Robust Acoustic Marker Recognition in Underwater Environments Using Curriculum Learning

The University of Tokyo, ◯ Yixue ZHU, Yusheng WANG
Wakachiku Construction Co., Ltd., Hiroshi TSUCHIYA, Makoto HIRAOKA
The University of Tokyo, Qi AN, Atsushi YAMASHITA

I. Introduction

Underwater sensing is a challenging process, while optical cameras can be used to detect underwater environments, their resolution is significantly limited by dark and turbid conditions. In contrast, acoustic cameras, also known as forward-looking sonar (FLS), can capture high-resolution images even in murky water.

Fiducial markers play a vital role in underwater robot navigation, serving as reference points for precise localization and mapping. Several acoustic fiducial marker systems have been proposed. Wang *et al.* developed ACMarker, a simple designed fiducial marker system for sonar imaging, enabling pose estimation in underwater environments, using image processing techniques such as edge detection [1]. On the other hand, Norman *et al.*. developed AcTag for automatic marker family generation [2]. However, in real-world images test, the True Positive Rate (TPR) was found to be below 50%, limiting reliable detection in real-world underwater conditions. These limitations arises from multiple factors, including detection algorithm, FLS imaging principle, and noise in environment. DeepTag, a framework for fiducial marker detection in optical cameras [3], achieves high robustness in detection and pose accuracy. However, its application in underwater environments is limited due to the inherent differences between optical and acoustic imaging.

In this marker system, we proposed: 1) an unique fiducial marker design features 20 key points (16 central points and 4 corner points). 2) a recognition flow for acoustic markers in FLS images. 3) a two-stage curriculum learning strategy that incorporates inherent FLS characteristics, such as secondary reflections and cross-talk.

II. Method

A. Marker design

Acoustic images are created by analyzing backscattered intensity, whose quality is influenced by the properties of the object material. For example, concrete diffusely scatters sound waves, while stainless steel regularly reflects them, offering distinct contrasts in sonar images, as shown in Figure 1. The proposed marker measures 350 mm × 350 m × 2 mm (length × width × thickness) and is composed of a grid of large and small circular holes. The diameter of each large circle is 60 mm, while the diameter of the small circles in the four corners is 20 mm. The distance between adjacent large circles is 22 mm, and the distance from the small circles to the edges of the marker is 7 mm. The design allows for some positional tolerance in the placement of the circles, as the recognizability of the pattern (e.g., detecting the arrangement of large circles) serves as the key feature for marker identification. The bottom layer is made of concrete, offering a diffusely scattering base, while the top layer is constructed from stainless steel perforated with numerous circular holes, which are easily excavated to meet practical fabrication requirements, to create consistent reflective patterns. The marker is secured using nails, which can be placed at any location capable of fixing the marker firmly. This provides practical adaptability during underwater deployment. Therefore, this design is also suitable for long-term use in underwater structures.

B. Learning framework

First, a two-stage curriculum learning is designed to detect the marker's position, which gradually increases the complexity of training data, allowing the model to adapt to complex scenarios. After obtaining marker position information and using the ID information, the network is further trained for ID recognition. To predict bounding box (bbox) and IDs of markers, You Only Look Once (YOLO) [4], a one-stage object detection framework that identifies object class categories and bounding boxes in a single forward pass, is applied.

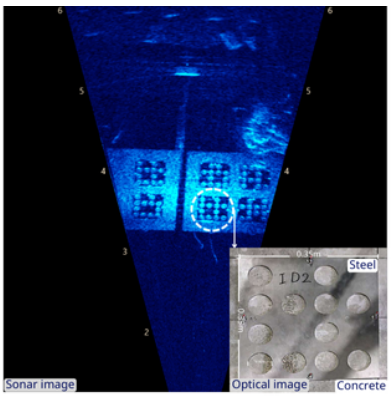

Fig. 1. Sonar and optical images of the proposed fiducial marker.

