Weng, Yang; Matsuda, Takumi; Sekimori, Yuki; Pajarinen, Joni; Peters, Jan; Maki, Toshihiro

Sim-To-Real Transfer for Underwater Wireless Optical Communication Alignment Policy between AUVs

Published in:
OCEANS 2022 - Chennai

DOI:
10.1109/OCEANSChennai45887.2022.9775437

Published: 19/05/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Weng, Y., Matsuda, T., Sekimori, Y., Pajarinen, J., Peters, J., & Maki, T. (2022). Sim-To-Real Transfer for Underwater Wireless Optical Communication Alignment Policy between AUVs. In *OCEANS 2022 - Chennai* (Ocean). IEEE. https://doi.org/10.1109/OCEANSChennai45887.2022.9775437

Sim-to-Real Transfer for Underwater Wireless Optical Communication Alignment Policy between AUVs

Yang Weng[1,*], Takumi Matsuda[1,2], Yuki Sekimori[1], Joni Pajarinen[3,4], Jan Peters[3], Toshihiro Maki[1]
1. Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
2. School of Science and Technology, Meiji University, Tokyo, Japan
3. Intelligent Autonomous Systems Laboratory, TU Darmstadt, Darmstadt, Germany
4. Department of Electrical Engineering and Automation, Aalto University, Espoo, Finland
*yangweng@iis.u-tokyo.ac.jp

***Abstract*—The underwater wireless optical communication (UWOC) technology provides a potential high data rate solution for information sharing between multiple autonomous underwater vehicles (AUVs). In order to deploy the UWOC system on mobile platforms, we propose to solve the optical beam alignment problem by maintaining the relative position and orientation of two AUVs. A reinforcement learning based alignment policy is transferred to the real world since it outperforms other baseline approaches and shows good performance in the simulation environment. We randomize the simulator and introduce the disturbances, aiming to cover the real distribution of the underwater environment. Soft actor-critic (SAC) algorithm, reward shaping based curriculum learning, and specifications of the vehicles are utilized to achieve the successful transfer. In the Hiratsuka sea experiments, the alignment policy was deployed on the AUV TriTON and successfully aligned with autonomous surface vehicle BUTTORI. It demonstrates a solution for combining the UWOC technology and AUVs team in the ocean investigation.**

***Index Terms*—underwater wireless optical communication, reinforcement learning, AUV, sim-to-real transfer**

I. Introduction

Compared with relying on a single and expensive AUV, the multiple AUVs deployment has many advantages in underwater exploration, such as observation efficiency, spatial scale, safety, and cost [1]. Several issues related to multiple AUVs, like formation control algorithms [2] [3], and cooperative localization methods [4] [5], have become attractive. It aims to make the operation of multiple AUVs become the standard in the upcoming years. The communication between multiple AUVs is a significant issue since the data rate of acoustic transmission is limited to 1 - 100 Kilobit per second (Kbps) [6]. During the joint investigation, it is not realistic for AUVs to share the collected observation data by acoustic communication.

The development of the UWOC technology provides a potential high data rate solution for information sharing between multiple AUVs [7]. However, UWOC requires the establishment and maintenance of a line-of-sight (LOS) link for communication. Underwater LOS link alignment is a complex problem involving the motion of the platforms, the observation of the target, and the control of the optical communication devices. In involving AUVs scenarios, maintenance becomes challenging due to the external disturbances and uncertainties in the AUV dynamic model. Previous researches proposed the beam pointing control system to steer the beam for scanning and link acquisition [8] [9]. Based on the detected light intensity, these methods rapidly adjust the beam pointing to maintain the LOS link. It attempts to eliminate the uncertainties in AUVs and the environmental disturbances with precise control of the optical devices.

Deep reinforcement learning algorithm has recently seen success in suppressing the impact of external disturbances and uncertainties in robotics motion planning [10] [11]. It is attractive to consider this complex beam alignment problem under a model-free reinforcement learning framework. We trained a reinforcement learning policy to keep two AUVs in a specific relative position and orientation for alignment. Compared with the previous researches, we manipulate the AUVs instead of relying on sophisticated optical devices, such as beam control servo and light intensity sensors. Besides, the navigation and energy saving issues can also be optimized through trial-and-error processes.

