IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

ACMarker: Acoustic Camera-based Fiducial Marker System in Underwater Environment

Yusheng Wang[1], Yonghoon Ji[2], Dingyu Liu[1], Yusuke Tamura[3], Hiroshi Tsuchiya[4], Atsushi Yamashita[1], and Hajime Asama[1]

***Abstract*— ACMarker is an acoustic camera-based fiducial marker system designed for underwater environments. Optical camera-based fiducial marker systems have been widely used in computer vision and robotics applications such as augmented reality (AR), camera calibration, and robot navigation. However, in underwater environments, the performance of optical cameras is limited owing to water turbidity and illumination conditions. Acoustic cameras, which are forward-looking sonars, have been gradually applied in underwater situations. They can acquire high-resolution images even in turbid water with poor illumination. We propose methods to recognize a simply designed marker and to estimate the relative pose between the acoustic camera and the marker. The proposed system can be applied to various underwater tasks such as object tracking and localization of unmanned underwater vehicles. Simulation and real experiments were conducted to test the recognition of such markers and pose estimation based on the markers.**

I. Introduction

Sensing in the underwater domain is a difficult task that has attracted the interest of researchers. Although techniques for optical camera-based sensing have been highly developed and applied to underwater environments, sonars are the only viable modality for extreme underwater environments with poor visibility. An acoustic camera, which is a forward-looking sonar, can acquire high-resolution images without concern for turbidity or illumination. This type of camera has been broadly employed in sensing and navigation applications using underwater robots, such as underwater mapping, mosaicking, and localization [1]–[3]. To construct a controllable environment for autonomous robots in a land environment, artificial landmarks such as ARTag, AprilTag, and ArUco are widely used in robotics and computer vision applications [4]–[6]. Fiducial marker systems can detect square markers, offer IDs, and estimate the six-degree-of-freedom (6DoF) relative pose between the camera and marker. Such systems have made great contributions to ground truthing, navigation in structured environments, and augmented reality (AR). It is necessary to build an acoustic camera-based fiducial marker system since the optical camera-based fiducial marker system is restricted by the limited visibility in underwater environments. Acoustic images are generated based on the backscattered intensity. Thus, it is possible to utilize the characteristic of an active sonar in which the backscattered intensity is influenced by the material of the target. By combining diffuse and specular reflection materials, it is possible to create a marker with patterns that can be recognized in an acoustic image. Currently, a circle marker can be robustly recognized in an acoustic image [7]. The information from the circle marker is the center point and identified ID; however, 6DoF pose estimation cannot be implemented. For 6DoF pose estimation, a square marker is considered since the four corners can be used to calculate the pose information. To the best of our knowledge, there are currently no square fiducial marker systems for acoustic cameras. Besides, the theory of the acoustic camera-based pose estimation problem is still in its early stages. In our previous research, we proved that 6DoF poses can be estimated with three or more points. However, it is necessary to improve the algorithm since pose estimation from a small number of points (e.g., three or four) is error-prone [8].

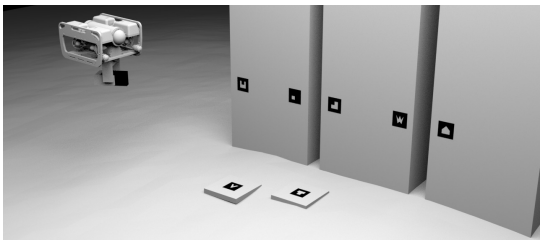


Fig. 1. Conceptual representation of proposed marker system. Metal markers can be placed on concrete structures or on the ground with a concrete or plaster base. This can be applied to underwater autonomous vehicle navigation and object tracking.

In this research, an acoustic camera-based fiducial marker system named ACMarker (marker for acoustic camera) is proposed, as shown in Fig. 1. The markers can be placed directly on the walls of an underwater structure or on the seabed with a concrete or plaster base. This facilitates the navigation of autonomous underwater vehicles (AUVs), as well as underwater structure inspection. The contributions of the system can be summarized as follows:

- We propose detection and ID identification methods based on simply designed square markers.

This work was supported by JSPS KAKENHI Grant Number 19K21547, 20K19898.

[1]Y. Wang, D. Liu, A. Yamashita, and H. Asama are with the Department of Precision Engineering, Graduate School of Engineering, University of Tokyo, Japan. {wang,liu,yamashita,asama}@robot.t.u-tokyo.ac.jp

[2]Y. Ji is with the School of Materials Science, JAIST, Japan. ji-y@jaist.ac.jp

[3]Y. Tamura is with the Department of Robotics, Graduate School of Engineering, Tohoku University, Japan. y.tamura@srd.mech.tohoku.ac.jp

[4]H. Tsuchiya is with the Research Institute, Wakachiku Construction Co., Ltd., Japan. hiroshi.tsuchiya@wakachiku.co.jp

- We propose a method to accurately and precisely estimate the 6DoF relative pose between the acoustic camera and the marker.
- Detection and pose estimation can be processed based on a single image and should work in real time.

The rest of the paper is organized as follows. In Section II, related works are introduced and compared to the proposed method. Section III introduces the acoustic camera models. Section IV explains the design of the marker and provides an overview of the proposed system. Section V explains the marker detection and ID identification method. Section VI describes the pose estimation method. Experiments and evaluations are presented in Section VII, followed by discussions in Section VIII. Finally, conclusions and future works are presented in Section IX.

II. Related Research

Lee et al. first tested acoustic camera-based artificial landmarks by combining diffuse and specular material, which inspired this research [7]. They designed circle 2D artificial landmarks with a probability-based recognition system. Although the landmarks can be recognized stably in an acoustic image, they require an image sequence for recognition. In other words, the camera must maintain a stationary position for a specific period for recognition. Furthermore, while circle markers are more robust than square markers, they cannot be used for 6DoF pose estimation because only the center point can be acquired. Westman et al. proposed a simultaneous localization and mapping (SLAM) method based on unconstrained landmarks [9]. This also requires an image sequence with a pose-graph SLAM framework to estimate the camera trajectory. Pyo et al. used pillars as 3D landmarks. They focused on the recognition of such landmarks by combining shadow information [10]. The 3D landmarks were specialized for the seabed, and were long-lasting and unaffected by biofouling. The marker designed in this research is a square 2D marker. It is made of metal and has a concrete or plaster base. It can be placed on a wall of an underwater structure or on the ground. More importantly, the main aim is to design a marker that can estimate the relative pose between the marker and camera, which is necessary for tasks such as visual tracking, underwater positioning, and SLAM.

