https://doi.org/10.7840/kics.2025.50.5.712

동적 환경에서의 강화학습기반 자율 수중 차량 변침점 생성 알고리즘에 관한 연구

노 지 민◆, 이 현 수*, 송 일 석**, 김 승 환**, 김 영 대**, 박 수 현***, 김 중 헌°

Reinforcement Learning-Based Autonomous Underwater Vehicle Waypoint Generation Algorithm in Dynamic Environments

Emily Jimin Roh◆, Hyunsoo Lee*, Ilseok Song**,
Seunghwan Kim**, Youngdae Kim**, Soohyun Park***, Joongheon Kim°

요 약

해양 과학 및 군사 작전에 대한 수요가 증가함에 따라 자율 수중 차량(autonomous underwater vehicle, AUV)의 중요성이 점차 커지고 있다. AUV는 유연한 각도 변화를 바탕으로 다양한 임무에서 성공률 높은 제어를 달성할 수 있다. 그러나 대부분의 기존 연구는 주로 단순한 환경에서의 실험과 AUV 작전에서의 고유한 방향성을 간과한다. 이에 본 논문은 AUV의 방향 정책 강화학습 알고리즘(directional policy reinforcement learning, DPRL)을 통한 변침점 생성 알고리즘 기반의 AUV 임무 수행 전략을 제안한다. 이때 장애물의 크기 및 위치, 목표물의 위치, 그리고 목표물에 대한 접근 각도를 나타내는 충격 각도를 무작위로 설정한 동적인 환경에서의 실험 결과를 통하여 본 알고리즘의 우수성을 입증한다.

키워드 : 자율 수중 차량, 강화학습, 변침점, 충격 각도
Key Words : Autonomous underwater vehicle(AUV), Reinforcement learning, Waypoint, Impact Angle

ABSTRACT

This paper proposes a method to optimize the autonomous torpedo maneuver path for reaching the target of torpedoes, which are explosive projectile weapons in naval operations. For flexible maneuvering of torpedoes, movement in various directions is considered. Also, the obstacles in the actual marine environment and the minimization of the waypoint that occurs when the angle of the torpedoes is changed considered to increase the efficiency of torpedo maneuvering. Consequently, this study presents the environment that reflects the action of the torpedo in various directions according to the maximum rotation angle. Torpedo maneuver strategy is formulated by applying a Markov Decision Process based reinforcement learning algorithm, Q-Learning. Compared to the general Q-Learning algorithm, the superiority of the proposed algorithm is assessed and its applicability in the actual marine environment, through the success rate of reaching the target point and the number of waypoints.

※ 이 논문은 2022년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(No. KRIT-CT-22-023-02, 잠수함용 지능형 임무지원시스템 통합자동화 기술 (잠수함 표적식별 및 교전지원 지능화 기술), 100%).
◆ First Author : Korea University Electrical and Computer Engineering, emilyjroh@korea.ac.kr, 학생회원
° Corresponding Author : Korea University Department of Electrical and Computer Engineering, joongheon@korea.ac.kr, 종신회원
* Korea University Department of Electrical and Computer Engineering, hyunsoo@korea.ac.kr, 학생회원
** LIGNex1, {ilseok.song, seunghwan.kim01, youngdae.kim}@lignex1.com
*** Sookmyung Women's University Division of Computer Science, soohyun.park@sookmyung.ac.kr, 정회원
논문번호 : 202411-300-A-RU, Received November 25, 2024; Revised December 26, 2024; Accepted January 2, 2025

Ⅰ. 서 론

해양 과학의 발전과 지능형 시스템의 발전에 따라, 자율 수중 차량(autonomous underwater vehicle, AUV)의 중요성이 점차 커지고 있다. 기존의 수중 차량과 달리, AUV는 외부의 인간 조작이나 유도 신호 없이 목표를 탐지하고 자율적으로 항해할 수 있는 핵심 시스템이다[1]. 이러한 특징을 바탕으로 AUV는 동적인 환경에서 인적 위험을 줄이고 안정적인 군사 작전, 수중 지형 탐사 및 해양 지역 모니터링과 같은 다양한 응용분야에서 활용된다[2]. 이에 따라 최근에는 특정 임무 수행을 위한 AUV의 제어 전략 수립 연구가 더욱 활발히 이루어지고 있다. AUV는 주로 다양한 해저 지형에서의 운용되는 시스템으로 무엇보다 동적인 환경에서도 안정적인 기동과 실시간으로 대처 가능한 제어 전략 수립이 필수적이다. 따라서 AUV의 제어 목표는 사전 데이터가 주어지지 않은 동적인 환경에서 빠른 시간 안에 다양한 위치를 갖는 목표물까지의 도달이다. 이때 AUV는 다양한 각도로의 제어가 가능하므로 특정 목표물이 주어지는 임무 수행에 있어 작전의 성공률을 향상시킬 수 있다[3]. 그러나 AUV의 방향성을 고려한 동적 환경에서의 적용을 위한 연구는 AUV의 행동 차원이 넓어짐에 따라 최적의 효율적인 경로로의 수렴이 어려우므로 지속적인 연구가 필요하다.