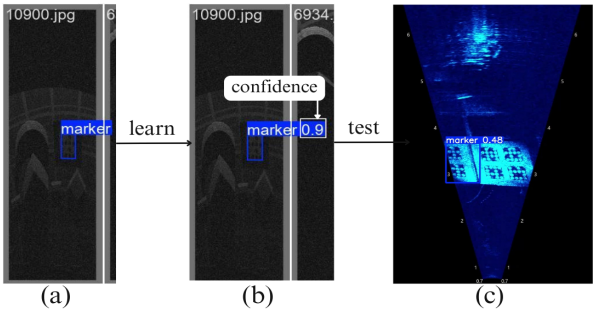

Fig. 2. Stage 1 for single marker detection. (a) a labeled image from the training dataset, where the marker's region is calculated. (b) training result including the predicted bounding box and confidence score for the marker. (c) test result on real-world images for marker detection.

Figure 2 illustrates the Stage 1 learning and validation process, using raw (r, θ) sonar images. Each training image contains only one marker and low Rayleigh noise. The FLS projection model maps objects to the (r, θ) plane, resulting in the loss of ϕ information. To simulate this imaging principle, the camera coordinates and marker coordinates from the simulator are used to calculate the actual (r, θ) values of the markers in the sonar images. The four corner points defining the region of interest (ROI) are transformed from the simulator's global coordinate system to the camera's local coordinate system through a coordinate transformation, taking into account the camera's position and orientation.

Figure 3 illustrates the learning process in Stage 2, incorporating real-world acoustic data augmented with multiple markers. Inherent FLS effects, such as secondary reflections and cross talk, are incorporated to simulate real-world underwater conditions. Our marker has a thickness of 2 mm, which contributes to secondary reflections, adding bright edges around the circular patterns of the marker. Cross talk arises because the sonar system emits ultrasonic beams simultaneously, which can interfere with each other when their reflected signals overlap or interact. The FLS images generated by the simulator are in (r, θ) format, which causes square markers to appear distorted. To address this, the (r, θ) images are first transformed into (x, y) images. The transformed fan-shaped images are then used for

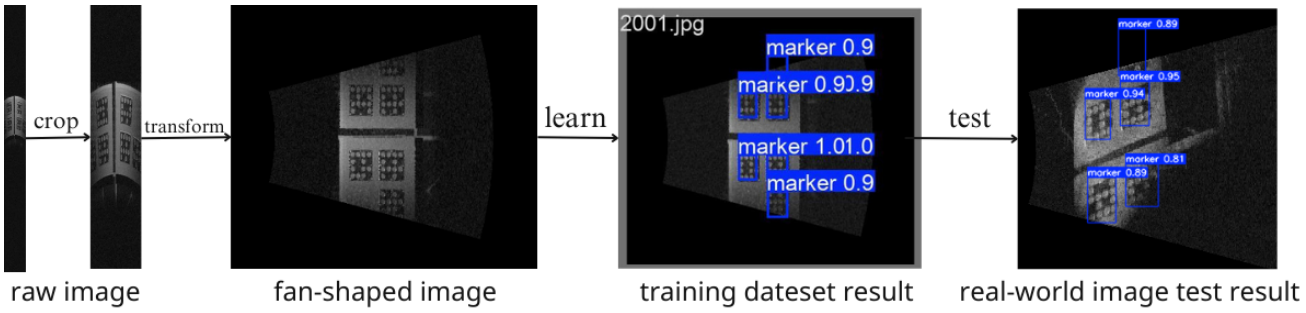

Fig. 3. Stage 2 for mutiple marker detection. The raw sonar images generated using Blender are first cropped and transformed into fan-shaped images. For learning, bounding box (bbox) labels is automatically calculated by the positional information from Blender. For testing, fan-shaped images converted from real-world data are used, producing predicted bboxes accompanied by confidence scores, indicating the reliability of the detection.

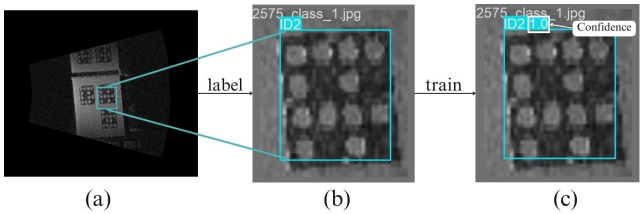

Fig. 4. ID detection process. (a) an image with predicted bounding boxes from Stage 2. (b) the cropped patch labeled with the corresponding ID information. (c) the model's prediction on training dataset.

