the GAN requires a segmentation label for the captured image. We modeled 3D shapes with the same shape and dimension as the objects used in the indoor water tank experiments. Then, we generated the label semantic map by simulating the images under the same

Table 1. Settings to capture training images

| Parameter | Value |
|---|---|
| Water tank size | 1.35 m x 3 m x 1.7 m (width x length x height) |
| Sonar position | 1.6 m, 0 m, 0.9 m (x, y, z) |
| Sonar tilt | 15 °, 20 °, 25 ° |
| Object translation | 0 m, 0.15 m (in the x-y plane) |

Table 2. Specifications of DIDSON

| Parameter | Value |
|---|---|
| Acoustic beam frequency | 1.8 MHz |
| Range field of view | 12 m |
| Azimuth field of view | 29 ° |
| Vertical beam angle | 14 ° |
| Image size | 512 x 96 |
| Maximum resolution | 0.3 ° |
| Frame rate | 4-21 fps |

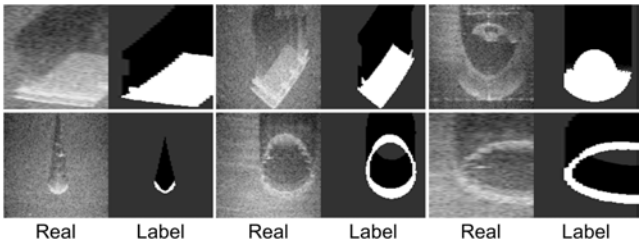



Fig. 5. Samples of training images pairs for GAN.

condition with the experimental setup. When developing the deep-learning-based segmentation, this process of creating segmentation labels is time-consuming because each pixel of the images should be annotated manually one by one. The developed sonar simulator could calculate the corresponding label automatically, reducing time to construct the dataset.

Finally, we manually cropped around the object for efficient training. We then applied the random scaling for data augmentation. Finally, by resizing the image patches to 128 by 128, we could construct a training data set consisting of 6,780 image pairs of real sonar images and its corresponding semantic map, like Fig. 5.

### *3.2 Training of the GAN for Segmentation*

We trained the GAN for 30 epochs with the constructed training dataset. The total training took about 1.2 hours when using single Graphics Processing Unit (GPU) Titan V. During the training, the generator and discriminator operated adversarially. As the generator generated a better output image, the discriminator got more accurate to distinguish between the real and generated images. Again, the more accurate the discriminator, the higher-quality image the generator produced to deceive the discriminator. Once the training is complete, the GAN generated a semantic map from a given sonar image. Then, the target object and its yaw angle can be identified by comparing the predicted semantic map with the prepared template images.

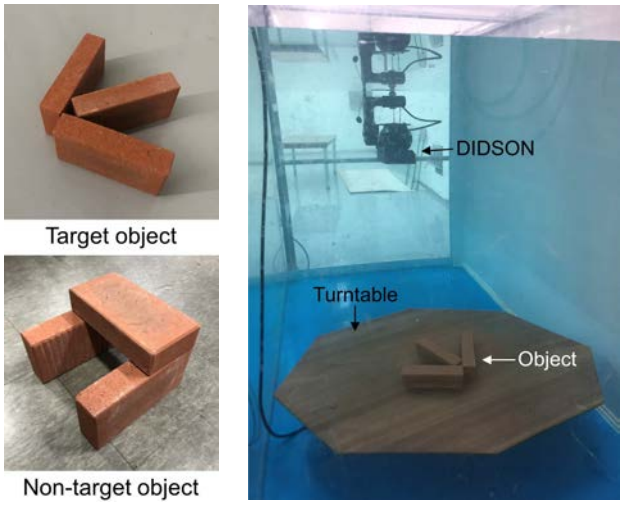
## 4. TARGET OBJECT AND YAW ANGLE RECOGNITION

### *4.1 Indoor water tank experiments*

We evaluated the proposed object and its yaw angle recognition method through indoor water tank experiments, as shown in Fig. 6. We used two objects to verify whether the proposed method can identify the target object. The target and non-target objects were made by piling three bricks into different shapes, like Fig. 6a. Moreover, to verify that the proposed object can recognize the yaw angle of the target object, we designed a turntable using a stepping motor and captured the sonar images rotating the object by ten-degree increments, as shown in Fig. 6b. We captured four sets of sonar images changing the tilt angle of DIDSON randomly. As a result, 144 sonar images were obtained for target and non-target objects, respectively. The proposed method can recognize the target object and its angle in these sonar images following the pipeline.

