Autonomous Surface Vehicle Control Method Using Deep Reinforcement Learning

Shang Zhang
College of Engineering
Ocean University of China
Qingdao, China
zhangshangouc@126.com

Rui Yang
College of Engineering
Ocean University of China
Qingdao, China
yangrui@ouc.edu.cn

Zhen Chen
College of Engineering
Ocean University of China
Qingdao, China
oucchenzhen@ouc.edu.cn

Ming Li
College of Engineering
Ocean University of China
Qingdao, China
limingneu@ouc.edu.cn

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327026

***Abstract*—Autonomous Surface Vehicle (ASV) provides a new platform for marine exploration and environmental monitoring. It is very important for ASV to have learning ability in the unknown environment. Reinforcement learning is a branch of machine learning. ASV can obtain control behaviors and improve adaptability and autonomy through environmental exploration. This paper carries out dynamic modeling on the four-thruster ASV, especially designs a controller using the DDPG (deep deterministic policy gradient) algorithm. Simulation results show that the DDPG controller can control MIMO (multiple-input multiple-output) nonlinear systems. After training, the ASV can perform fixed-point control, sinusoidal trajectory tracking, and the "reconfigurable" experiments of multiple ASVs in the absence of water flow or water flow interference. The algorithm proposed in this paper has strong robustness and lays a foundation for the research of cooperative control of multiple ASVs.**

Keywords—ASV, DDPG, fixed-point control, trajectory tracking, reconfigurable

I. INTRODUCTION

As countries attach importance to marine development, the demand for maritime safety protection, hydrometeorological information collection, and scientific research has increased significantly. Autonomous Surface Vehicle (ASV) is an important platform for ocean exploration, whose advantages are small size and flexible movement. In the "Roboat" project, the main goal is to enable multiple "Roboats" to build bridges and stages, which can be connected into a floating platform, and to transport goods and waste in the canals of Amsterdam [1]. As shown in Fig. 1, multiple ASVs are combined into different shapes and can adapt to different operating scenarios. They can quickly form flexible transportation channels, build automatic pontoons, and form operating platforms. They can also resist greater interference and improve the stability of system. This paper focuses on the modeling and controller design of the four-thruster ASV and shows the "reconfigurable" capabilities of the four ASVs.

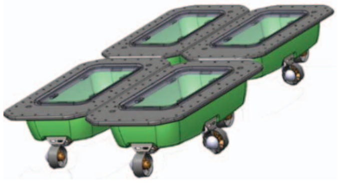


Fig. 1. Multiple ASVs are combined into different shapes

Multiple ASVs with "reconfigurable" capabilities require a reasonable number of thrusters. The ASV in [2] uses one thruster for driving and one steering gear to achieve steering, which is also a common scheme in current research. The disadvantage is that ASV has a large turning radius and cannot adjust the attitude alone. The ASV designed in [3] is equipped with two thrusters. The thrusters are bilaterally symmetrical, and the forward and steering are realized by controlling the thrust respectively. Although the attitude of ASV can be adjusted flexibly, it cannot complete the sway motion. Compared with [2] and [3], ASV adopts the four-thruster scheme in [4]. The thrusters are distributed in an “X” shape. ASV can realize surging, swaying, and yawing movements, and the flexibility is greatly improved. The four thrusters of the ASV in [5] are distributed in a "+" shape. The ASV has strong maneuverability and higher propulsion efficiency, and a nonlinear model predictive control (NMPC) scheme is proposed, which can control accurately in the indoor and outdoor environment. Reference [1] proposes a reconfigurable feedback control system for multiple ASVs using four-thruster ASV. The platform consists of multiple rectangular vessels where each is capable of latching to a pre-defined point of another vessel. ASV is a multi-input, multiple-output nonlinear system, and developing a high-performance controller is a huge challenge. However, traditional PID control, nonlinear model predictive control, and other traditional control strategies have high requirements on the environment model and don’t have the ability to explore and learn the unknown environment. With the gradual development of theory and technology, especially in reinforcement learning and deep learning, the development of an unmanned system has been greatly improved. The Q network was designed by observing the state and action space, and a deep reinforcement learning method is proposed for ASV control in [6]. Reference [7] proposes a deep learning model that can directly learn control strategies from high-dimensional perception inputs for reinforcement learning, and with the help of relevant theories of reinforcement learning, there has been some progress in researching autonomous navigation and obstacle avoidance. Reference [8] proposes a decentralized multi-agent obstacle avoidance algorithm based on deep reinforcement learning, which can effectively apply online learning to offline learning. In order to improve the stability and flexibility of ASV's movement, and has the ability of position and attitude control, trajectory tracking, and

