A Comparative Study of Reinforcement Learning-based Collision Avoidance for Maritime Autonomous Surface Ships

Liangbin Zhao
Institute of High Performance Computing (IHPC)
*Agency for Science, Technology and Research (A*STAR)*
Singapore
zhao_liangbin@ihpc.a-star.edu.sg

Xingrui Yu
Institute of High Performance Computing (IHPC)
*Agency for Science, Technology and Research (A*STAR)*
Singapore
yu_xingrui@cfar.a-star.edu.sg

Xiuju Fu
Institute of High Performance Computing (IHPC)
*Agency for Science, Technology and Research (A*STAR)*
Singapore
fuxj@ihpc.a-star.edu.sg

***Abstract*— The efficacy of reinforcement learning has been substantiated in the development of intelligent modules for achieving autonomous collision avoidance in maritime autonomous surface ships (MASS). However, the performance of reinforcement learning algorithms with different configurations varies for making decisions. The evaluation and comparison of different configurations of reinforcement learning algorithms in this application remain challenging due to the absence of standardized or consensually adopted testing methodologies. In light of this, we proposed a simulation-based evaluation framework with three hierarchical metrics, namely, collision-free achievement, deviation angle for path-following, and avoidance time consumed, to enable the evaluation of reinforcement learning-based collision avoidance approaches. Comparative experimental analyses were conducted on six configurations of reinforcement learning algorithms with distinct reward designs across three typical vessel encounter scenarios in a simulated environment. Results indicated that by employing a potential-based design in intermediate rewards, specifically through the calculation of the course deviation, a notable enhancement in path-following can be achieved. Additionally, the weighted sum approach in the final reward design has also been demonstrated to effectively enhance the respective performance. The evaluation framework proposed, and our comparative experiments provided valuable insights and reference for evaluating collision avoidance algorithms in unmanned ship navigation and the reward designs employed in reinforcement learning within this specific application scenario.**

Keywords—Reinforcement Learning, Collision Avoidance, Simulated Testing, Maritime Autonomous Surface Ships

I. INTRODUCTION

The need for heightened efficiency and operational safety has driven the evolution of diverse levels of automation implemented in maritime sector. Given that human factors remain the predominant contributors to collision incidents in maritime traffic, the development of autonomous ship navigation systems, capable of reducing the probability of human errors and lowering labor costs, emerges as a typical technological focus within the industry. Integrated with other onboard sensors, autonomous ship navigation systems is designed to autonomously control or assist vessels in safely reaching their destinations in a manner that is relatively more accurate and efficient. Ships equipped with such systems that can operate, to varying degrees, independently of human interaction are referred to as Maritime Autonomous Surface Ships (MASS) by the International Maritime Organization (IMO)[1].

The navigation control of maritime vessels is a multi-objective complex decision-making process. Its automated solutions not only need to address the trade-offs between path tracking and collision avoidance but also consider the intricate kinematic characteristics of ships. Differing from other conventional methods that require extensive calculations involving ship dynamics models, Reinforcement Learning (RL) emerges as an effective alternative for autonomous control[2]. The RL agent directly acquires the end-to-end connection between observations and actions through the principle of trial and error. In recent years, various scholars have proposed different RL-based methods to demonstrate its effectiveness in autonomous navigation for maritime vessels. For example, the Proximal Policy Optimization (PPO) RL algorithms[3] has been adopted by Meyer[4] and Zhao[5] in their solution for vessel collision avoidance. The implementation was successful in their respective simulation environments. Additionally, Deep Q-Network (DQN) has been employed by Woo[6] to design the collision avoidance decision module for unmanned surface vehicle, and they have successfully conducted real ship experiments involving multiple encounters.