Plane-based resection for the acoustic camera has not been fully studied thus far. Few research studies have focused on pose estimation based on a planar target. Negahdaripour proposed a method to realize plane-based resection and lens distortion removal simultaneously, i.e., calibration [11]. However, numerous points are required, and the pose estimation result may be influenced by the initial guess. Brahim et al. used a covariance matrix adaptation evolution strategy algorithm (CMA-ES) to verify the acoustic camera model by conducting plane-based resection [12]. They used 18 points over the acoustic image. However, the algorithm was also influenced by the initial value. In our previous study, we proposed a method to obtain the initial value using a weak-perspective camera model and conduct nonlinear optimization for refinement; this was named the planar acoustic n point (AnP) [8]. However, if the target is small with only a few points (e.g., 0.25 m $\times$ 0.25 m with four points), only 5DoF can be estimated accurately. Although the reprojection error is small, the estimated pose may have a decimeter-level error. More constraints are necessary during optimization to further improve this result. Additional information such as an elevation-angle scope limitation and illuminated area [13] can be included. In this paper, we propose a particle-filter-based optimization method with additional constraints to improve the result.

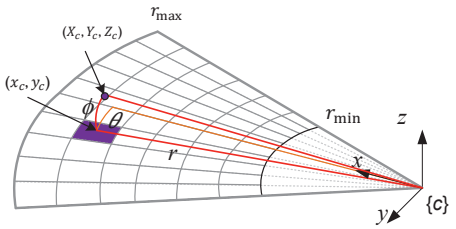


Fig. 2. Acoustic camera projection model.

III. Acoustic Camera Model

A 3D point in the camera coordinate system can be represented as (r, θ, ϕ) and can be transferred into Euclidean coordinates (X_c, Y_c, Z_c) based on the following equation:

$$\begin{bmatrix} X_c & Y_c & Z_c \end{bmatrix}^\top = \begin{bmatrix} r\cos\phi\cos\theta & r\cos\phi\sin\theta & r\sin\phi \end{bmatrix}^\top. \tag{1}$$

The 3D point is projected in an area in the acoustic image denoted as (x_c, y_c) in the Euclidean image coordinates. As shown in Fig. 2, the projected point can be represented by

$$\begin{bmatrix} x_c & y_c \end{bmatrix}^\top = \begin{bmatrix} r\cos\theta & r\sin\theta \end{bmatrix}^\top. \tag{2}$$

The acoustic camera projection model can be written as follows by combining Eqs. (1) and (2):

$$\begin{bmatrix} x_c & y_c & 0 \end{bmatrix}^\top = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_c & Y_c & Z_c \end{bmatrix}^\top, \tag{3}$$

where $\alpha = \frac{1}{\cos\phi}$.

For acoustic cameras such as DIDSON and ARIS, the scope of the elevation angle ϕ is between $\phi_{\min} = -7^\circ$ and $\phi_{\max} = 7^\circ$. In other words, the value of α is between 1 and 1.0075. Thus, α can be approximated to 1 so that the projection can be considered an orthographic projection [2][14]. If α is considered a constant larger than 1, then the acoustic camera model can be seen as a weak-perspective camera model.

IV. Overview

A. System Overview

An overview of the proposed system is shown in Fig. 3. The input is the sonar signal matrix, which is a bearing-range image; this is denoted as a raw image in this paper. The output is the detected marker with the ID and the estimated

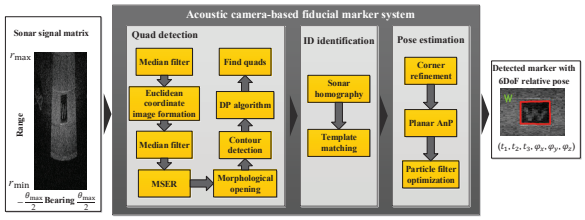
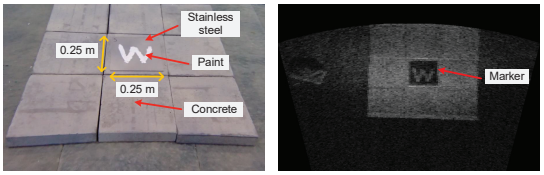


Fig. 3. System overview. Input is sonar signal matrix, and output is detected marker with relative 6DoF pose. The entire system can be separated into quad detection, ID identification, and pose estimation.



(a) (b)

Fig. 4. Fiducial markers: (a) marker in water tank and (b) marker in acoustic image. Stainless steel with concrete offers sufficient contrast for detection.

6DoF pose between the camera and marker coordinates. The system can be divided into three parts. Similar to optical camera-based marker systems [4]–[6], we initially detect quadrilaterals (i.e., quads) in the image. Then, we remove the projection distortion in the pattern and conduct template matching to verify the ID information. Finally, we estimate the pose based on the corners of the quads and the additional constraints.

B. Fiducial Markers

Fiducial markers for acoustic cameras can be created by combining diffuse and specular reflection materials. Concrete, which is one of the most common materials for man-made underwater structures, possesses outstanding backscattering ability. After several tests, we found that stainless steel is one of the most ideal specular materials. Such materials can generate stable dark regions in acoustic images. Combinations of such materials can provide sufficient contrast for marker detection in the acoustic image. In this research, we designed the fiducial markers shown in Fig. 4. Characters were painted on the stainless steel as IDs; this can be changed by using other diffuse materials or by simply cutting patterns in the stainless steel. The stainless steel can be directly placed on man-made concrete structures such as bridges and caissons. If a marker is placed in other environments such as the seafloor, it is better to have a concrete or plaster base.