AUV의 경로 계획을 위한 많은 연구는 주로 2단계 방식으로 진행된다[4]. 첫 번째 단계에서는 수평 2차원 경로를 계산하여 목표 지점까지의 주요 경로를 결정하며, 두 번째 단계에서는 수집된 환경 데이터를 활용하여 수직 방향(깊이)을 조정함으로써 장애물 회피와 에너지 효율성을 극대화한다. 이에 본 연구에서는 이러한 2단계 방식 중 첫 번째 단계에 초점을 맞추어, 수평 2차원 경로에서 최적의 경로 도출에 중점을 두었다. 또한, 일정 시간에 AUV 제어가 가능한 최대회전각도를 두어 유연성을 높여, 최종적으로 다양한 방향으로의 유연한 대처가 가능한 AUV의 목표물까지의 임무 수행 전략을 수립하고자 한다. 이때 본 논문에서는 고정된 데이터셋에 의존적이지 않고 불확실한 환경에서 수집된 경험을 통해 학습하는 Markov decision process (MDP)[5]기반 강화학습을 적용하여 AUV의 방향 정책을 학습하는 proximal policy optimization (PPO)기반의 directional policy reinforcement learning (DPRL) 알고리즘을 통한 변침점 생성기반의 AUV 임무 수행 전략을 제안한다.

그림 1은 AUV의 임무 수행 전략의 전체 시스템 개념도이다. 본 알고리즘의 목표는 동적 환경에서 장애물을 회피하며 정해진 목표물까지의 기동이다. AUV는

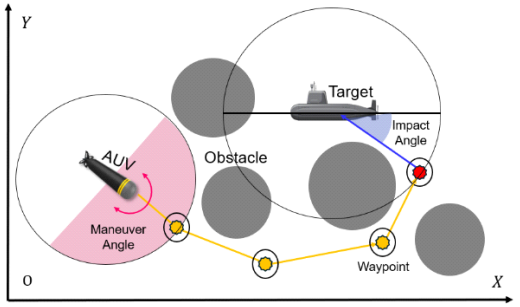


그림 1. 동적 환경에서의 AUV 변침점 기반 임무 수행 시스템 개념도
Fig. 1. Overview of AUV waypoint generation-based mission performance in a dynamic environment

일정 시간마다

그림 1에서 붉은색으로 표현된 각도만큼을 자유롭게 회전할 수 있으며, 각도 전환이 필요할 때 변침점을 생성한다. 또한 목표물에 대한 접근 각도를 나타내는 충격각도를 설정함으로써 실제 임무에서 목표물에게 AUV의 위치 노출의 위험성을 최소화하는 제어 전략을 수립하고자 한다. 따라서 최종적으로 본 DPRL 알고리즘이 deep deterministic policy gradient (DDPG)[6] 알고리즘과 비교하여 보상 함수 수렴성과 목표 지점까지의 변침점 생성, 총 거리 및 성공률, 마지막으로 임무 수행 추론 결과 비교를 통하여 제안하는 DPRL 알고리즘의 우수성과 실제 해양 환경에서의 적용 가능성을 평가한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 본 논문에서 다루는 AUV 모델링에 대하여 다루고, 3장에서는 제안하는 동적 환경에서의 강화학습 기반 AUV 임무 수행의 변침점 생성을 위한 DPRL 알고리즘에 대하여 살펴본다. 4장에서는 제안하는 알고리즘에 대한 성능 평가를 통하여 본 알고리즘의 우수성을 확인하고, 마지막으로 5장에서 결론을 도출한다.