However, the performance of reinforcement learning algorithms with different configurations varies and the evaluation and comparison of reinforcement learning algorithms in autonomous ship navigation remain challenging due to the absence of standardized or consensually adopted testing methodologies[7]. Larsen [8] conducted a detailed comparison of various difficulty levels in simulated environments to evaluate the performance of different RL algorithms in autonomous navigation. The conclusion drawn was that the PPO algorithm demonstrated superior robustness to changes in the complexity of the environment. Their comprehensive study focuses on the performance of unmanned vessels in tracking known paths. The evaluation perspective is more inclined towards the goal of overall navigation. The main evaluation metrics adopted include average progress for path-following, cross-track error-based path adherence, and others. For the performance in short-term collision avoidance scenarios encountered during the process, there is a lack of more fine-grained evaluation. To address this, we proposed a simulation-based evaluation framework to further assess and compare the performance of different reinforcement learning configurations in the autonomous

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00012

navigation of maritime vessels, particularly regarding collision avoidance capabilities. We conducted comparative experiments as part of our contribution, providing valuable insights and references for evaluating collision avoidance algorithms in autonomous navigation and understanding the reward designs employed in reinforcement learning within this specific context.

II. Comparative Evaluation Experiment

A. Overview of the evaluation framework for autonomous collision avoidance

Collision avoidance, as a crucial stage in the autonomous navigation process of unmanned ships, is defined as the process wherein a vessel safely restores itself to the navigation state along the predefined route after encountering collision risks within a certain timeframe. The collision avoidance algorithms still have the dual primary goals of collision avoidance and path tracking recovery, alongside the secondary goal of optimizing strategies during the avoidance process. Therefore, recognizing the distinct priorities among them, this study proposes an evaluation framework based on three hierarchical metrics through simulation, namely, collision-free achievement, deviation angle for path-following, and avoidance time consumed, as shown in Fig. 1. It is worth noting that the evaluation of avoidance strategies could be multifaceted, considering factors such as adherence to the collision regulations (COLREGs) or good seamanship practices. we only utilizes avoidance time consumed as a demonstration for quantitative assessment in efficiency.

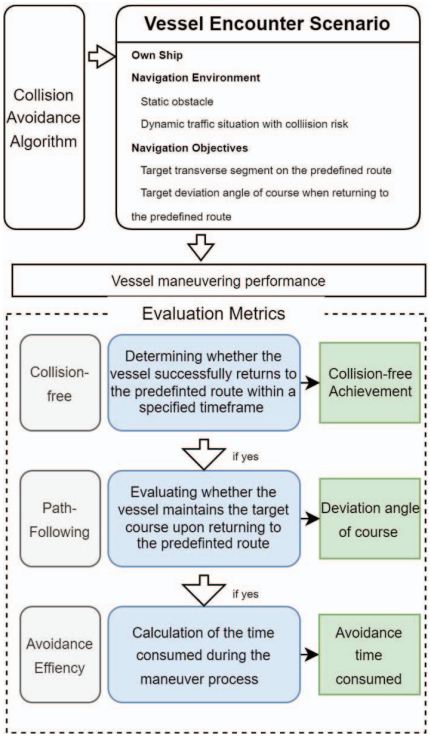

Fig. 1. Evaluation framework.

Our evaluation framework is established based on simulated ship encounter scenarios. The own ship under test in this scenario have two primary objectives. The first one to safely return to the predefined route without collision within a specified timeframe. The representation of the destination is a specific transverse segment on the predefined route. The length of the transverse segment depends on the target threshold setting for the deviation distance of route. The second objective is to maintain alignment with the predefined direction as closely as possible when returning to the route, ie. minimizing the course deviation angle. Two metrics, named Collison-free achievement and deviation angle of course, are utilized to evaluate the performance of these two primary objectives.

Besides, a designated range for the deviation angle threshold can be set to qualitatively evaluate whether the objective of path-following is achieved. When the collision avoidance algorithm guides the own ship to achieve the two aforementioned objectives in a test scenario, the time consumed during the avoidance process is considered as an evaluation metric for the avoidance efficiency.

B. Basic Configuation of Evaluation Experiment

The basic configurations for constructing the evaluation experiment are as follows.