V. Marker Detection and ID Identification

A. Quad Detection

The raw information read from the sensor is an $r - \theta$ matrix. First, a median filter is used to filter the noise in the raw image matrix. Then, the raw image matrix is transferred into a Euclidean coordinate image based on Eq. (1). If a straight line in 3D space is projected onto the acoustic image, it can be approximated as a straight line in the 2D image

(a) (b) (c) (d) (e)

Fig. 5. Quad detection: (a) original marker, (b) after MSER binarization, (c) after morphological opening, (d) after contour detection, and (e) after DP algorithm.

(a) (b) (c) (d) (e)

Fig. 6. Sonar homography: (a) detected quad, (b) affine transform, (c) binarization based on Otsu method, (d) discarding pixels on edge, and (e) template generated from another frame.

[15]. This can be also proven by approximating the acoustic camera model to an orthographic projection model or weak-perspective projection model, as explained in Section III. Then, maximally stable extremal regions (MSER) is used for binary segmentation [16]. MSER can stably detect regions of high contrast in the acoustic image. Since the size of the marker and the scale of the acoustic image (i.e., range distance per sample) are known, it is possible to adjust the parameters of MSER detection based on the size of the marker. This effectively results in a small number of region candidates. Thresholding is not an ideal method of detecting feasible regions because the intensities of the pixels are not evenly distributed owing to heavy noise and multiresolution.

Next, a morphological opening process is conducted to smooth the edges of the regions and remove small objects before contour detection. Similar to most optical camera-based fiducial marker systems, the Douglas-Peucker (DP) algorithm is used to refine the polygon after the contours are detected [17]. Polygons with four vertices are detected as quads. Since the projected marker can only be a parallelogram, constraints can be added to the angles and the side lengths of the quads to further limit the number of marker candidates. Figure 5 shows an example of the aforementioned quad-detection process.

B. ID Identification

This subsection introduces the method of recognizing the ID information from the markers after marker candidates are detected. First, the projection distortion is removed based on sonar homography. Then, the ID can be extracted by template matching.

1) Distortion Removal with Sonar Homography: Sonar homography has been studied by several research groups; it is the main component of sonar mosaicking. Hurtos et al. applied isometry with orthographic projection approximation and a zero-roll-angle assumption to sonar homography [2]. However, such a pose assumption cannot be applied to this research because the roll angle is not always zero. Negahdaripour et al. proposed a non-approximated sonar homography method with no assumptions [18]. The homog-

raphy matrix $\mathbf{H}$ can be written as follows:

$$\mathbf{p}^{'} = \begin{bmatrix} \gamma q_{11} & \gamma q_{12} & \psi q_{13} \\ \gamma q_{21} & \gamma q_{22} & \psi q_{23} \\ 0 & 0 & 1 \end{bmatrix} \mathbf{p} = \mathbf{H}\mathbf{p}, \tag{4}$$

where $\gamma = \frac{\cos\phi}{\cos\phi^{'}}$ and $\psi = r\frac{\sin\phi}{\cos\phi^{'}}$. q_{ij} denote the components in matrix $\mathbf{Q} = \mathbf{R} - \mathbf{t}\mathbf{n}^{\top}$. $\mathbf{R}$ and $\mathbf{t}$ are the rigid body transformations, and $\mathbf{n}$ is the surface normal vector for the plane. The homography matrix $\mathbf{H}$ varies from point to point in the acoustic image, and elevation angle ϕ is unknown from the acoustic images. It is necessary to carry out nonlinear optimizations to solve $\mathbf{H}$ for each point. In this research, sonar homography is used to recognize the IDs of the markers for which precision of the transformation is not strictly required. Some assumptions can be made to simplify the model. Assuming α in the acoustic camera model is a constant over one acoustic image, γ is also a constant. Similarly, we can regard ψ as a constant over the marker in the image because the fiducial markers shown in the acoustic image can be considered as a small area. For instance, considering ARIS EXPLORER 3000 with a 0.25 m $\times$ 0.25 m marker, the maximum change in ψ is $0.25\text{m} \times \frac{\sin 7^{\circ}}{\cos 7^{\circ}} = 0.03075\text{m}$, which is the most extreme situation. Then, the homography matrix $\mathbf{H}$ is approximated as a standard affine transformation matrix. We transfer the quad into a square for the following ID identification.

2) Template Matching: The marker candidates are then transferred into square images. Here, we apply the Otsu method to binarize the images [19]. The noise on the edge of the images is filtered by setting the intensities to zero. Then, template matching is conducted to recognize the patterns. Since the template and pattern are binary images of the same size, the Hamming distance D_H is used to compute the similarity score S between pattern $\mathbb{I}_p$ and template $\mathbb{I}_t$.

$$S = 1 - \frac{D_H(\mathbb{I}_t, \mathbb{I}_p)}{W \times H}, \tag{5}$$

where W and H are the width and height of the marker in the image, respectively. The pattern with a score larger than the threshold is considered the ID. In this research, the threshold is set to 0.80 with a marker size of 80 pixels $\times$ 80 pixels. Figure 6 shows an example of sonar homography. The template is generated from an arbitrary frame, as shown in Fig. 6(e).

VI. Pose Estimation

A. Corner Refinement

Once the quads are detected, the vertices are used as corner features for pose estimation. However, severe noise, image distortion, and nonlinear projection may influence the corner detection accuracy. Although we assume that in Euclidean coordinate images, lines in 3D space are projected as lines in 2D images, for high-precision tasks, the approximation may influence the result. The multiresolution characteristic of the Euclidean coordinate image may also deteriorate the result of corner detection. It is possible that the vertices are away from the real corners. We found that refining the subpixels in the raw image may improve the result. This process involves iteratively searching the corners in a defined window size, beginning with the vertices of the quad. This process is essential since the corner position from the quad is not enough reliable.

