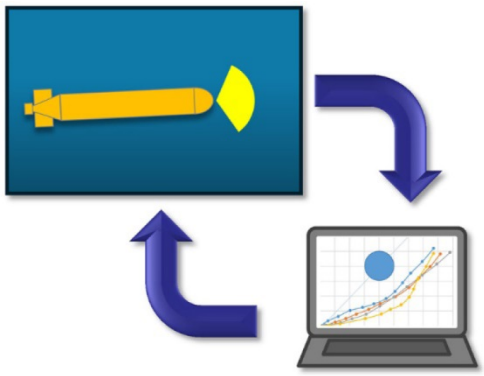
Advanced Technology of Deep Reinforcement Learning for Autonomous Vehicle



YUSUKE YASHIRO*1 KAZUKI EGUCHI*2
YOSUKE NAKAGAWA*1

Deep reinforcement learning conducts learning while automatically collecting data through iterative trials. It has attracted attention in recent years as a method for seeking optimal action policy and its applications have been increasing. However, in order to apply deep reinforcement learning to autonomous vehicle products that move under automatic control while satisfying multiple objectives, such as moving to a target position, avoiding obstacles, etc., there is the problem that the learning conditions must be appropriately adjusted before starting the learning. Digital innovation headquarter has taken on this challenge as a solution for our mission of smarter product and devised an advanced method to realize the easier application of deep reinforcement learning to autonomous vehicles and the more efficient improvement of the product performance, such as shortening of obstacle avoidance maneuvers. The effectiveness of avoiding obstacles efficiently under multiple obstacle conditions is confirmed by simulation and this report also presents the verification results.

1. Introduction

Autonomous vehicle products have been spreading throughout society, whether it's on the land, at sea, or in the air, due to their possibility of satisfying recent manpower-saving needs for mobility products[(1)]. To ensure that an autonomous vehicle fulfills its role, it is necessary to adjust various parameters within the onboard automatic control system in advance to satisfy multiple control objectives. In recent years, the tasks required for products have become more challenging and the number of parameters that their control system needs to consider in advance has tended to increase due to the requirements, such as that the product needs to operate safely even in unexpected situations. These adjustments are obstacles to achieving manpower savings in vehicle operation.

On the other hand, data-driven technologies such as deep learning have begun to become popular with the background of the progress of information technology, and their application to the automatic adjustment of control systems has been being considered. In particular, deep reinforcement learning, which automatically acquires data and learns optimal measures, is expected and studied as a method of automatic construction and tuning adjusting of control systems. However, this method has many items that need to be adjusted, such as parameters related to the internal deep neural network and reward functions that determine what is optimal in learning. Therefore, there is a risk that simply applying this method will only replace the work needed for adjusting the parameters in the control system with the work needed for adjusting the parameters used for learning.

This report introduces an advanced method to facilitate the application of deep reinforcement learning that we devised to solve this problem. First, it is necessary to estimate the reward function, which is difficult to adjust, so that it satisfies the optimality required by the designer. To do this, adversarial inverse reinforcement learning[(2)] was used. It is a method to derive the reward function

*1 CIS Department, Digital Innovation Headquarters, Mitsubishi Heavy Industries, Ltd.
*2 Chief Staff Manager, CIS Department, Digital Innovation Headquarters, Mitsubishi Heavy Industries, Ltd.

and also learning result at the same time using expert data as training data. Meanwhile, by adding a selector that selects and switches multiple learning results to be used to the control system, the work needed for adjusting deep reinforcement learning was reduced and it was made possible to select appropriate learning results even in an environment that was not expected at the time of the learning, in order to establish efficient control. We verified the effectiveness of the devised method by applying it to the simulation of obstacle avoidance maneuvers by autonomous underwater vehicles (AUV), which have been used in a wider range of fields, including natural science research, security, etc. in this report. This technology contributes to the realization of advanced and safe mobilities in the information society by widely deploying to autonomous vehicle products in our group.