- **Own Ship.** In our simulation environment, a three-degree-of-freedom (3-DOF) mathematical model is used to simulate the ship maneuvering motion of the own ship according to the selected action. The parameters utilized are derived from a validated 330-meter oil tanker.
- **Navigation Environment.** The test scenarios include three kinds of typical encounters between two vessels, namely overtaking, head-on and crossing, based on the angle between the course of ship, as shown in Fig. 2.
- **Navigation Objectives.** In all test scenarios, the own vessels start navigating along a straight predefined route, and the objectives is to safely return to the original route within 10 minutes and a distance of 1.75 kilometer traveled on the route. The decision interval is 10 seconds. The threshold for determining the achievement of arrival and path-following is set at 200 meters (deviation distance) and 15° (deviation angle), as shown in Fig. 2.

C. RL-based ship collision avoidance

In this study, the decision-making unit RL algorithm, with its different configurations, is the subject we need to compare and evaluate. Although various types of reinforcement learning algorithms, such as DQN, PPO, TD3, and SAC, have been proposed for ship avoidance decision-making, PPO is the most widely adopted algorithm. A comparative study [8] also demonstrated its advantages in terms of implementation robustness and decision performance. The deployment experience of RL in our simulation environment also demonstrated the advantages of PPO. Compared to other algorithms, PPO can achieve relatively stable and successful convergence in various scenarios. Therefore, In our comparative experiments, we selected PPO as the RL algorithm, placing particular emphasis on performances with different reward designs.

In addition, regarding the design of the state vector for short-term encounter situation between two vessels, the indicators employed in this study include basic dynamic information of the own ship and other ships in the environment (such as coordinates, speed, direction), the

own ship's relative relationship to the predefined route and destination (such as cross-track error, deviation angle error), and the own ship's relative relationship to other ships in the environment (such as relative bearing, relative position, relative speed). Besides, in this study, the action space is simplified into three strategies of rudder acceleration, namely, steering to port, steering to starboard, and not steering.

D. Reward Design of RL for comparison

The reward signal is the driving force behind RL algorithms, directly influencing the learning outcomes and behavior of the intelligent agent. In the context of ship collision avoidance, this paper summarizes six fundamental scoring methods for achieving goals and combines them into six different reward designs through weighted sum for comparative evaluation, as shown in the tables below.

TABLE I. SCORING METHOD FOR GOALS

Stage	*Goals*	*Scoring method* [e]
End of the navigation	Collision-free arrival (G_{cf})	$G_{cf} = \begin{cases} w_{cf} & \text{if arrival} \\ -w_{cf} & \text{if collision} \end{cases}$
	Optimal Course when arrival (G_{pf})	$G_{pf} = \begin{cases} w_{pf} & \text{if path-following is achieved} \\ 0 & \text{otherwise} \end{cases}$
	Arrival effiency (G_{ef})	[a] $G_{ef} = w_{ef}(1 - \frac{T_{\text{cost}}}{T_{\max}})$
Process of navigation	To be near the route (G_{xte})	[b] $G_{xte}^{(t)} = w_{\text{intermediate}}(1 - \frac{XTE^{(t)}}{XTE_{\max}})$
	To follow the heading of the route (G_{cd})	[c] $G_{cd}^{(t)} = w_{\text{intermediate}}(\frac{1}{2} + \frac{\cos(\theta^{(t)})}{2})$
	Avoid collision (G_{cr})	[d] $G_{cr}^{(t)} = \begin{cases} -w_{\text{cr}} & \text{if collision risk exits} \\ 0 & \text{otherwise} \end{cases}$

[a] T_{cost} is the time consumed. $T_{\max}$ is the threshold of maximum time.

[b] XTE is the cross-track error to route. $XTE_{\max}$ is the threshold of maximum cross-track error.

[c] θ is the course deviation angle to route.

[d] collision risk exits when DCPA (distance at closest point of approach) index less than 200 meters and relative distance less than 900 meters.