B. Planar AnP

Defining the marker coordinate as the world coordinate, a 3D point in world coordinates is denoted by $\mathbf{p}_w$, and in camera coordinates by $\mathbf{p}_c$. Pose $\mathbf{R}$ and $\mathbf{t}$ in this study are defined as follows:

$$\mathbf{p}_c = \mathbf{R}\mathbf{p}_w + \mathbf{t}. \tag{6}$$

The pose estimation problem can be formed as an estimation of $\mathbf{R}$ and $\mathbf{t}$, based on the following equation:

$$\hat{\mathbf{R}}, \hat{\mathbf{t}} = \underset{\mathbf{R},\mathbf{t}}{\operatorname{argmin}} ||f(\mathbf{R}\mathbf{X} + \mathbf{t}) - \mathbf{Y}||_F^2, \tag{7}$$

where $f(.)$ is the acoustic camera model, and F denotes the Frobenius norm. For $\mathbf{X} \in \mathbb{R}^{3\times n}$ and $\mathbf{Y} \in \mathbb{R}^{2\times n}$, each column in $\mathbf{X}$ holds the coordinates of a known 3D point in world coordinates, and each column in $\mathbf{Y}$ holds the coordinates of an observed point in the image. To acquire the initial value for the nonlinear optimization in Eq. (6), we define $\mathbf{t} = [t_1, t_2, t_3]^{\top}$, $\mathbf{t}_{12} = [t_1, t_2]^{\top}$, and $\mathbf{M}$, which is from the first two rows of the rotation matrix $\mathbf{R}$. If α is considered a constant [8], the problem can be formed as

$$\hat{\mathbf{M}}, \hat{\mathbf{t}_{12}} = \underset{\mathbf{M},\mathbf{t}_{12}}{\operatorname{argmin}} ||\alpha(\mathbf{M}\mathbf{X} + \mathbf{t}_{12}) - \mathbf{Y}||_F^2. \tag{8}$$

Assuming $\alpha = 1$ converts the problem into an orthographic n point (OnP) problem for the coplanar-point case, this may lead to a sub-Stiefel Procrustes problem [20], which is difficult to solve analytically. In contrast, the assumption of $\alpha > 1$ may lead this problem into a plane-based resection on the weak-perspective camera model, from which a clean closed-form solution can be acquired. We apply the method proposed by Bartoli et al. to solve the plane-based resection on the weak-perspective camera model; this is shown as Algorithm 1 [21]. Here, $\mathbf{1}$ refers to a vector of all ones, and $\odot$ is Hadamard product. λ_1 refers to the eigenvalue and rank_1 refers to rank-1 decomposition.

The 5DoF pose can be estimated by linearizing the acoustic camera model. Two solutions can be acquired from the initial 5DoF estimation. Assuming that there is a priori knowledge of the signs of the rotation angles, in this research we leave one solution for further processing. In addition, the markers symmetric to the imaging plane can generate the same acoustic image. It is acceptable to consider one solution since they are symmetric.

The last DoF t_3 cannot be acquired from the weak-perspective camera model. The following optimization is conducted to obtain the initial value of t_3. Since θ is independent of t_3, it is possible to change the cost function in Eq. (6) into an equivalent form as follows:

$$\hat{t}_3 = \underset{t_3}{\operatorname{argmin}} \sum_{i=1}^{n} (\sqrt{{}^iX_c^2 + {}^iY_c^2 + {}^iZ_c^2} - \sqrt{{}^ix_c^2 + {}^iy_c^2})^2. \tag{9}$$

Algorithm 1: Initial 5DoF Estimation

Input: $\mathbf{X} \in \mathbb{R}^{3\times n}$, $\mathbf{Y} \in \mathbb{R}^{2\times n}$

Output: α, $\mathbf{R}_1$, $\mathbf{t}_1$, $\mathbf{R}_2$, $\mathbf{t}_2$

$\mathbf{x} \in \mathbb{R}^3$, $\mathbf{y} \in \mathbb{R}^2$,$\mathbf{x}\leftarrow \frac{1}{n}\mathbf{X1}$, $\mathbf{y}\leftarrow \frac{1}{n}\mathbf{Y1}$

$\mathbf{X}^{'} \leftarrow \mathbf{X} - \mathbf{x1}^\top$,$\mathbf{Y}^{'} \leftarrow \mathbf{Y} - \mathbf{y1}^\top$

$(\mathbf{U},\Sigma,\mathbf{V}) \leftarrow \mathrm{SVD}(\mathbf{X}^{'})$

$\mathbf{Z}\leftarrow \det(\mathbf{U})\mathbf{Y}^{'}[\mathbf{v}_1\ \mathbf{v}_2]$

$\mathbf{S}\leftarrow \mathrm{diag}(\Sigma_{11},\Sigma_{22})$

$\mathbf{B} \leftarrow \mathbf{ZS}^{-1}$

$\mathbf{u} \in \mathbb{R}^2$, $\alpha^2 \leftarrow \lambda_1(\mathbf{BB}^\top)$, $\mathbf{u} \leftarrow \mathrm{rank}_1(\mathbf{I} - \frac{1}{\alpha^2}\mathbf{BB}^\top)$

$\mathbf{Q}\leftarrow [\frac{1}{\alpha}\mathbf{B}\ \mathbf{u}]$

Denote the first and the second row in $\mathbf{Q}$ by $\mathbf{q}_1$ and $\mathbf{q}_2$, $\mathbf{Q}_1 \leftarrow [\mathbf{q}_1^\top\ \mathbf{q}_2^\top\ (\mathbf{q}_1 \times \mathbf{q}_2)^\top]^\top$

$\mathbf{Q}_2 \leftarrow \mathbf{Q}_1 \odot \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$

$\mathbf{R}_1 \leftarrow \mathbf{Q}_1\mathbf{U}^\top$, $\mathbf{R}_2 \leftarrow \mathbf{Q}_2\mathbf{U}^\top$