This work was supported by the National Natural Science Foundation of China (No. 51709245), the National Key Research and Development Program of China (No. 2017YFC1405203), Qingdao Post-Doctoral Fund (No. 861705040024).

the "reconfigurable" ability of multiple ASVs, this paper designs a DDPG controller, and analyzes the results of simulation experiments to prove the effectiveness of the proposed control strategy.

The structure of this paper is as follows: firstly, the kinematics and dynamics model of the four-thruster ASV are established, and the water flow interference model is added to the dynamics model; secondly, the DDPG controller is designed and the reward function is given; then, the DDPG controller is trained, and the simulation experiments of fixed-point control, trajectory tracking and " reconfigurable " of multiple ASVs are designed, and the experimental results are analyzed; finally, we draw conclusions and put forward the next research direction.

II. Establishment of ASV's Model

A. *ASV overall structure analysis*

In this paper, a small ASV with four thrusters is used for experimental analysis, and the cooperation of four thrusters can ensure the flexibility and stability of the ASV. We establish the fixed coordinate system *E-XYZ* and the motion coordinate system $O-\xi\eta\varsigma$ for ASV. The structure of the system is shown in Fig. 2.

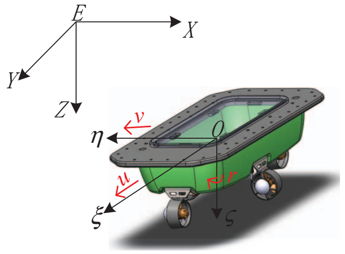


Fig. 2. The structure of the ASV system

The origin of the fixed coordinate system *E-XYZ* is taken at a certain point on the surface of the water, *EX* and *EY* are on the horizontal plane, and *EZ* points to the center of the earth. In the fixed coordinate system, the position and attitude of ASV are described by$[x \quad y \quad \psi]^T$; the origin of the motion coordinate system is taken at the center of gravity of the ASV. The movement of ASV along$O\xi$, $O\eta$ and rotation around $O\varsigma$ is described by $[u \quad v \quad r]^T$. The forces and moments corresponding to the three directions are described by$[\tau_u \quad \tau_v \quad \tau_r]^T$.

ASV completes the surging (*u*), swaying (*v*) and yawing (*r*) movements by the cooperation of four thrusters(T_1~T_4). As shown in Fig. 3, the thrusters are located at the midpoints of the four sides of the ASV. The thrusters are distributed in a "+" shape, and the arrows indicate the positive thrust direction.

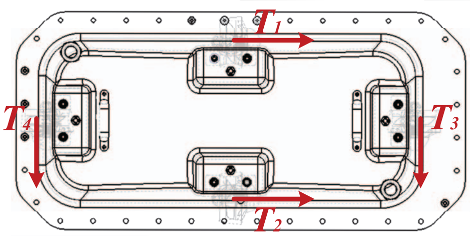


Fig. 3. Distribution structure of four thrusters

For the horizontal direction control of ASV, the force and torque generated by the thrusters can be expressed as:

$$\begin{bmatrix} \tau_u \\ \tau_v \\ \tau_r \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ L/2 & -L/2 & W/2 & -W/2 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} \tag{1}$$

Where, *L* is the length of the ASV, *W* is the width of the ASV, and f_1~f_4 are the thrust forces of the four thrusters.