2. Explanation of proposed method

Figure 1 shows the flow of learning and application of the learning results to a control system. First, expert data to be used as reference behaviors in adversarial inverse reinforcement learning is prepared. The expert data may be artificially created based on the behaviors required for the vehicle product. By adversarial inverse reinforcement learning, the reward function is estimated and automatically derived together with the learning result so that the learning result becomes closer to the expert data, which eliminates the work needed for adjusting the reward function. For example, in the case where a vehicle moves toward the target position while avoiding obstacles, attempting to shorten the time to reach the target position generates a trade-off of increasing the risk of passing nearby obstacles, which may cause a collision. However, by using expert data and adversarial inverse reinforcement learning, the balance between the time and avoidance implicitly sought by the designer can be estimated and reflected in the learning.

Deep reinforcement learning proceeds with learning so as to obtain higher reward values, but the finally derived learning result (method) depends on the random number conditions used. It is also not always suitable for product situations different from those at the time of the learning. Therefore, by using multiple learning results simultaneously for the control including results where the reward value is not the highest, the robustness of the control is improved. In this research, a simple method called a selector was used for selecting learning results, rather than a data-driven method such as ensemble learning with which multiple learning results are derived while learning of proper use of the different results. The selector executes a ranking function predetermined by the user for each control cycle and compares each learning result numerically. The learning result deemed optimal at that point is used and the selection will be redone in the next control cycle. Due to this procedure, continued use of learning results that may be inappropriate depending on conditions such as the positional relationship between the vehicle and obstacles is avoided.

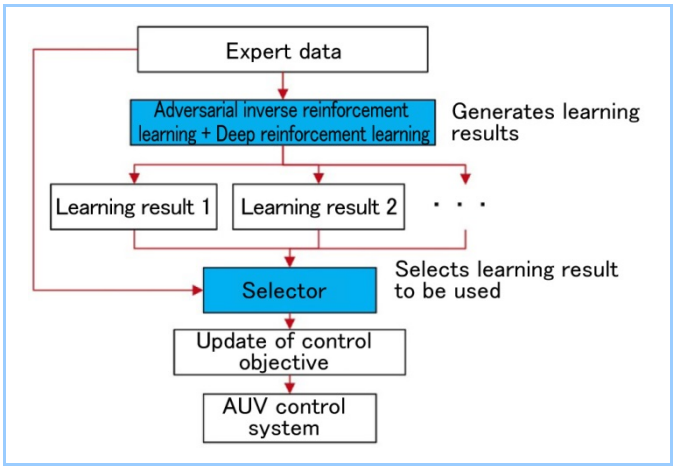


Figure 1 Flow of proposed method

3. Obstacle avoidance control system

Figure 2 shows a block diagram of obstacle avoidance control for an AUV using multiple learning results and a selector. Each learning is performed in a normalized space and the result is stored as a position data array. When the vehicle detects an obstacle, the data array is scaled

according to the distance to the obstacle and then used. The AUV normally moves toward the target position, but when triggered by detecting an obstacle, the initial target position is temporarily overwritten by the position data array. The learning result is used as the control objective, so the existing control system can be used as it is. The data array is used starting with the first position and switched in every control cycle according to the comparison based on the selector's ranking function. For an example of the ranking function, in this research, the distance between a position in each position data array and the initial target position was used to select a position closer to the target position while avoiding obstacles. When the vehicle reaches the selected position, the selector selects the next position, and these procedures are repeated. When the end of the data array is reached, the avoidance is considered to be finished, and again the vehicle moves toward the initial (before-overwriting) target position.

As the number of data that can be used at the same time in the selector increases, it becomes easier to select an appropriate option depending on the situation, but it takes more time to learn to prepare the data, so the example in this report uses four data. In this regard, future development of adversarial inverse reinforcement learning and deep reinforcement learning technology are required to reduce learning time.