$\mathbf{t}_1 \leftarrow \frac{1}{\alpha}\mathbf{y} - \mathbf{Q}_1\mathbf{U}^\top\mathbf{x}$, $\mathbf{t}_2 \leftarrow \frac{1}{\alpha}\mathbf{y} - \mathbf{Q}_2\mathbf{U}^\top\mathbf{x}$

The Levenberg–Marquardt (LM) algorithm is used to estimate t_3. Then, the estimated values can be used as the initial guess or a final optimization. In our previous research, we used the LM algorithm with no constraints. If the corner points are accurately measured, then LM may lead to an accurate and precise result. However, it is difficult to detect the corner position accurately and precisely owing to multiresolution and heavy noise. The pose estimation problem is ill-posed and there may be a large error in t_3. Additional information can be added as constraints to refine the result, including $\phi_{\max}$ and $\phi_{\min}$. This paper proposes a particle-filter-based method for local optimization.

C. Particle-Filter-Based Optimization

Directly optimizing the corner reprojection error based on the LM algorithm cannot guarantee that the marker is within in the field of view, which may lead to a large error in the pose. Fixed values of $\phi_{\max}$ and $\phi_{\min}$ are an important characteristic of the acoustic camera. These can be used to check whether the markers are in the field of view of the acoustic camera. Furthermore, if the marker is placed on a flat surface, the ϕ angle constraint can generate an illuminated area (IA) [13] in the acoustic image, as shown in Fig. 7. IA is the bright area in the acoustic image when sensing a flat surface. This area is generated from the limitation of the ϕ angle scope. $\phi_{\max}$ and $\phi_{\min}$ may lead to two boundaries: upper boundary b_u and lower boundary b_l, as shown in Fig. 7(a). However, owing to the complex phenomenon of ultrasound and the low signal-to-noise ratio, precise detection of this area is difficult because the signal on the side of the image is weak. In this study, we use the middle point of the boundaries for pose estimation. The middle point of the upper and lower boundaries can be described by ranges $\mathfrak{R}_u$ and $\mathfrak{R}_l$, respectively, since $\theta = 0$.

$$\mathfrak{R}_u = \frac{t_z}{-\cos\phi_{\max}\sin\varphi_y + \sin\phi_{\max}\cos\varphi_y\cos\varphi_x}, \quad (10)$$

where t_z is from $\mathbf{t}_c = -\mathbf{R}^\top\mathbf{t} = [t_x, t_y, t_z]^\top$, which is the camera pose in the marker coordinate. Similarly, $\mathfrak{R}_l$ can be computed by changing $\phi_{\max}$ to $\phi_{\min}$. The detection of middle points is based on the method proposed in [13]. After the images are binarized, small objects on the images are removed, and then a 1D search is conducted along $\theta = 0$ to find the middle points.

The initial pose $\hat{\mathbf{w}} = [\hat{t}_1, \hat{t}_2, \hat{t}_3, \hat{\varphi}_x, \hat{\varphi}_y, \hat{\varphi}_z]^\top$ can be estimated from the planar AnP. For particle-filter-based optimization, particles are first generated randomly around the initial value. Denoting the pose of the i-th particle by $\mathbf{w}_i$, the initial position for $\mathbf{w}_i$ can be written as

$$\mathbf{w}_i = \hat{\mathbf{w}} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon}_k \sim \mathcal{U}(-0.1, 0.1), \quad (11)$$

where $\boldsymbol{\epsilon}_k$ is the k-th element in vector $\boldsymbol{\epsilon}$, and $\mathcal{U}$ refers to a uniform distribution. For the sampling step, a new particle set is generated from the past particle set according to

$$\begin{aligned} &\mathbf{w}_i^{(n)} = \mathbf{w}_i^{(n-1)} + \boldsymbol{\mu}, \\ &\boldsymbol{\mu} \sim \mathcal{N}(0, \mathrm{diag}(\sigma_{t_1}^2, \sigma_{t_2}^2, \sigma_{t_3}^2, \sigma_{\varphi_x}^2, \sigma_{\varphi_y}^2, \sigma_{\varphi_z}^2)), \end{aligned} \quad (12)$$

where $\mathcal{N}$ refers to a normal distribution. Note that $\sigma_{t_3}^2$ should be larger than the other variances because the estimated t_3 has a large uncertainty. In this study, we set $\sigma_{t_3} = 0.2$, and the other variances to 0.1. For importance weighting, the cost function for pose estimation is written as follows:

$$d^2 = ||f(\mathbf{RX}+\mathbf{t}) - \mathbf{Y}||_F^2 + \lambda(\mathfrak{R}_l - \hat{\mathfrak{R}}_l)^2 + \lambda(\mathfrak{R}_u - \hat{\mathfrak{R}}_u)^2, \quad (13)$$

where $\hat{\mathfrak{R}}_l$ and $\hat{\mathfrak{R}}_u$ are the measured values, and $\mathfrak{R}_l$ and $\mathfrak{R}_u$ are the estimated values. λ is a weight to balance the reprojection and IA errors. Denoting the error of the initial value by d_{initial}^2, the probabilistic weights ω_i are computed as follows:

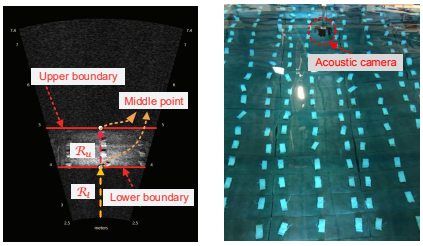
$$\omega_i = \exp\left(-\frac{d^2}{d_{\text{initial}}^2}\right). \quad (14)$$

If ϕ_m is beyond $[\phi_{\min}, \phi_{\max}]$, then we directly set the probabilistic weight ω to zero. Here, (r_m, θ_m, ϕ_m) denotes the corner in the camera coordinates. A sampling importance resampling (SIR) particle filter is used in this study. During resampling, we calculate the number of particles by multiplying the standard deviations and the hyperparameter density. If the number of particles is less than the threshold, the optimization may be considered to be converged. Empirically, after 1∼2 iterations, the process converges. The weighted average pose of the particles is used as the optimized result.