B. *Establishment of ASV dynamic model*

According to the dynamic formula of a rigid body in fluid proposed by Fossen [9], the kinematics and dynamics model of ASV are established as:

$$\begin{gathered} \dot{\eta} = J(\upsilon)\upsilon \\ M\dot{\upsilon} + C(\upsilon) + D(\upsilon) + g(n) = \tau \end{gathered} \tag{2}$$

Where,

η is the position and attitude vector of ASV in the fixed coordinate system:

$$\eta = [x \quad y \quad \psi]^T \tag{3}$$

$J(\psi)$ is the transformation matrix from the motion coordinate system to the fixed coordinate system:

$$J(\psi) = \begin{bmatrix} cos\,\psi & -sin\psi & 0 \\ sin\psi & cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

υ is the linear velocity and angular velocity vector of ASV in the motion coordinate system:

$$\upsilon = [u \quad v \quad r]^T \tag{5}$$

M is the inertial matrix of ASV:

$$M = diag(M_{11}, M_{22}, M_{33}) \tag{6}$$

Assuming that the origin of the motion coordinate system coincides with the center of gravity of the ASV, $C(\upsilon)$ is the obliquely symmetric matrix of the Coriolis and centripetal terms of the ASV:

$$C(\upsilon) = \begin{bmatrix} 0 & 0 & -M_{22}v \\ 0 & 0 & M_{11}u \\ M_{22}v & M_{11}u & 0 \end{bmatrix} \tag{7}$$

$D(\upsilon)$ is the positive-semidefinite drag matrix-valued function:

$$D(\upsilon) = diag(D_{11} \quad D_{22} \quad D_{33}) \tag{8}$$

τ is the combined force of the thrusters and the flow force:

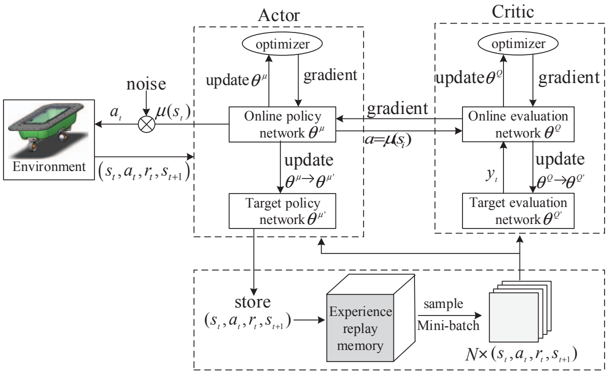
$$\begin{gathered} \tau = \tau_{pro} + \tau_{env} \\ \tau_{pro} = [\tau_u \quad \tau_v \quad \tau_r]^T \\ \tau_{env} = [\tau_x \quad \tau_y \quad 0]^T \\ \tau_x = 0.5 \times \rho_d \times C_d \times S_x \times |v_x| \times v_x \\ \tau_y = 0.5 \times \rho_d \times C_d \times S_y \times |v_y| \times v_y \end{gathered} \tag{9}$$

Where, τ_x, τ_y is the force produced by the water flow along the *X* and *Y* axes to ASV in the fixed coordinate system; ρ_d is the density of water; C_d is the resistance coefficient; S_x, S_y are the projections of ASV facing the *X* and *Y* directions respectively; v_x, v_y is the velocity of water flow in the *X* and *Y* directions.

III. Controller Design Based on DDPG Algorithm

The DDPG algorithm is a model-free, online reinforcement learning method [10]. The DDPG controller is an actor-critic reinforcement learning model, which selects the actions to perform based on the current state and calculates the best strategy to maximize long-term rewards. The observation space of the DDPG controller can be continuous or discrete space, and the output of the controller is continuous space.

The structure diagram of the system model based on DDPG controller is shown in Fig. 4.



Actor
Critic
optimizer
optimizer
update θ^{μ}
gradient
update θ^{Q}
gradient
noise
a_t
$\mu(s_t)$
Environment
(s_t, a_t, r_t, s_{t+1})
Online policy network θ^{μ}
gradient
$a=\mu(s_i)$
Online evaluation network θ^{Q}
update $\theta^{\mu}\rightarrow\theta^{\mu'}$
y_t
update $\theta^{Q}\rightarrow\theta^{Q'}$
Target policy network $\theta^{\mu'}$
Target evaluation network $\theta^{Q'}$
store (s_t, a_t, r_t, s_{t+1})
Experience replay memory
sample
Mini-batch
$N\times(s_t, a_t, r_t, s_{t+1})$

Fig. 4. Structure diagram of system model based on DDPG controlles

For reinforcement learning problems, rewards are used to evaluate the performance of current strategies [11]. Therefore, we define a reward function that allows the DDPG controller to evaluate state. The reward function is defined as shown in (10).