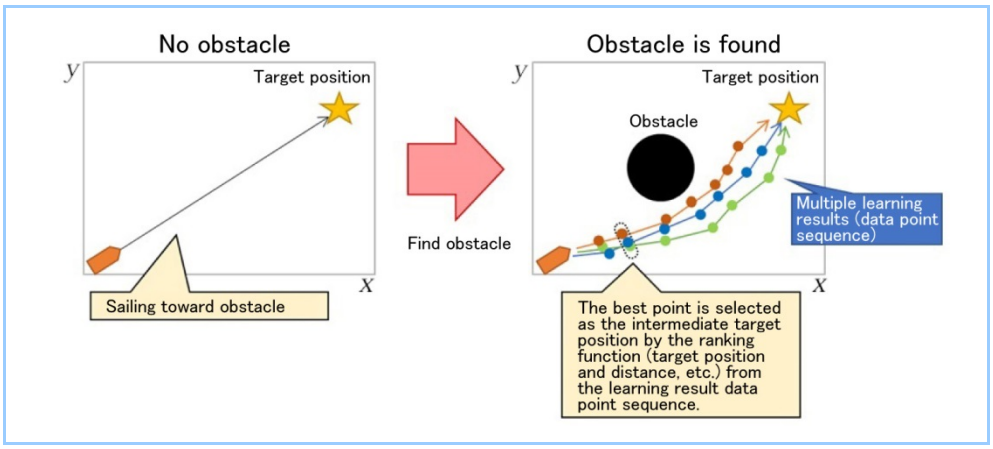


Figure 2 Obstacle avoidance control using proposed method

4. Verification with AUV simulation

Figure 3 shows the setup of the AUV simulation. The AUV used for this verification goes and turns using the stern thruster and rudders, acquires its own position using the inertial navigation system (INS) model, and detects the surface of an obstacle using only one forward looking sonar (FLS) model to obtain the distance between the AUV and the obstacle surface.

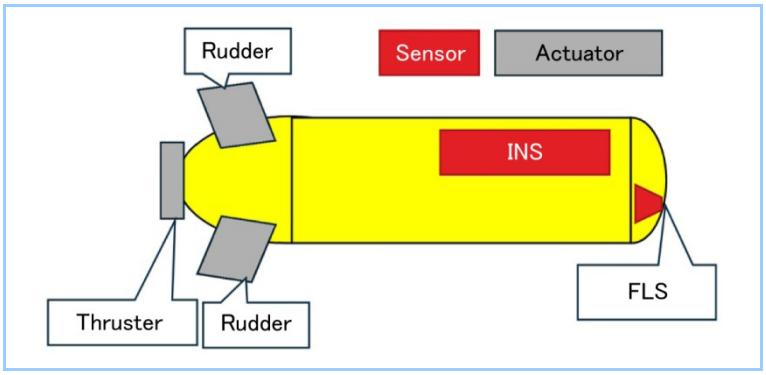


Figure 3 AUV model for verification with simulator

Figure 4 shows the simulation result. As a comparison counterpart to the proposed method, a method of adding a temporary target position a certain distance to the side of an obstacle if detected was used. For the proposed method, a total of four sets of data were used: three avoidance trajectory data sets resulted from the learning and one expert data set for adversarial inverse reinforcement learning.

First, in the base case (a) where the total route length and the obstacle size were the same as those at the time of the learning, the additional distance for the temporary target position was adjusted so that the comparison counterpart method was able to reach the target position somewhat earlier than the proposed method. Then, the two methods were compared in the evaluation case (b). The evaluation case had a larger obstacle, so the environment for the AUV was different from that at the time of the learning.

In the evaluation case, due to the difference in the obstacle size, the method of adding a temporary target position at a certain distance to the side of an obstacle was not able to make appropriate avoidance with the avoiding distance suitable for the base case, and repeated meandering composed of avoiding, finishing avoidance, detecting the obstacle again, and avoiding, resulting in the required time of 704 seconds. On the other hand, the proposed method generated and passed through the avoidance route smoothly without meandering, resulting in a faster arrival at the target position in 500 seconds. This is because, during the move, the selector selected a data array that brought the vehicle closest to the target position among the position data array given as the avoidance trajectory and thereby unnecessary turns were reduced. The same superiority of the proposed method was confirmed in evaluation cases with the conditions different from those for case (b), which indicates that the proposed method establishes a robust avoidance method that prevents the avoidance behavior from deteriorating significantly for different obstacles.