VII. Experiment

A. Simulation Experiment

1) Simulation Environment: The synthetic images were generated with the open-source Blender software. The acoustic camera can be simulated based on a ray-tracing model [22][23]. We set the attenuation of the ray strength based on the inverse square law. To model the reflection, two kinds



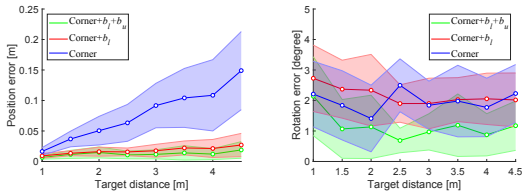
(a) (b)

Fig. 7. Illuminated area (IA): (a) upper and lower boundaries in acoustic. Owing to fixed aperture angle, illuminated area can be seen in image. (b) Environment where (a) was captured.

of materials were set: diffuse reflection-based material and specular reflection-based material. We prepared markers with five types of patterns by combining the two materials. The Hamming distances between the marker patterns are larger than a certain value. The size of the markers is 0.3 m $\times$ 0.3 m. The backscattered rays form a grayscale image. With depth information, it is possible to generate a synthetic acoustic image [22]. The resolution of the image is 0.003 m. The program is listed in our GitHub[1]. It is worth mentioning that the detected corner positions are not sufficiently accurate owing to the multiresolution and nonlinear projection characteristics. The performance of the marker may be influenced by the distance to the marker and the orientation of the marker. Inspired by AprilTag [5], we also conducted two types of experiment. First, we fixed the orientation of the marker and investigated the relationship between the pose estimation error and the distance. Second, we fixed the distance of the marker and checked the relationship between pose estimation error and rotation angle. It should be noted that the orientation factor is different from that for optical markers. Two aspects influence the result: the angle between the normal vector of the marker and the sonar heading, and the angle between the normal vector of the marker and the normal vector of the imaging plane. To simplify, we place the marker at a fixed position, and rotate the marker along the x-axis of the acoustic camera (roll) or the y-axis of the acoustic camera (pitch). For pose estimation, the minimum and maximum number of particles were set to 3000 and 5000, respectively. For each image, five particle filter iterations were taken for local optimization. The similarity score was maintained at approximately 96%. The position error for b_l and b_u for the synthetic image is approximately 0.003 m.

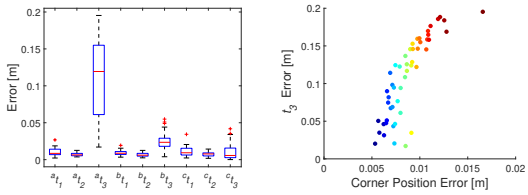
2) Distance Factor: For the experiment on distance factor, we tested the distance from 1 to 4.5 m. It is assumed that all markers are completely in the field of view. Pitch angle equals to zero is not an ideal configuration since the backscattered intensity is too weak, and the upper boundary locates at the infinite. The basic configuration was set as roll = 0, pitch = 30°. Then, for each distance, we randomly generated 50 images with markers at different positions. For distances smaller than 1 m, there is no complete marker

[1]https://github.com/sollynoay/ACMarker



(a) (b)

Fig. 8. Relationship between pose estimation error and marker distance: (a) position error, (b) rotation error.



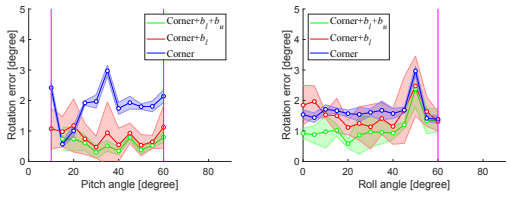
(a) (b)

Fig. 9. Error in the case distance = 2.5 m: (a) error in t_1, t_2, t_3, where a, b, and c refer to corner only, corner+b_l, and corner+b_l+b_u, respectively; (b) relationship between t_3 and corner position error. Color refers to the corner position error.

under current configuration.

For quad detection and ID identification, the detection accuracy is 100% and all the IDs are true positive. The root mean square error (RMSE) was used to evaluate the pose estimation result. Figure 9 shows the result of the distance factor experiment. It is known that with corner information only, the position error increases when the distance of the marker increases. This can be effectively improved by combing IA information. With both b_u and b_l, it is possible to acquire ground-truthing level results for the position and orientation. If we only use b_l information, the position error can be significantly improved, whereas the orientation error maintains the same level when using the corner only. To further analyze the position error, taking distance = 2.5 m as an example, Figure 9(a) shows the error for t_1, t_2 and t_3 estimated by (a) corner only, (b) corner+b_l, and (c) corner+b_l+b_u. It is known that for pose estimation by corner only, although t_1 and t_2 are accurate, there is a large error in t_3. Figure 9(a) shows the relationship between the corner position error and t_3 error. For Spearman's correlation, r = 0.8608 and p = 0, it can be said that the t_3 estimation result is related to the corner position error.

3) Orientation Factor: During the orientation factor test, the marker was set at a fixed position. Since we used a particle filter, the result may differ slightly for each run. We repeated the program five times to acquire the average performance. Figure 10(a) indicates the factor of pitch angle. The region between the two pink lines is the detectable region. For pitch angle smaller than 10° or larger than 60°, the marker cannot be detected well by the current system design. This is because of the quad detection failure. In contrast, for the factor of roll angle, we maintained the pitch angle at 30° and changed the roll angle. From Fig. 10(b), it is known that the detectable region is from 0 to 60°.



(a) (b)

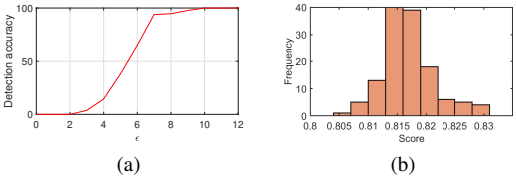
Fig. 10. Simulation experiment on orientation factor: (a) rotation estimation result for changing pitch angle, (b) rotation estimation for changing roll angle. Pink lines refer to the boundaries of the detectable scope.