Due to the limitations of gathering data from a real environment, reinforcement learning algorithms usually use a simulation environment to train agents. However, the gap between the simulation and real environment degrades the performance once the policy is implemented in the real world [12]. In this research, we propose to deploy a learned policy on real AUVs for optical beam alignment. We randomize the simulation environment and introduce the disturbances, aiming to cover the real distribution of the underwater environment data. The curriculum learning and reward shaping techniques are utilized to improve stability in the real environment. An AUV and its operating system are developed to implement this policy in the real environment. The success of the sea experiment proves that the beam alignment policy learned from the simulation environment can be applied to the real environment.

The remaining of the paper is organized as follows. The underwater optical beam alignment problem is modeled in Section II. In Section III, the sim-to-real work is presented for implementing the policy on real AUVs. Then, a reinforcement

learning policy is trained for alignment task and evaluated in Section IV. In Section V, the alignment experiments are conducted in the real environment. Concluding remarks are given in Section VI.

II. Problem Formulation

An alignment method is presented for two AUVs to establish the LOS link. The reinforcement learning algorithm searches for an optimal policy to complete this alignment task.

A. Beam Alignment

Optical signal has limited propagation distance and strong directivity. The wireless optical communication requires an AUV with a transmitter to emit a light signal to the receiver of another AUV. It is assumed that the receiver is omnidirectional in this research. We propose to solve the alignment problem by maintaining the relative position and orientation. The acoustic navigation is used for observing the states of the alignment target. As long as the position error between two AUVs is still within the coverage area of the optical beam, the LOS link can be successfully established.

As shown in Fig. 1, the alignment task is considered at a horizontal plane with a horizontal position $[x, y]$, surge velocity u, sway velocity v, yaw orientation ψ, and yaw angular velocity r because the pressure sensor can provide an accurate determination of absolute depth [8]. The AUV that transmits optical signals is regarded as the transmitting AUV. The AUV that receives optical signals is defined as the receiving AUV. The motion in roll and pitch orientation is ignored. The transmitting AUV is the agent we discussed in the reinforcement learning algorithm. In the alignment task, the receiving AUV is expected to be close to the optimal point of beam coverage area, where it can detect the high intensity optical signals. The optimal point can be determined by the light field distribution of the optical beam [13]. The alignment distance d_Δ is defined as the distance between the optimal point and the center of the receiving AUV.

The acoustic navigation, including the bearing only ranging [14] and the two-way travel time (TWTT) ranging [15], is used by the transmitting AUV to observe and track the receiving AUV. The states of the receiving AUV $[x^R, y^R, \psi^R, u^R, v^R, r^R]$ can be shared to the transmitting AUV through acoustic signals. The bearing only ranging is a scalable method. The receiving AUV can broadcast the acoustic signal, and all the platforms can measure the relative bearing angle α^{TR}. Both relative bearing angle and relative distance l^{TR} can be measured if the transmitting AUV requests the TWTT ranging. The scalability of the TWTT ranging is weak, and it is better to reduce the usage of this method in multiple AUVs operations. For saving energy, we also hope the optical transmitter can be turned off when it is impossible to establish the LOS link.

B. Alignment Policy

The goal of the alignment policy is to control the transmitting AUV to shorten the alignment distance d_Δ for maintaining the LOS link while also reducing the use of acoustic channel resources and optical devices.

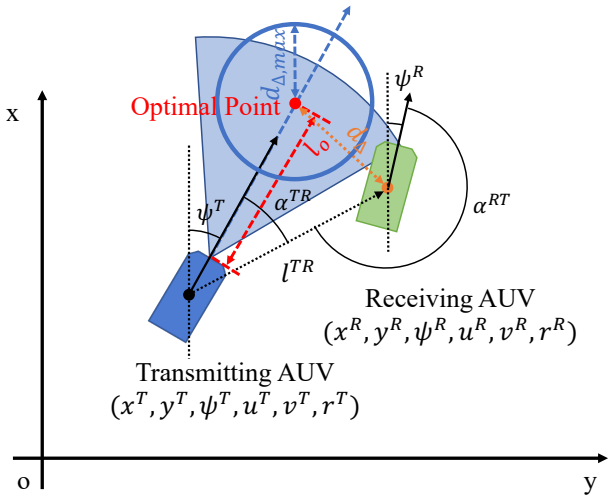


Fig. 1. Relative relationship for underwater optical beam alignment. The blue sector is the coverage area of the optical beam emitted by the transmitting AUV. The receiving AUV needs to be located in this blue sector and detect the optical signals for communication.