### *4.2 Template Image Simulation*

The proposed method first generated the template images of the target object using the simulator. We can determine how many template images to generate in degrees, depending on the desired resolution of the yaw angle estimation. We herein generated 36 template images in 10-degree increments. Fig. 7 shows samples of the template images.




(a) Target and non-target. (b) Experimental setup.

Fig. 6. Indoor water tank experiments to acquire test images for the proposed method.

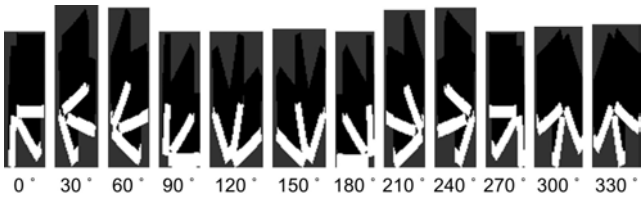



Fig. 7. Samples of simulated template images according to angles.

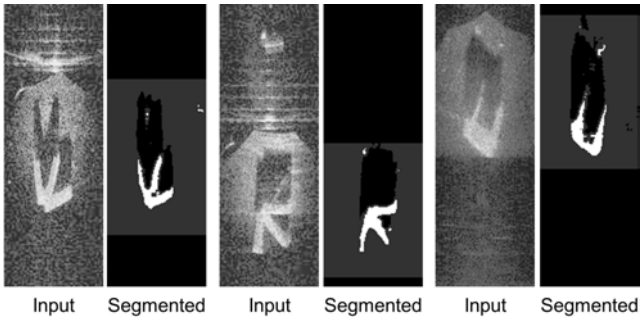



Fig. 8. Results of the GAN-based segmentation of the sonar images.

### *4.3 Sonar Image Segmentation*

Then, the trained GAN generated the semantic map of the input images like Fig. 8. For the quantitative analysis of the GAN-based pre-processing method, we measured the peak signal to noise ratio (PSNR) for one set of target sonar images. The PSNR increased by 10.36 dB from 25.57 dB to 35.93 dB after applying the proposed method. It shows that the degradation effects of the sonar images were removed effectively. Thus, we could extract more reliable information from the generated semantic map.

We also measured the segmentation intersection over union (IOU) of the generated semantic map. Since a sonar sensor can provide reliable information about the range and azimuth of underwater terrain, it is essential whether the proposed method can localize the semantic information of the scene well. The calculated segmentation IOU was 0.638. It means that more than two-thirds of the actual semantic information and the generated semantic maps overlapped, indicating that the generated semantic map represents the shape and position of the object well.

### *4.4 Recognition Results*

Finally, we can identify the underwater target object and its angle through template matching. Table 3 compares the correlation with the template image of the target sonar image and the non-target sonar images. The table shows the average correlation values in four test sets between the specific angle of sonar images and template images, and the thresholding method proposed by Otsu (Otsu 1979) was applied. The highest correlation is measured when the actual angle of the object and the angle of the template image match. The overall correlation values of the non-target object were lower than the value of the target object. As a result, the proposed method recorded the target and yaw angle identification accuracy of 92.01 % by setting the optimal threshold.

We then compared the results of the template matching before and after applying the segmentation to verify the effect of the proposed GAN-based preprocessing method like Table 4. The correlation values are much noisy when the GAN is not applied. A cause that the correlation is noisy before applying segmentation can be the low SNR of the sonar image. The proposed GAN can remove the degradation effects and generate semantic maps that contain only essential information of a scene. Therefore, the proposed method can obtain more reliable information from the images and improve the accuracy of the recognition.

Table 3. Matching results between the template images and the segmented images

| | | **Template Image** | | | |
|---|---|---|---|---|---|
| | | **0 °** | **90 °** | **180 °** | **270 °** |
| **Target** | **0 °** | **0.6837** | 0 | 0 | 0 |
| | **90 °** | 0 | **0.7697** | 0 | 0 |
| | **180 °** | 0 | 0 | **0.5776** | 0 |
| | **270 °** | 0 | 0 | 0 | **0.7162** |
| **Non-target** | **0 °** | 0.5715 | 0.4892 | 0.5201 | 0.5537 |
| | **90 °** | 0 | 0.5463 | 0 | 0 |
| | **180 °** | 0 | 0.5238 | 0.5328 | 0.4867 |
| | **270 °** | 0 | 0.4884 | 0.5134 | 0 |