B. Real Experiment

In a real experiment, we tested the performance of recognition of the marker and relative pose estimation in the water tank shown in Fig. 7. The acoustic camera ARIS EXPLORER 3000 was mounted on the AR2 rotator. The position information of the camera was measured with a ruler, and the orientation information was from the rotator. The ARIS EXPLORER 3000 operated at 3.0 MHz so that the resolution of the image was 0.003 m. $r_{\min}$ and $r_{\max}$ were set to 0.98 and 4.36 m, respectively. The size of the raw image was 128 pixels $\times$ 1483 pixels, and the size of the Euclidean coordinate image was 802 pixels $\times$ 1512 pixels. A stainless steel board of size 0.25 m $\times$ 0.25 m was used in this research. Two types of markers were tested: one with "w" painted on the metal board, and the other with a wooden square stuck onto the metal board. In total, three acoustic videos were recorded and denoted by Cases 1, 2, and 3. The frame rate of the acoustic videos was 3.0 Hz. We used the same parameter settings of the particle filter for the simulation experiment.

- Case 1: W marker, roll = 2.3°, pitch = 30°, 42 frames at the same position.
- Case 2: W marker, roll rotates from 8.6° to −9.8°, pitch = 30°, 131 frames.
- Case 3: Square marker, roll = 0, pitch = 33°, 51 frames at the same position.

Initially, for quad detection, the detection rate reached 100% in Cases 1 and 2 under the current parameter settings. However, some quads that could not be detected in Case 3, which is because of secondary reflection due to the thickness of the wooden board. Still, the detection rate reached 90%. The result of quad detection is basically influenced by the MSER detection and DP algorithm. In this research, we discuss the parameters of the DP algorithm further. The boundaries for markers are highly distorted for sonar images, as shown in Fig. 5(d), ϵ is a parameter that refers to the largest distance between the input contour and the approximated contour in the DP algorithm. It can be seen as a parameter on how much the contour should be modified. Figure 11(a) shows the relationship between ϵ and the quad detection accuracy. It is known that when ϵ is larger than 10, all quads can be detected successfully. Since the resolution of the image is 0.003 m, the scale of the contour modified is approximately 0.03 m. Since we only tested two IDs, all recognized IDs are true positive. An important parameter that influences the recognition rate is the threshold of the similarity score. In Fig. 11(b), we show the histogram of the similarity score of the 131 frames in Case 2. For the experiment, we set the threshold to 0.8. Then, for pose estimation, since $\mathbf{t}$ was not measured directly for the ground truth, we also evaluated $\mathbf{t}_c$. The pose estimation results are shown in Table I and II, where RE refers to the reprojection error of the corners. Since b_u is larger than the effective range, only b_l is used as constraint to refine the result. We reach the same conclusion with the simulation experiment. Further, we tested the corner position error for Case 1, and found out that the corner error is approximately 0.02 m, which is larger than that of the simulation experiment. In Case 1, we also compared the initial guess and the optimized result using corners. The average pitch error significantly decreased from 6.03° to 1.54°, whereas using the LM algorithm may lead to an error of up to 6.27°. The roll angle estimation result during roll rotation in Case 2 is shown in Fig. 12. Although there is some noise, the proposed method can successfully track the motion of the marker.



(a) (b)

Fig. 11. Detection test based on Case 2: (a) test on effect of ϵ in DP algorithm, and (b) histogram of the similarity scores.

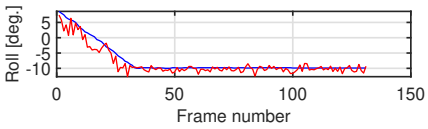


Fig. 12. Tracking test: roll-angle estimation in Case 2. The red line is the estimated result and the blue line is the ground truth.

TABLE I

RESULTS 1 IN REAL EXPERIMENT

Case		Detection rate	$\mathbf{t}_c$ [m]	RPY [degree]	RE [m]
1	Corners	42/42	0.102 ± 0.061	2.46 ± 1.45	0.0086 ± 0.0017
	Corners + b_l	42/42	0.077 ± 0.099	2.29 ± 0.80	0.0139 ± 0.0030
2	Corners	131/131	0.049 ± 0.034	1.78 ± 0.63	0.0094 ± 0.0024
	Corners + b_l	131/131	0.044 ± 0.031	1.78 ± 0.69	0.0131 ± 0.0032
3	Corners	46/51	0.164 ± 0.140	2.81 ± 1.94	0.0109 ± 0.0023
	Corners + b_l	46/51	0.077 ± 0.068	1.78 ± 1.15	0.0112 ± 0.0025

TABLE II

RESULTS 2 IN REAL EXPERIMENT

Case		$\mathbf{t}$ [m]	$\mathbf{t}_{12}$ [m]	$\mathbf{t}_{12}$ [m]
1	Corners	0.145±0.006	0.012 ± 0.005	0.251 ± 0.010
	Corners + b_l	0.026 ± 0.007	0.010 ± 0.007	0.039 ± 0.015
2	Corners	0.092 ± 0.024	0.020 ± 0.004	0.157 ± 0.042
	Corners + b_l	0.029 ± 0.006	0.020 ± 0.004	0.041 ± 0.012
3	Corners	0.022 ± 0.011	0.015 ± 0.010	0.027 ± 0.022
	Corners + b_l	0.014 ± 0.004	0.014 ± 0.007	0.009 ± 0.007

VIII. DISCUSSIONS

A. Marker size and design

In the simulation experiment, we also tested markers with sizes of 0.05 m, 0.1 m, 0.15 m, and 0.2 m. It is known that the minimum size of the marker is 0.2 m under the

(a) (b) (c)

Fig. 13. Cutting pattern on metal board: (a) metal board with concrete base, (b) marker in acoustic image, and (c) coding pattern for ID recognition.

current system setting. MSER detection is one of the main reasons why small markers cannot be detected; this requires improvement in the future. In the water tank, we also cut patterns in the metal board. Cutting pattern in the metal with a concrete base has a better contrast. And some coding techniques [5] can be used for ID recognition as shown in Fig. 13(c) in the future.