The state space of the agent is defined as:

$$s = [\hat{x}_\Delta, \hat{y}_\Delta, \cos\psi^R, \sin\psi^R, u^R, v^R, r^R, \cos\hat{\psi}^T, \sin\hat{\psi}^T] \tag{1}$$

where $[x_\Delta, y_\Delta]$ is the alignment distance. The variables with hat symbols are updated by a particle filter estimator. All variables in the state space are one-dimensional and continuous.

The action space of the agent is as follows:

$$a = [u^T, r^T, i_{twtt}, i_{op}] \tag{2}$$

where the boolean variables i_{twtt} and i_{op} represent whether the transmitting AUV requests for TWTT ranging, and whether to turn on the optical transmitter in the current timestep, respectively.

We propose the reward function of the form:

$$\begin{aligned} r(s,a) = &- \rho_1(1+\rho_2 i_{twtt})(1+\rho_3 i_{op}) d_\Delta^{\frac{1}{2}} \\ &- \rho_4 u_\Delta - \rho_5 r_\Delta + \rho_6 i_{done} \end{aligned} \tag{3}$$

where ρ_1 to ρ_6 are coefficients. The u_Δ and r_Δ represent the relative velocities in surge and yaw. A boolean variable i_{done} is used to indicate if the alignment task is completed.

III. Sim-to-Real

A policy learned from the simulation environment needs to be deployed on the real AUVs. It is unrealistic to design a simulator that perfectly matches the real environment. The gap between the simulation and the real world may degrade the adaptability of the policy in the real environment. The following technologies are used to transfer the knowledge learned in the simulation to the real world.

Algorithm 1 Soft Actor-Critic [16]

Input:
Initial parameters of critic and actor networks θ_1, θ_2, ϕ
Initial weights of target networks $\overline{\theta}_1 \leftarrow \theta_1, \overline{\theta}_2 \leftarrow \theta_2$
Empty replay buffer $\mathcal{D} \leftarrow \emptyset$
for each iteration **do**
  for each environment step **do**
    Sample action by $a_t \sim \pi_\phi(a_t|s_t)$
    Sample transition state by $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$
    Store samples by $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
  end for
  for each gradient step **do**
    Update critic by $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    Update policy by $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    Adjust temperature by $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$
    Update target by $\overline{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\overline{\theta}_i$ for $i \in \{1, 2\}$
  end for
end for
Output: θ_1, θ_2, ϕ

A. Reinforcement Learning Algorithm

As listed in Algorithm 1, the SAC algorithm presented by Haarnoja *et al.* [16] is selected to search for an optimal policy that can collect not only the maximum cumulative reward, but also the maximum entropy. The objective function is as follows:

$$\pi^* = \arg\max_\pi \sum_{t=0}^{\infty} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[\sum_{l=t}^{\infty} \gamma^{l-t} \mathbb{E}_{s_l\sim p, a_l\sim\pi} M(s_t, a_t)] \tag{4}$$

and

$$M(s_t, a_t) = r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))|s_t, a_t \tag{5}$$

where $\mathbb{E}$ is the expectation operation and $\mathcal{H}$ is the entropy term.

The entropy term is a measure of randomness, which encourages the policy to explore more widely. It also provides a robust framework that minimizes the need for hyperparameter tuning when transferring to a real environment [17].

In the alignment task, the policy needs to generate the surge and yaw angular velocity commands for AUV thrusters. The thrusters cannot accurately perform actions when the high frequency jitter occurs in the velocity command. One of the advantages of the SAC algorithm is that it correlates the exploration temporally and can output smoothing actions for thrusters [17].

B. Domain Randomization and Disturbances

Domain randomization is an approach to bridge the reality gap for reinforcement learning. When it is impossible to make the simulation environment match the real environment, we can highly randomize the simulator. We design the episode based alignment experiments that can be efficiently repeated by the simulator. The domain randomization is implemented in the simulator. The initial position and orientation of the AUVs are randomized in each episode. The velocity of the transmitting AUV is determined by the policy, while the velocity of the receiving AUV is randomly generated. With enough variability in the simulator, the real world may appear to the model as just another variation [18].