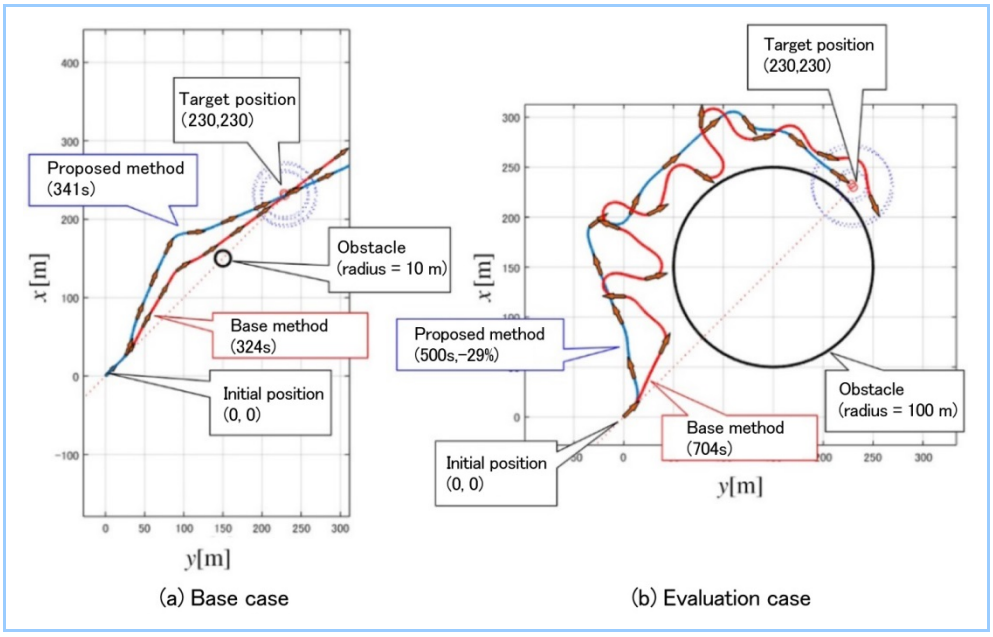


Figure 4 Result of verification with simulation

5. Conclusion

We devised an advanced technology that improves the applicability of deep reinforcement learning by eliminating the work needed for adjustment with the aim of applying it to autonomous vehicle products with multiple control objectives, and verified its effectiveness through simulations. With this method, the reward function, which is difficult to adjust, is automatically estimated by adversarial inverse reinforcement learning, and learning results that are close to the desired expert data are obtained. At the same time, the selector switches the learning results to be used and makes it possible to obtain appropriate control even in environments with obstacles that are different from those at the time of the learning. As a result, it is expected that the operation, including learning and control, of autonomous vehicle products, which are required to be deployed even in unknown environments, can be safely proceeded with and that the products contribute to the realization of more advanced tasks.

One of the remaining important issues is to consider in advance how many variations of learning results should be prepared, together with the operational environment of the product, in order to avoid the need for relearning after the product is put into operation. These variations are affected by learning time, however since reinforcement learning algorithms continue to develop

day by day, it is necessary to follow the latest trends in learning algorithms to see if there is a method that can learn faster. We will continue to utilize the Digital Innovation Headquarters system to centrally promote information collection and product technology development in order to satisfy the expectations of customers who use vehicle products.

This method was developed in collaboration with Chiba University. We would like to express our gratitude to Professor Sachiyo Arai, Associate Professor Tadanao Zanma, and their students for their cooperation.

References

(1) Kazuki Eguchi et al., Low-speed, Low-altitude AUV Control, Mitsubishi Heavy Industries Technical Review Vol.58 No.1 (2021)
(2) J. Fu, et al, Learning Robust Rewards with Adversarial Inverse Reinforcement Learning, ICLR 2018 (2018)
(3) Yashiro. Y, et al, Development of Applicable Reinforcement Learning Compensator Using Ranking Information for AUV, IEEE OCEANS 2022 (2022)