$$
\begin{aligned}
r_1 &= \begin{cases} 10 \;\; if\; e_d < 1 \;or\; e_\psi < \frac{\pi}{6} \\ 0 \;\;\; if\; e_d \geq 1 \;or\; e_\psi \geq \frac{\pi}{6} \end{cases} \\
r_2 &= \alpha e^{-e_d} + \beta e^{-e_\psi} \\
r_3 &= \chi(|f_1| + |f_2| + |f_3| + |f_4|) \\
r &= r_1 + r_2 + r_3
\end{aligned}
\tag{10}
$$

Where, α, β, χ is the weight coefficient of the reward function; $(x_d \quad y_d \quad \psi_d)^T$ is the target position and attitude of ASV; $(x \quad y \quad \psi)^T$ is the current position and attitude of ASV; $e_d = \sqrt{(x_d - x)^2 + (y_d - y)^2}$represents the distance from the target point; $e_\psi = \sqrt{(\psi_d - \psi)^2}$ represents the angle to the target attitude.

r_1 guides the ASV to near the target point with a large positive reward, and can accelerate the convergence of the DDPG network;

r_2 evaluates the distance to the target position and attitude. We define a reward in the form of an exponential function. α, β is positive, and the greater the e_d or e_ψ , the smaller the reward value obtained, conversely, the greater the reward value obtained. r_2 can guide ASV to approach the target position and attitude;

r_3 evaluates the energy consumption of the thrusters. By restricting the size of the thrust, it is conducive to stabilizing the change of thrust and can avoid excessive energy consumption in practical applications;

r is the total reward value obtained by taking each action.

IV. Simulation Experiment

In this section, we use the reinforcement learning toolbox in MATLAB to establish the DDPG control system of ASV and design the training environment. Based on the training controller, the fixed-point control and trajectory tracking experiments of ASV are carried out. At the same time, we also conduct a "reconfigurable" demonstration of multiple ASVs, verifying that the DDPG control strategy proposed in this paper is feasible and effective for ASV control.

A. Simulation conditions and training methods

The model parameters of ASV [5] are shown in TABLE I. The parameters of ASV's fluid damping formula are shown in TABLE II.

TABLE I. The model parameters of ASV

M_{11}	M_{22}	M_{33}	D_{11}	D_{22}	D_{33}	L/m	W/m
13.0	23.3	1.3	6.0	7.1	0.8	1.0	0.5

TABLE II. The parameters of ASV's fluid damping formula

ρ_d	C_d	S_x	S_y	v_x	v_y
$10^3 kg/m^3$	0.8	$0.06m^2$	$0.12\ m^2$	$0.3m/s$	$0.3m/s$

Computer configuration: CPU Inter Xeon(R) E5-2640, memory 64GB; MATLAB 2019b; Reinforcement Learning Toolbox.

The training environment of the DDPG controller is: the controlled object is ASV, the initial position is randomized in a circle with a radius of 10m, and the attitude angle range is [0,360°]. The goal of training is to drive ASV from the initial state to the origin, and the attitude angle is $\psi = 0°$. The schematic diagram of the training environment is shown in Fig. 5. The simulation parameters are: the maximum training period is 20,000; the sampling time in a training period is 0.4s, the simulation time is 40s, the reward function weight coefficient α is 1, β is 0.5, and χ is -0.1.

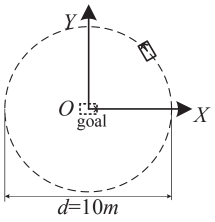


Fig. 5. Schematic diagram of DDPG controller training environment

After training, we can get the training process of the DDPG controller as shown in Fig. 6, the total training time is about 5 hours.