The perturbations are introduced into each timestep. These perturbations can be caused by external disturbances and the uncertainties of vehicles. On the sensing part, the measured velocities in surge, sway, and yaw are the mixing results of real velocities and Gaussian noises, whose standard deviations are 0.1, 0.1, and 1, respectively. On the manipulation part, the real surge, sway, and yaw angular velocities are derived from the policy actions mixed with the same Gaussian noises.

C. Reward Shaping Based Curriculum Learning

Curriculum learning is an extension of transfer learning, where the goal is to gradually changes the task from simple to complex [19]. We consider the task in a simulation environment as a simple task, while the alignment in the real environment is a complex task. Through curriculum learning, the agent finally obtains the ability to complete the alignment task in the real environment. In the alignment task, the transmitting AUV first learns to track the target, and then reduces the use of acoustic channel resources and optical devices. In the future, the experimental data collected in the real world can be used to train the policy again.

The coefficients in (3) need to be tuned according to the importance of different controlling objectives. The reward function with coefficients are proposed as follows:

$$r_1(s, a) = -0.01 d_\Delta^{\frac{1}{2}} - 0.01 u_\Delta - 0.002 r_\Delta \tag{6}$$

and

$$\begin{aligned} r_2(s, a) = &- 0.01 d_\Delta^{\frac{1}{2}} (1 + 9 i_{twtt})(1 + i_{op}) \\ &- 0.01 u_\Delta - 0.002 r_\Delta + 10 i_{done} \end{aligned} \tag{7}$$

IV. Implementation

The sample data collected from the simulator is used for the alignment policy. The learned policy is compared with the heuristic baseline approach before deploying on real AUVs.

A. Policy Training

The SAC algorithm is implemented with the OpenAI Stable Baselines toolkit [20]. The neural networks use Multilayer Perceptron (MLP) structure. The parameters used in the reinforcement learning algorithm are given in Table I. The simulation environment is developed through the OpenAI Gym interface [21].

The action i_{twtt} and i_{op} are used for saving the acoustic resource and energy, which are not considered in this sea experiment. The agent learns 3×10^6 timesteps of sample data with reward r_1.

TABLE I
THE PARAMETERS CONFIGURED BY THE REINFORCEMENT LEARNING ALGORITHM

Parameter	Symbol	Value
Layer of MLP		2
Neuron of MLP		64
Discount factor	γ	0.99
Learning rate	λ	0.0003
Buffer size		50000
Batch size		64

B. Policy Evaluation

To evaluate the significance of the reinforcement learning based method, we compared learned policy with a heuristic baseline approach. The heuristic approach is derived from the motion planning method used in previous experiments by Maki *et al.* [22]. The performance of this heuristic method is verified by the sea experiments.

The details of the comparison with the heuristic approach are presented in the previous research [23]. The reinforcement learning approach proposed in this research outperforms the heuristic approach in alignment efficiency and energy saving.

V. EXPERIMENTS

In order to evaluate whether the alignment policy can be deployed on real AUVs, we implemented the policy in the water tank and sea experiments.

A. Preparation

The learned policy is deployed on the hovering AUV Tri-TON, and the specifications of the vehicle are given in Table II. In the real experiments, the AUV Tri-TON is considered as the transmitting AUV, and the autonomous surface vehicle BUTTORI is used as the receiving AUV. According to the specifications of the AUV Tri-TON, the maximum surge and yaw angular velocities are set to 0.2 m/s and 0.2 rad/s, respectively. The TWTT ranging between AUV Tri-TON and BUTTORI is performed every 6 seconds. No global navigation satellite system (GNSS) or radio communications are used in the experiments.

B. Water Tank Testing

The alignment experiments are tested in the water tank. The size of the water tank is 8 meters long, 8 meters wide, and 8 meters deep. As shown in Fig. 2, the AUV Tri-TON and the autonomous surface vehicle BUTTORI are deployed for experiments.

The BUTTORI keeps stationary in the alignment experiment. The AUV Tri-TON performs the actions generated by the learned policy, including the commands of surge and yaw angular velocities. The AUV Tri-TON is required to align with the BUTTORI and maintain the relative position and orientation.