B. Computation time

Using a PC with an Intel(R) Core(TM) i7-8750H CPU @ 2.20 GHz, it takes approximately 125 ms for one image in the real experiment. The detection and ID identification process has a duration of approximately 90 ms; this is because MSER is time-consuming. Currently, the system reaches a performance of 8 Hz under the current resolution. It should be mentioned that the path to the target in the acoustic image is blank, and it is possible to trim the image to increase the speed. For multiple markers, the ID identification and pose estimation process can be computed in parallel.

IX. CONCLUSIONS

In this paper, we proposed a fiducial marker system named ACMarker. The system includes detection of the marker, ID identification, and pose estimation based on the marker. The Experiment results proved that 5DoF can be estimated accurately and precisely with corner information. If IA information is included, 6DoF can be estimated successfully. The marker can be applied to tasks such as AR, visual localization, and landmark-based SLAM, which facilitate underwater research. In future work, the detection bottleneck will be examined to increase the robustness and computation speed. The segmentation based on MSER is slow and unstable. Moreover, some coding techniques may be considered for ID recognition. Learning-based methods will be explored to increase the performance of the system. One of the drawbacks of the current pose estimation method is that t_3 cannot be accurately estimated without IA information, whereas IA is based on the assumption that the marker locates on a flat surface. Future work may also include the detection of IA from non-flat surface, or the use of additional constraints to refine the pose estimation result. It was found that the backscattered intensity of the diffuse material on the marker changed with changing pose between the marker and the acoustic camera. Considering the intensity information (i.e., photometric information) may improve the current pose estimation method. Currently, placing markers on existed structures may require human effort. It may also be interesting to realize automatic placement of the markers based on underwater vehicles in the future.

REFERENCES

[1] J. Wang, T. Shan, and B. Englot, "Underwater terrain reconstruction from forward-looking sonar imagery," *Proc. IEEE Int. Conf. on Robot. Autom.*, pp. 3471–3477, May 2019.

[2] N. Hurtos, D. Ribas, X. Cufi, Y. Petillot, and J. Salvi, "Fourier-based registration for robust forward-looking sonar mosaicing in low-visibility underwater environments," *J. Field Robot.*, vol. 32, no. 1, pp. 123–151, 2015.

[3] 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.

[4] M. Fiala, "Artag, a fiducial marker system using digital techniques," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 590–596, Jun. 2005.

[5] E. Olson, "Apriltag: A robust and flexible visual fiducial system," *Proc. IEEE Int. Conf. on Robot. Autom.*, pp. 3400–3407, May 2011.

[6] S. Garrido-Jurado, R. Munoz-Salinas, F. Madrid-Cuevas, and M. Marin-Jimenez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognit.*, vol. 47, no. 6, pp. 2280–2292, Jun. 2014.

[7] Y. Lee, J. Choi, N. Y. Ko, and H.-T. Choi, "Probability-based recognition framework for underwater landmarks using sonar images," *Sensors*, vol. 17, no. 9, p. 1953, Sep. 2017.

[8] 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," *Proc. MTS/IEEE conf. OCEANS 2020 Singapore*, Aug. 2020.

[9] E. Westman, A. Hinduja, and M. Kaess, "Feature-based slam for imaging sonar with under-constrained landmarks," *Proc. IEEE Int. Conf. on Robot. Autom.*, pp. 3629–3636, May 2018.

[10] J. Pyo, H. Cho, and S. Yu, "Beam slice-based recognition method for acoustic landmark with multi-beam forward looking sonar," *IEEE Sens. J.*, vol. 17, no. 21, pp. 7074–7085, Nov. 2017.

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

[12] N. Brahim, D. Guériot, S. Daniel, and B. Solaiman, "3d reconstruction of underwater scenes using didson acoustic sonar image sequences through evolutionary algorithms," *OCEANS 2011 IEEE-Spain*, pp. 1–6, Jun. 2011.

[13] 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.

[14] B. T. Henson and Y. V. Zakharov, "Attitude-trajectory estimation for forward-looking multibeam sonar based on acoustic image registration," *IEEE J. Oceanic Eng.*, vol. 44, no. 3, pp. 753–766, Jul. 2019.

[15] N. T. Mai, H. Woo, Y. Ji, Y. Tamura, A. Yamashita, and H. Asama, "3-d reconstruction of line features using multi-view acoustic images in underwater environment," *Proc. IEEE Int. Conf. Multisens. Fusion and Integr. for Intell. Sys.*, pp. 312–317, Nov. 2017.

[16] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vision Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[17] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *The Canadian Cartographer*, vol. 10, no. 2, pp. 112–122, 1973.

[18] S. Negahdaripour, "Visual motion ambiguities of a plane in 2-d fs sonar motion sequences," *Comput. Vis. Image Underst.*, vol. 116, no. 6, pp. 754–764, Jun. 2012.

[19] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cyber.*, vol. 9, no. 1, pp. 62–66, 1979.

[20] J. R. Cardoso and K. Zietak, "On a sub-stiefel procrustes problem arising in computer vision," *Numer. Linear Algebra Appl.*, vol. 22, no. 3, pp. 523–547, Feb. 2015.

[21] A. Bartoli and T. Collins, "Plane-based resection for metric affine cameras," *J. Math. Imaging Vis.*, vol. 60, no. 7, pp. 1037–1064, 2018.

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

[23] N. T. Mai, Y. Ji, H. Woo, Y. Tamura, A. Yamashita, and H. Asama, "Acoustic image simulator based on active sonar model in underwater environment," *Proc. 15th Int. Conf. Ubiquitous Robots*, pp. 775–780, Jun. 2018.