[e] W_{cf} =10, W_{pf} =5, W_{ef} =5, $W_{\text{intermediate}}$=0.125

TABLE II. REWARD DESIGNS

Type	*Design*	*Scoring method*					
		G_{cf}	G_{pf}	G_{ef}	$G_{xte}^{(t)}$	$G_{cd}^{(t)}$	$G_{cr}^{(t)}$
Final Reward	Fr-Cf	√					
	Fr-CfPf	√	√				
	Fr-CfPfEf	√	√	√			
Final Reward + Intermediate Reward	Fr-Cf_Ir-Xte	√			√		√
	Fr-Cf_Ir-XteCd	√			√	√	√
	Fr-CfEf_Ir-XteCd	√		√	√	√	√

III. Results and Discussion

Using all the setups mentioned above, we train RL agents based on Gymnasium python library in three simulated scenarios. It is worth mentioning that the model training is conducted using unseen simulated scenarios of the corresponding encountering types. The test scenarios are demonstrated in Fig 2. And we calculate averages of metrics from 100 tests conducted after 10,000 episodes to evaluate performance, as shown in tables and figures below. The trajectory in blue was made by own ship. And the red trajectory represents the path of the other ship, which would result in a collision if own ship doesn't make an avoidance maneuver.

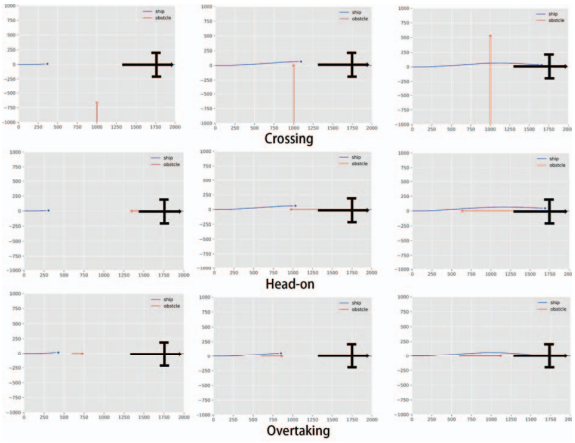

Fig. 2. Example of RL-based ship avoidance (black arrow and line segment is the navigation objectives)

Firstly, in our experiments, all the RL algorithms consistently achieve the fundamental goal of avoiding collisions and reaching to the target destination with a success rate exceeding 90%. These results clearly demonstrate the effectiveness of directly setting the target score in the final reward, i.e., G_{cf}, in this testing environment. Secondly from Tab. III and Fig. 3, it can be observed that the two below reward designs have proven to be effective in enhancing the performance of path-following.

- Directly incorporating the score of deviation angle at arrival into the final reward, i.e. G_{pf}.
- Adding the scores for course deviation during the process in intermediate reward, i.e. G_{cd}.

Besides, from Tab. IV and Fig. 4, it is can be observed that incorporating reward scores for minimizing the total time into the final reward effectively enhances the performance in avoidance efficiency, i.e. G_{ef}.

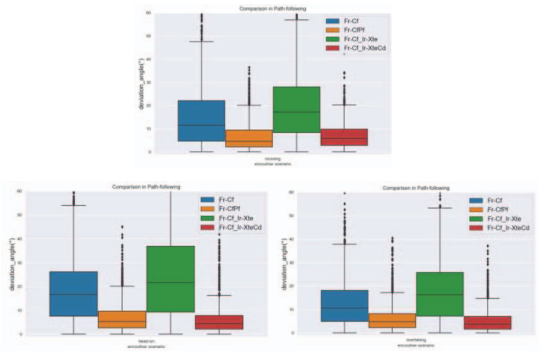

Fig. 3. Comparision of avoidance performance in Path-following

TABLE III. PERFORMANCE RESULT IN PATH-FOLLOWING

Scenarios	Reward design	Average achieving rate of collision-free	Average achieving rate of path-following (w.r.t. 15°)	Average time consumed (second)
Crossing	Fr-Cf	0.89	0.90	351.35
	Fr-CfPfEf	0.96	0.93	**350.10**
	Fr-Cf_Ir-Xte	0.92	0.91	354.60
	Fr-CfEf_Ir-XteCd	0.93	0.94	**351.05**
Head-on	Fr-Cf	0.90	0.90	352.96
	Fr-CfPfEf	0.93	0.93	**352.39**
	Fr-Cf_Ir-Xte	0.96	0.93	351.52
	Fr-CfEf_Ir-XteCd	0.94	0.94	**349.17**
Overtaking	Fr-Cf	0.95	0.93	349.77
	Fr-CfPfEf	0.94	0.94	**349.74**
	Fr-Cf_Ir-Xte	0.98	0.96	354.25
	Fr-CfEf_Ir-XteCd	0.93	0.94	**350.31**