TABLE II
AUV TRI-TON SPECIFICATIONS

Parameter	Value (Device)
Size	1.40 m (L) × 1.33 m (H) × 0.76 m (W)
Mass	230 kg
Max. speed	0.5 m/s
Max. depth	800 m
Duration	8 hours
Thruster	100 W thruster × 5
Battery	LiIon 26.6 V 25 Ah × 4
Ground velocity	Teledyne RDI Navigator 1200 kHz (DVL)
USBL	SeaTrac X150
FOG	JAE JG-35FD
Depth	Mensor DPT6000
Main computer	UP Core
Operation system	Ubuntu 20.04
Middleware suite	Robot Operating System
CPU	Intel Atom x5-z8350

The alignment distance d_{Δ} of one episode is plotted in Fig. 3. When the policy is activated, it controls the AUV to approach the target, and the alignment distance keeps decreasing. At the 50 second of the experiment, the alignment distance is close to 0, which indicates that the vehicle can establish a LOS link with the target. The AUV Tri-TON successfully maintained the required relative position and orientation with BUTTORI for more than 250 seconds.

C. Sea Experiments

As shown in Fig. 4, we conducted the sea experiments at Hiratsuka Port, Japan. Compared with the water tank, the disturbance and uncertainty in the marine environment are significant. In addition, we will move the target BUTTORI in the experiments to test whether the AUV Tri-TON can track and maintain alignment with the target.

D. Results

One of the sea experiments is presented in Figs. 5 and 6. The AUV Tri-TON is represented by the blue triangle, and the red star marker is the autonomous surface vehicle BUTTORI. The particle filter estimation results of the AUV Tri-TON are depicted by blue dots. The orange point is the optimal point defined in Fig. 3 for optical communication. The yellow circle represents the result of acoustic ranging.

The parameters listed in the bottom left corner are the time in the experiment (second), the position of the AUV Tri-TON (meter), the standard deviation of particle filter estimation results in the position, the yaw orientation of the AUV Tri-TON ψ^T (degree), the standard deviation of particle filter estimation results in yaw, the surge velocity command generated by reinforcement learning policy (meter per second), the surge velocity measured by DVL (meter per second), the

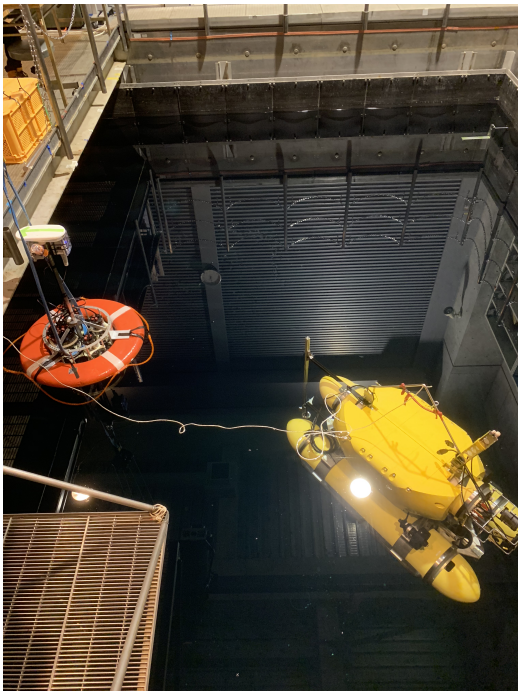


Fig. 2. Water tank experiments. The AUV Tri-TON (yellow) and the autonomous surface vehicle BUTTORI (orange) are deployed in the water tank.

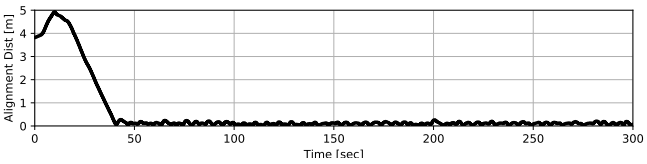


Fig. 3. The alignment distance d_{Δ} in one episode.

Fig. 4. Sea experiments at Hiratsuka Port. The AUV Tri-TON (yellow) and the autonomous surface vehicle BUTTORI (orange) are deployed in the sea environment.

yaw angular velocity command generated by reinforcement learning policy (radian per second), the yaw angular velocity measured by FOG (radian per second), and the alignment distance d_{Δ} (meter). The parameters listed in the top left corner are the number of TWTT ranging in the experiment, the time when the latest ranging results are received (second), the position of the BUTTORI (meter), the relative bearing angle measured by the USBL device in the AUV Tri-TON (degree), the relative bearing angle measured by USBL device in the BUTTORI (degree), and the relative distance (meter).