TABLE I
PERFORMANCE METRICS FOR DIFFERENT STAGES

	Stage 1	**Stage 2**	
		w/o Stage 1	**with Stage 1**
True Positives ↑	256	2133	**2250**
False Positives ↓	14	24	**0**
Precision ↑	0.948	0.989	**1.000**
Recall ↑	0.109	0.905	**0.954**
Confidence (Avg.) ↑	0.48	0.89	0.89

training and testing.

Figure 4 illustrates the ID learning process. First, the bounding box information predicted in Stage 2 is used to crop the fan-shaped image containing multiple markers into patches, each containing a single marker. These patches are then assigned predefined IDs as classification targets. Finally, the model learns to classify the IDs based on these labeled patches, enabling accurate identification of individual markers.

III. Experiment

Simulator data generation was conducted using Blender 3.6.11, where markers were placed on two rectangular blocks to simulate their placement on underwater concrete structures. Variations were introduced in marker positions, material reflectivity, camera angles, noise levels, and effects such as secondary reflections to replicate real-world underwater conditions. These adjustments were tailored to different curriculum learning stages.

We employed the YOLOv11n model to ensure high efficiency and accuracy. For the original (r, θ) images, partially visible markers often result in out-of-bound errors, preventing effective learning. By transforming (r, θ) images into (x, y) images, the resulting transformed images include padded black edges, allowing all partially visible markers to be correctly labeled and included in the training process.

The results are summarized in Table I, demonstrating the performance metrics across different stages of the curriculum learning stages. For Stage 1, 8,000 images were used for training. The model achieved a mAP@0.5 of 0.995 and an F1 score of 1 at a confidence threshold of 0.856 in the training set, demonstrating the ability to identify almost all positive cases with high accuracy. However, on real-world FLS images from water tank experiments, the Stage 1 model exhibited limitations, including incorrect bounding boxes that did not align with the markers, highlighting the need for further improvements.

For Stage 2, the model was trained on a smaller dataset of 1,000 images and achieved a mAP@0.5 of 0.995 with an F1 score of 1 at a confidence threshold of 0.000. Using Stage 2 alone (without Stage 1), the system achieved a precision of 0.9889 and a recall of 0.9046. When incorporating Stage 1 learning into Stage 2, the model further improved its performance, achieving a precision of 1.0000 and a recall of 0.9542. Furthermore, the system demonstrated a processing speed of 0.7 ms for preprocessing, 3.4 ms for inference, and 0.7 ms for postprocessing per image, making it suitable for real-time detection applications.

In addition to marker detection, the ID detection process was designed to recognize six unique marker patterns (IDs) from sonar images. Using the predicted bounding boxes from Stage 2, the markers were cropped into patch images and labeled with their corresponding IDs to create the training dataset. The classification model trained on these labeled patches achieved a mAP@0.5 of 0.995 across all six classes in the training dataset, ensuring robust and accurate ID recognition.

IV. Conclusion

We proposed an acoustic camera marker family featuring numerous key points to enhance recognition robustness. A comprehensive acoustics simulator database was constructed, consisting of sonar images and corresponding labels. The Curriculum learning strategy employed in this study improved the model's robustness, effectively bridging the gap between synthetic and real-world datas. For future work, we aim to extend this system by implementing the acoustic n-point method for real-time 6DoF pose estimation.

References

[1] Yusheng Wang, Yonghoon Ji, Dingyu Liu, Yusuke Tamura, Hiroshi Tsuchiya, Atsushi Yamashita, and Hajime Asama. Acmarker: Acoustic camera-based fiducial marker system in underwater environment. *IEEE Robotics and Automation Letters*, 5(4):pp. 5018–5025, 2020.

[2] Kalin Norman, Daniel Butterfield, and Joshua G. Mangelson. Actag: Opti-acoustic fiducial markers for underwater localization and mapping. *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2023)*. pp. 9955-9962, 2023.

[3] Z. Zhang, Y. Hu, G. Yu, and J. Dai. Deeptag: A general framework for fiducial marker design and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(03):2931–2944, 2023.

[4] Ultralytics YOLO 11.0.0, 2024.