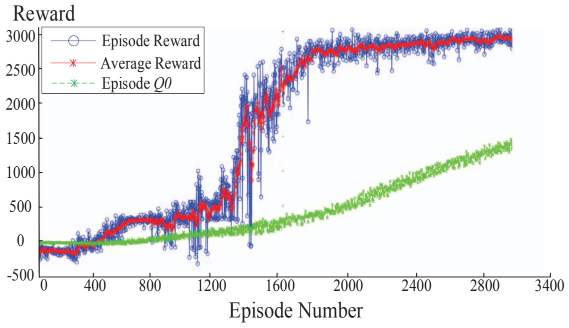
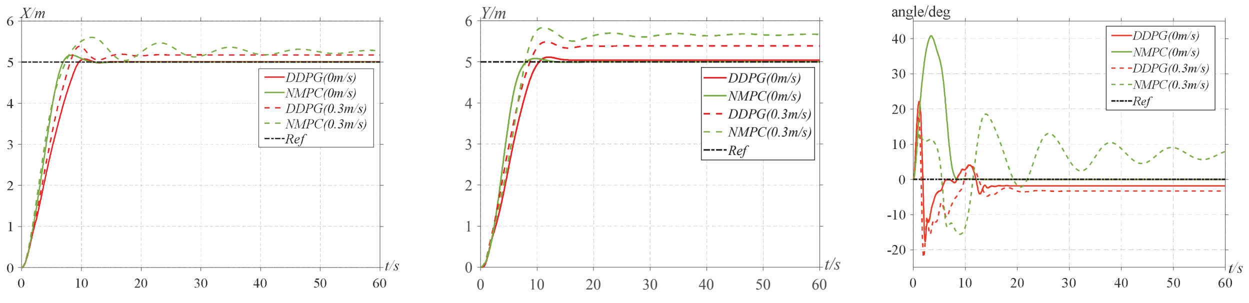


Fig. 6. Training process of DDPG controller



(a) Simulation results of *X*-axis (b) Simulation results of *Y*-axis (c) Simulation results of ψ

Fig. 7. Comparison of simulation results of fixed-point control under static water and 0.3*m/s* velocity disturbance

From Fig. 6, we can see that the blue line (episode reward) represents the reward value obtained in each training episode; the red line (average reward) represents the average reward value of 10 training episodes; the green line (episode Q0) represents the long-term reward value of the discount factor obtained after each training episode. In the 0~1000 training episodes, the DDPG controller starts to learn the training environment, but the results of the movement are poor, and the reward value is low. In the 1200~2000 training episodes, the ASV starts to adopt a better action strategy, and the reward value gradually increases. After 2000 training episodes, the reward value converges to a certain range, and the controller achieves the control goal, which indicates that the reward function design is reasonable.

B. *Fixed point control and trajectory tracking*

In this section, the control effect of DDPG controller is discussed under the static water environment and the disturbance of water flow, and the parameters of NMPC controller are adjusted to the optimal state for comparative analysis of the results. The initial position and attitude of ASV are (0,0,0), and the target position and attitude is (5,5,0). Each thrust range is limited to [- 5N, 5N]. The comparative analysis results on the three degrees of freedom of *X*, *Y* and ψ are shown in Fig. 7.

According to the analysis of Fig. 7, both DDPG and NMPC algorithms can accomplish the goal of fixed-point control. In an environment without water flow, both controllers can achieve no steady-state error on the *X* and *Y* axes. The DDPG controller has a steady-state error of 2° in attitude control because the weighting coefficient of the reward function $\alpha > \beta$, we pay more attention to reducing the steady-state error of the position. The NMPC controller is superior to the DDPG controller in the rise time, but the amount of overshoot is larger, especially the attitude control of ASV, and the DDPG controller makes the fluctuation smaller. In the presence of water flow interference, both controllers have steady-state errors in fixed-point control, but the accuracy and stability of DDPG controller are better than NMPC controller; When the flow velocity reaches 0.3*m/s*, the magnitude of the force generated by the water flow can be obtained from (10) τ_x is 4.32*N*, τ_y is 8.64*N*, because the maximum thrust force that ASV can generate in the $O\xi$ and $O\eta$ direction is 10*N*. Therefore, the ratio of the environmental interference force τ_x in the *X* axis direction to the maximum thrust of the thruster in the $O\xi$ direction is 43.2%, and the ratio of the environmental interference force τ_y in the *Y* axis direction to the maximum thrust in the $O\eta$ direction of the thruster is 86.4%, which is also the reason why the two controllers have steady-state errors in fixed-point control.