In the alignment task, AUV Tri-TON starts to approach BUTTORI with the guidance of acoustic navigation. At the 17.0 second of the experiment, the alignment distance d_{Δ} shown in Fig. 5(b) is 0.34 meters. The relative relationship between Tri-TON and BUTTORI is available for establishing the LOS link. The AUV can keep the alignment distance at about 1 meter. The yaw angle of AUV Tri-TON is not well controlled by policy, which is the reason for the large alignment distance. As shown in Fig. 5(c), the largest deviation occurs at 97.0 second. The vehicle inertia and the thrusters delay are the main reasons.

At the 123.0 second of the experiment, we move the BUTTORI to test if the Tri-TON can track the target. The alignment distance increases during the tracking process. Due to the delay of acoustic transmission, the yellow circle represented by acoustic ranging cannot coincide with the Tri-TON. At 199.0 second of the experiment, we stop controlling BUTTORI, and it is only affected by the currents. As shown in Fig. 6(c), the AUV can align with the target, and the alignment distance is 0.20 meters.

E. Discussion

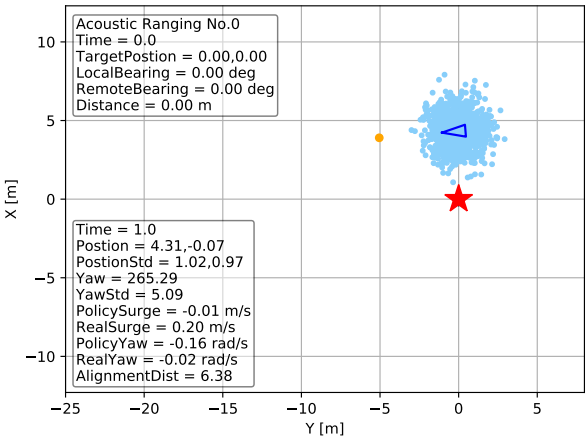
The alignment policy learned from the simlation environment is deployed on the real AUVs. In the Hiratsuka Port experiments, this policy can manipulate the AUV Tri-TON to align with the target BUTTORI.

The results of sea experiments show that the policy does not handle the delay from the thrusters and acoustic ranging well. In the future, we plan to repeat the alignment experiments in the water tank to obtain the statistics of the thrusters and acoustic ranging delay, which can be used to improve the policy. The thrusters delay will be taken into account in the simulation environment. The particle filter will consider the delay from acoustic ranging in state estimation. In addition, the data collected in the sea experiments can also be utilized to retrain the alignment policy.

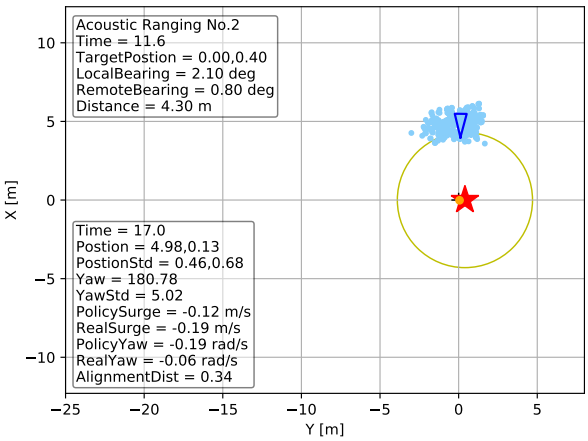
In order to implement the UWOC in actual scenarios, it is necessary to pay attention to the adaptability of the policy in different marine environments. The alignment policy needs to be evaluated under more complex sea conditions.