TABLE IV. PERFORMANCE RESULT IN EFFCIENCY

Scenarios	Reward design	Average achieving rate of collision-free	Average deviation angle (°)	Average achieving rate of path-following (w.r.t. 15°)
Crossing	Fr-Cf	0.88	14.96	0.59
	Fr-CfPf	0.89	**6.59**	**0.90**
	Fr-Cf_Ir-Xte	0.92	19.17	0.44
	Fr-Cf_Ir-XteCd	0.92	**6.96**	**0.91**
Head-on	Fr-Cf	0.91	18.46	0.46
	Fr-CfPf	0.90	**6.96**	**0.90**
	Fr-Cf_Ir-Xte	0.87	24.12	0.36
	Fr-Cf_Ir-XteCd	0.96	**6.17**	**0.93**
Overtaking	Fr-Cf	0.95	12.72	0.66
	Fr-CfPf	0.95	**6.24**	**0.93**
	Fr-Cf_Ir-Xte	0.94	17.77	0.47
	Fr-Cf_Ir-XteCd	0.98	**5.14**	**0.96**

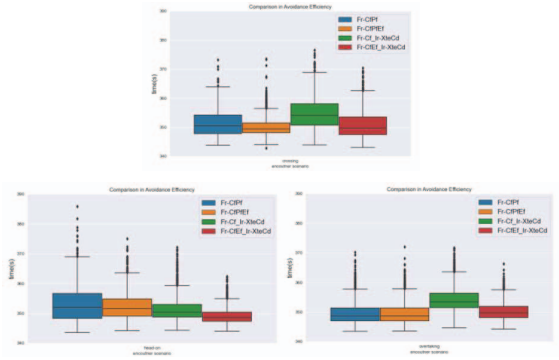

Fig. 4. Comparision of avoidance performance in efficiency

IV. CONCLUSION

Based on the evaluation framework we proposed, experiments comparing RL agent with different reward designs revealed that, by employing a potential-based design in intermediate rewards, particularly through the score of course deviation, a significant improvement in path-following performance can be achieved. Additionally, the weighted sum approach in the final reward design has also been demonstrated to effectively enhance the respective performance.

Considering more complex and diverse scenarios, such as encounters involving multiple vessels, in the testing framework will be one of the key aspects for further evaluating the generalization ability and robustness of RL-based collision avoidance algorithms in the future. Additionally, exploring how to train more effective RL agent and comparing their strengths and weaknesses with traditional collision avoidance algorithms will be the subsequent focus of this research.

ACKNOWLEDGMENT

This research was funded by Programme MARS (Programme of Maritime AI Research in Singapore) with funding grant number SMI-2022-MTP-06 by Singapore Maritime Institute (SMI).

REFERENCES

[1] https://www.imo.org/en/MediaCentre/HotTopics/Pages/Autonomous-shipping.aspx

[2] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[3] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[4] Meyer E, Robinson H, Rasheed A, et al. Taming an autonomous surface vehicle for path following and collision avoidance using deep reinforcement learning. IEEE Access, 2020, 8: 41466-41481.

[5] Zhao L, Roh M I. COLREGs-compliant multiship collision avoidance based on deep reinforcement learning. Ocean Engineering, 2019, 191: 106436.

[6] Woo J, Kim N. Collision avoidance for an unmanned surface vehicle using deep reinforcement learning. Ocean Engineering, 2020, 199: 107001.

[7] Burmeister H C, Constapel M. Autonomous collision avoidance at sea: A survey. Frontiers in Robotics and AI, 2021, 8: 739013.

[8] Larsen T N, Teigen H Ø, Laache T, et al. Comparing deep reinforcement learning algorithms' ability to safely navigate challenging waters. Frontiers in Robotics and AI, 2021, 8: 738113