After training, the DDPG controller also has the ability of trajectory tracking. When the given target is a sinusoidal trajectory and water flow interference is added in the *X*-axis direction, the tracking result is shown in Fig. 8.

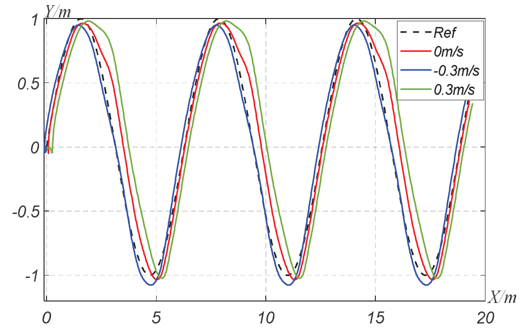


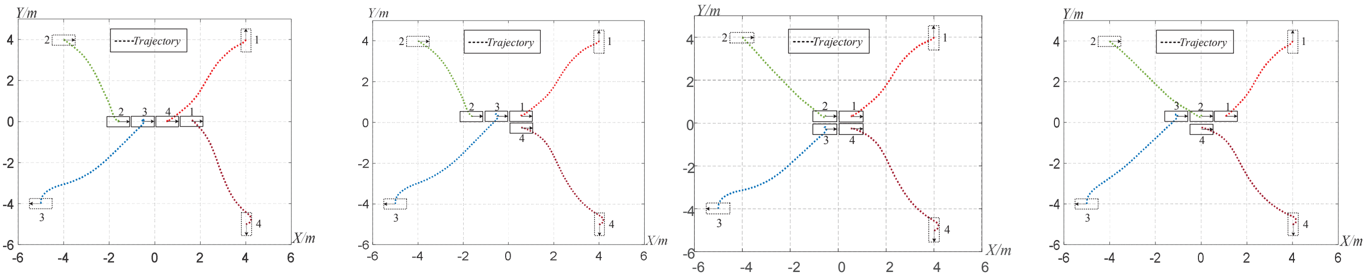
Fig. 8. Comparison of simulation results of trajectory tracking result under static water and ±0.3*m/s* velocity disturbance

According to the analysis of Fig. 8, when there is no water flow interference, the maximum tracking error is 0.55*m*, and the maximum error mainly occurs between the sinusoidal periods[$\pi/2 \quad 3\pi/2$]; When $x_d - x > 0$, the tracking effect is better, when $x_d - x < 0$, the tracking effect is poor, the reason for this phenomenon is that the environment training of $x_d - x < 0$ is insufficient during DDPG training, and the control effect can be improved after continuing training. There are ±0.3*m/s* water flow in the *X*-axis direction, which has a greater disturbance to the ASV, and the DDPG controller can also complete trajectory tracking. When the flow velocity is 0.3*m/s*, the maximum tracking error is 0.8*m*; when the flow velocity is -0.3*m/s*, the maximum tracking error is 0.1*m*, and the maximum error occurs between the sinusoidal periods [$\pi/2 \quad 3\pi/2$]. Therefore, sufficient training of the DDPG controller can improve control accuracy.

C. *The "reconfigurable" experiments of multiple ASVs,*

In this section, the "reconfigurable" simulation experiment of four ASVs is conducted. All ASVs use DDPG controller, which are combined into target shapes by designing the arrangement and movement sequence of ASVs. The experimental results are shown in Fig. 9(a) ~ (d). The numbers of the four ASVs are 1~4, and the initial position and attitude of ASVs are (4, 4, 90°) , (-4, 4, 90°) , (-5, -4, 180°) , (4, -5, -90°) respectively. The arrow indicates the orientation of ASV, and the interference of obstacles, water flow and other ASVs are not considered in the experiment.

As shown in Fig. 9, the target shapes in the experiment are "line" shape, "L" shape, "four sides" shape, and "T" shape. It can be obtained that the DDPG controller can independently



(a) "line" shape

(b) "L" shape

(c) "four sides" shape

(d) "T" shape

Fig. 9. The results of the *"*reconfigurable*"* experiments

plan the movement path, so that the ASV reaches the target position from different initial states to complete the "reconfigurable" experiment.