VI. Conclusion

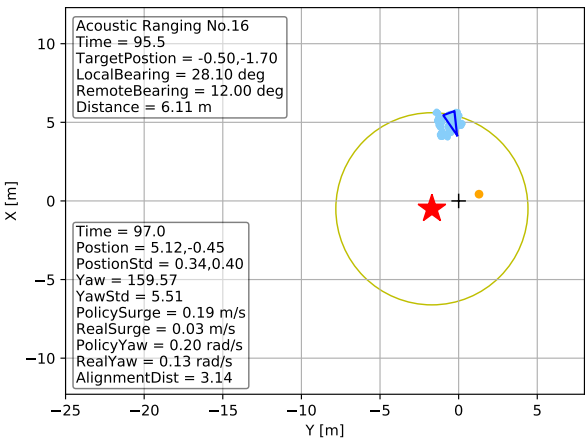
We trained a reinforcement learning policy for establishing the LOS link between two AUVs in a simulation environment and planned to deploy it on real AUVs. The sim-to-real methods we discussed in this research reduce the gap between



(a) Sea experiment at 1.0 s

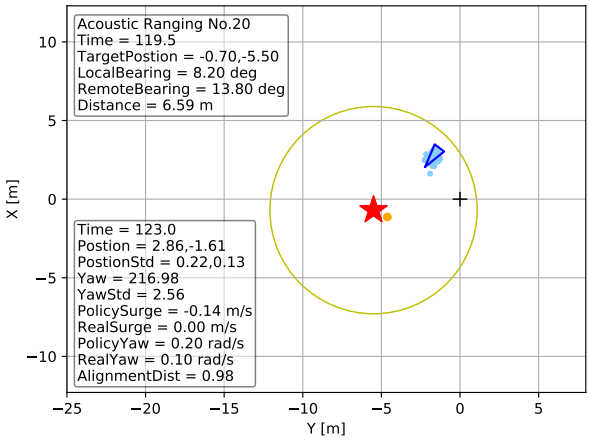


(b) Sea experiment at 17.0 s

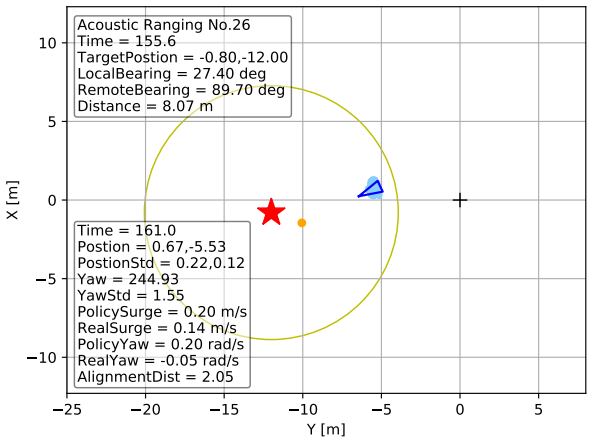


(c) Sea experiment at 97.0 s

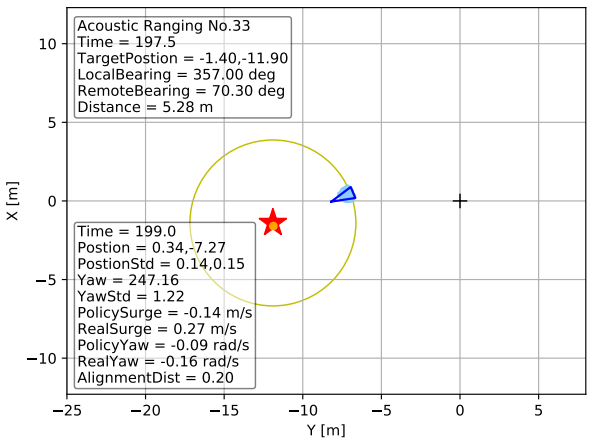
Fig. 5. The states of AUV in the sea experiment (a) 1.0 s, (b) 17.0 s, and (c) 97.0 s. The AUV Tri-TON is represented by the blue triangle, and the sharp corner of the triangle is the head of the vehicle. The red star marker is the autonomous surface vehicle BUTTORI.



(a) Sea experiment at 123.0 s



(b) Sea experiment at 161.0 s



(c) Sea experiment at 199.0 s

Fig. 6. The states of AUV in the sea experiment (a) 123.0 s, (b) 161.0 s and (c) 199.0 s. The AUV Tri-TON is represented by the blue triangle, and the sharp corner of the triangle is the head of the vehicle. The red star marker is the autonomous surface vehicle BUTTORI.

the simulation and the real environment, allowing us to transfer the learned policy to the real world. In sea experiments, the alignment policy successfully manipulates the AUV TriTON to align with the target BUTTORI. It demonstrates that conveniently training the alignment policy in a simulation environment and deploying it on a real AUV is suitable for underwater optical communication research.