Table 4. Matching results before applying the GAN-based segmentation

| | | **Template Image** | | | |
|---|---|---|---|---|---|
| | | **0 °** | **90 °** | **180 °** | **270 °** |
| **Target** | **0 °** | 0.5972 | 0.4570 | 0 | 0 |
| | **90 °** | 0.4480 | 0.6446 | 0 | 0.4597 |
| | **180 °** | 0.4689 | 0.4709 | 0.4726 | 0 |
| | **270 °** | 0.4398 | 0 | 0.4530 | 0.5735 |

## 5. CONCLUSION

This paper proposed a method to detect the target object from the underwater sonar image and further to estimate the yaw angle at which the object is placed through template matching and the GAN-based segmentation. The proposed method can recognize the target object and estimate its yaw angle in sonar images captured in the water tank. Moreover, the GAN-based segmentation can improve the accuracy of the proposed method. The estimated angle information is helpful to recognize the target object more reliably and can be applied for operations of a sensing platform such as localization and navigation.

## REFERENCES

Belcher, E., Hanot, W., and Burch, J. (2002). Dual-frequency identification sonar (didson). In *Proceedings of the 2002 Interntional Symposium on Underwater Technology (Cat. No. 02EX556)*, 187–192. IEEE.

Cao, X., Togneri, R., Zhang, X., and Yu, Y. (2018). Convolutional neural network with second-order pooling for underwater target classification. *IEEE Sensors Journal*, 19(8), 3058–3066.

Cho, H., Gu, J., and Yu, S.C. (2015). Robust sonar-based underwater object recognition against angle-of-view variation. *IEEE Sensors Journal*, 16(4), 1013–1025.

Etter, P.C. (1995). Underwater acoustic modeling: principles, techniques and applications.

Isola, P., Zhu, J.Y., Zhou, T., and Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Joe, H., Kim, J., and Yu, S.C. (2019). Sensor fusion-based 3d reconstruction by two sonar devices for seabed mapping. In *2019 the 12th IFAC Conference on Control Applications in Marine Systems*. Elsevier.

Kim, B., Kim, J., Cho, H., Kim, J., and Yu, S.C. (2019). Auv-based multi-view scanning method for 3-d recon-

struction of underwater object using forward scan sonar. *IEEE Sensors Journal*.

Kim, D., Shin, J.U., Kim, H., Kim, H., Lee, D., Lee, S.M., and Myung, H. (2016). Development and experimental testing of an autonomous jellyfish detection and removal robot system. *International Journal of Control, Automation and Systems*, 14(1), 312–322.

Kim, J., Sung, M., and Yu, S.C. (2018a). Development of simulator for autonomous underwater vehicles utilizing underwater acoustic and optical sensing emulators. In *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, 416–419. IEEE.

Kim, T., Kim, J., and Byun, S.W. (2018b). A comparison of nonlinear filter algorithms for terrain-referenced underwater navigation. *International Journal of Control, Automation and Systems*, 16(6), 2977–2989.

Maki, T., Horimoto, H., Ishihara, T., and Kofuji, K. (2019). Autonomous tracking of sea turtles based on multibeam imaging sonar: Toward robotic observation of marine life. In *2019 the 12th IFAC Conference on Control Applications in Marine Systems*. Elsevier.

Muñoz-Vázquez, A.J., Ramírez-Rodríguez, H., Parra-Vega, V., and Sánchez-Orta, A. (2017). Fractional sliding mode control of underwater rovs subject to non-differentiable disturbances. *International Journal of Control, Automation and Systems*, 15(3), 1314–1321.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), 62–66.

Pyo, J., Cho, H., and Yu, S.C. (2017). Beam slice-based recognition method for acoustic landmark with multi-beam forward looking sonar. *IEEE Sensors Journal*, 17(21), 7074–7085.

Pyo, J. and Yu, S.C. (2019). Development of radial layout underwater acoustic marker using forward scan sonar for auv. In *2019 IEEE Underwater Technology (UT)*, 1–6. IEEE.

Ribas, D., Ridao, P., Tardós, J.D., and Neira, J. (2008). Underwater slam in man-made structured environments. *Journal of Field Robotics*, 25(11-12), 898–921.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sualeh, M. and Kim, G.W. (2019). Simultaneous localization and mapping in the epoch of semantics: A survey. *International Journal of Control, Automation and Systems*, 17(3), 729–742.

Yu, S.C. (2008). Development of real-time acoustic image recognition system using by autonomous marine vehicle. *Ocean Engineering*, 35(1), 90–105.