In the process of the ASV movement, water flow disturbance is one of the factors that affect its stability. To verify the ability of ASV to be "reconfigurable" at different flow rates, the experiment takes the "four sides" as the target shape, and sets the initial flow rate at 0 *m/s* and increase 0.1 *m/s* each time until the rate reaches 0.3 *m/s*; at the same flow rate, the direction of the water flow is initially set to 0°, increasing 45° each time until the direction of the water flow reaches 360°. A total of 36 sets of experiments are conducted. The results are shown in Fig. 10.

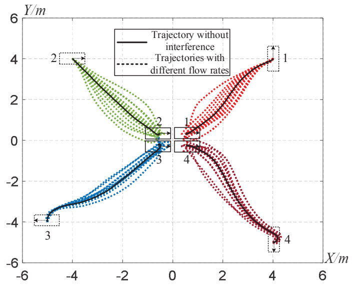


Fig. 10. The " reconfigurable " ability of multiple ASVs at different flow rates and directions

It can be seen from Fig. 10 that ASV has strong robustness under different flow rates and can reach the target position through different trajectories to complete "reconfigurable" experiment. However, because there is no communication between the ASVs, the DDPG controller does not add obstacles to the observation signal for obstacle avoidance learning in the experiment, so when approaching the target position, it may cause collisions between the ASVs. After obstacle avoidance learning, the DDPG controller can effectively solve this problem.

V. Conclusion

In this paper, a controller design method based on the DDPG algorithm is proposed and verified by simulation experiments on ASV. It can be seen that in the absence of water flow and water flow interference, this method can accurately perform fixed-point control and a good trajectory tracking effect, at the same time, the controller proposed in this paper can also make multiple ASVs have "reconfigurable" functions. Since the obstacle avoidance is not considered in the experiment, the " reconfigurable " ability still lacks autonomy. In the future, we will continue to optimize the DDPG controller. By adding static obstacles and dynamic obstacles to the environment, multiple ASVs can autonomously plan their paths to achieve the goal of "reconfiguration".

References

[1] S. Park, E. Kayacan, C. Ratti, and D. Rus, "Coordinated control of a reconfigurable multi-vessel platform: Robust control approach," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 4633-4639.

[2] Y. Lu, G. Zhang, L. Qiao, and W. Zhang, "Adaptive output-feedback formation control for underactuated surface vessels," International Journal of Control, vol. 93, no. 3, pp. 400-409, 2020.

[3] J. Woo, C. Yu, and N. Kim, "Deep reinforcement learning-based controller for path following of an unmanned surface vehicle," Ocean Engineering, vol. 183, pp. 155-166, 2019.

[4] J. Paulos, N. Eckenstein, T. Tosun, J. Seo, J. Davey, J. Greco et al., "Automated self-assembly of large maritime structures by a team of robotic boats," IEEE Transactions on Automation Science and Engineering, vol. 12, no. 3, pp. 958-968, 2015.

[5] W. Wang, L. A. Mateos, S. Park, P. Leoni, B. Gheneti, F. Duarte et al., "Design, modeling, and nonlinear model predictive tracking control of a novel autonomous surface vehicle," in 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 6189-6196.

[6] H. Sharma, T. Sebastian, and P. Balamuralidhar, "An efficient backtracking-based approach to turn-constrained path planning for aerial mobile robots," in 2017 European Conference on Mobile Robots (ECMR), 2017, pp. 1-8: IEEE.

[7] V. Mnih et al., "Playing atari with deep reinforcement learning," arXiv preprint arXiv:1312.5602, 2013.

[8] Y. F. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning," in 2017 IEEE international conference on robotics and automation (ICRA), 2017, pp. 285-292.

[9] T. I. Fossen, "Guidance and control of ocean vehicles," University of Trondheim, Norway, Printed by John Wiley & Sons, Chichester, England, ISBN: 0 471 94113 1, Doctors Thesis, 1999.

[10] Y. Ma, W. Zhu, M. G. Benton, and J. Romagnoli, "Continuous control of a polymerization system with deep reinforcement learning," Journal of Process Control, vol. 75, pp. 40-47, 2019.

[11] M. Zhu, Y. Wang, Z. Pu, J. Hu, X. Wang, and R. Ke, "Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving," Transportation Research Part C: Emerging Technologies, vol. 117, p. 102662, 2020.