REFERENCES

[1] R. Cui, S. S. Ge, B. V. E. How, and Y. S. Choo, "Leader–follower formation control of underactuated autonomous underwater vehicles," *Ocean Engineering*, vol. 37, no. 17-18, pp. 1491–1502, 2010.

[2] E. Fiorelli, N. E. Leonard, P. Bhatta, D. A. Paley, R. Bachmayer, and D. M. Fratantoni, "Multi-auv control and adaptive sampling in monterey bay," *IEEE journal of oceanic engineering*, vol. 31, no. 4, pp. 935–948, 2006.

[3] C. Yuan, S. Licht, and H. He, "Formation learning control of multiple autonomous underwater vehicles with heterogeneous nonlinear uncertain dynamics," *IEEE transactions on cybernetics*, vol. 48, no. 10, pp. 2920–2934, 2017.

[4] A. Bahr, J. J. Leonard, and M. F. Fallon, "Cooperative localization for autonomous underwater vehicles," *The International Journal of Robotics Research*, vol. 28, no. 6, pp. 714–728, 2009.

[5] Z. J. Harris and L. L. Whitcomb, "Preliminary evaluation of cooperative navigation of underwater vehicles without a dvl utilizing a dynamic process model," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[6] J.-g. Huang, H. Wang, C.-b. He, Q.-f. Zhang, and L.-y. Jing, "Underwater acoustic communication and the general performance evaluation criteria," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 8, pp. 951–971, 2018.

[7] S. Zhu, X. Chen, X. Liu, G. Zhang, and P. Tian, "Recent progress in and perspectives of underwater wireless optical communication," *Progress in Quantum Electronics*, p. 100274, 2020.

[8] N. D. Hardy, H. G. Rao, S. D. Conrad, T. R. Howe, M. S. Scheinbart, R. D. Kaminsky, and S. A. Hamilton, "Demonstration of vehicle-to-vehicle optical pointing, acquisition, and tracking for undersea laser communications," in *Free-Space Laser Communications XXXI*, vol. 10910. International Society for Optics and Photonics, 2019, p. 109100Z.

[9] P. B. Solanki, S. D. Bopardikar, and X. Tan, "Active alignment control-based led communication for underwater robots," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1692–1698.

[10] R. Cui, C. Yang, Y. Li, and S. Sharma, "Adaptive neural network control of auvs with control input nonlinearities using reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 6, pp. 1019–1029, 2017.

[11] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[12] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 737–744.

[13] S. Chandrasekhar, "Dover books on intermediate and advanced mathematics," 1960.

[14] J. Vaganay, P. Baccou, and B. Jouvencel, "Homing by acoustic ranging to a single beacon," in *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158)*, vol. 2. IEEE, 2000, pp. 1457–1462.

[15] N. H. Kussat, C. D. Chadwell, and R. Zimmerman, "Absolute positioning of an autonomous underwater vehicle using gps and acoustic measurements," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 1, pp. 153–164, 2005.

[16] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[17] T. Haarnoja, V. Pong, K. Hartikainen, A. Zhou, M. Dalal, and S. Levine, "Soft actor critic—deep reinforcement learning with real-world robots," 2018, https://bair.berkeley.edu/blog/2018/12/14/sac.

[18] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[19] S. Narvekar and P. Stone, "Learning curriculum policies for reinforcement learning," *arXiv preprint arXiv:1812.00285*, 2018.

[20] A. Hill, A. Raffin, M. Ernestus, A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Stable baselines," https://github.com/hill-a/stable-baselines, 2018.

[21] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[22] T. Maki, H. Mizushima, H. Kondo, T. Ura, T. Sakamaki, and M. Yanagisawa, "Real time path-planning of an auv based on characteristics of passive acoustic landmarks for visual mapping of shallow vent fields," in *OCEANS 2007*. IEEE, 2007, pp. 1–8.

[23] Y. Weng and T. Maki, "Observability analysis of underwater wireless optical communication alignment between auvs," in *OCEANS 2021- San Diego*. IEEE, 2021, pp. 1–6.