Ⅱ. AUV 모델링

실제 해양환경에 적용을 고려하기 위하여 이번 장에서는 AUV의 운동방정식에 대하여 다룬다. 본 논문에서 사용한 AUV 동역학 모델은 그림 2에서의 6 자유도 운동을 하는 AUV 모델의 roll ϕ, pitch θ에 대한 운동을 무시하고 수평 운동인 surge u, sway v, yaw ψ 운동에 집중하여 3자유도 모델로 근사화하여 활용한다. AUV 모델의 운동방정식은 Fossen (2011)[7]에 의해 제시된 방정식을 사용했다. AUV는 지구 지표에 고정되어 있으며 x축이 진북, z 축이 지구 중심을 가리키는 north-east-down 좌표계로 표현되며, 그 중심이 무인 잠

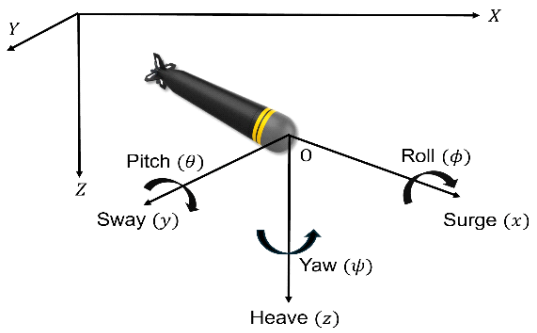


그림 2. 지구고정좌표계 및 AUV고정좌표계의 AUV 모델
Fig. 2. AUV model in body-fixed frame and earth-fixed frame

수정의 부력의 중심인 AUV고정좌표계로 표현할 수 있다. AUV 모델에 대한 3 자유도 동역학 식은 다음과 같다.

$$\nu = [u\ v\ r]^T, \quad \eta = [x\ y\ \psi]^T \tag{1}$$

$$J(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

$$\dot{\eta} = J(\eta)\nu \tag{3}$$

식 (1)의 ν에서, u와 v는 x, y 방향으로의 선속도를 나타내고, r은 각속도를 의미한다. 또한, η의 x, y는 선형 위치를, ψ는 오일러 각도를 의미한다. AUV의 식 (2)의 회전행렬 J는 식 (3)을 통해 AUV고정좌표계의 ν를 지구고정좌표계의 $\dot{\eta}$로 변환할 수 있다.

이러한 모델링을 바탕으로 본 연구에서는 변침점을 최소화하는 경로를 찾고자 한다. 각도 변환 시 생성되는 변침점은 [8]에 따라 추가적인 추진력을 요구하며 이동거리 외에 불필요한 에너지 소비를 유발하므로 변침점이 많을수록 전체 에너지 소모량이 늘어나 AUV의 효율이 감소하게 된다. 또한 각 변침점에서의 방향 전환은 [9]에 따라 일정 시간이 소요되며, 변침점 개수가 많을수록 이러한 시간이 누적된다. 따라서, 변침점이 증가할수록 AUV가 목표 지점에 도달하기까지의 효율성 및 임무 수행 시간이 길어지게 된다. 따라서 본 연구에서는 변침점을 최소화하는 최적 경로를 찾고자 한다.

Ⅲ. 동적 환경에서의 강화학습기반 AUV 제어

3.1 강화학습 개요

강화학습은 기계학습의 한 유형으로, 지도학습이나 비지도학습과는 다르게 고정된 데이터셋에 의존하지 않고, 불확실한 환경에서 수집된 경험을 통해 학습이 진행된다[10]. 따라서 강화학습의 경우 데이터셋 없이도 다양한 환경을 탐색할 수 있다는 장점이 있다[11]. 그러므로 강화학습은 에이전트가 환경에서 상호작용을 통해 시행착오를 겪으며 생성되는 경험을 통하여 학습된다. 에이전트는 현재 상태에서 행동을 선택하고, 그 결과로 얻은 보상에 따라 다음 상태로 전이된다. 이러한 반복적인 과정에서 에이전트는 보상을 기반으로 최적의 정책을 학습할 수 있다. 이때 정책이란 에이전트가 현재 상태를 입력으로 받았을 때 취할 행동을 결정하는 함수이다. 따라서 에이전트는 높은 보상을 받을 수 있는 최적의 정책을 학습하여 추론할 수 있도록 하는 것이 매우 중요하다. 이와 같이 시간에 따른 의사결정을 다루는 문제를 순차적 의사결정(sequential decision making) 문제라고 하며, 강화학습은 MDP를 통해 다양한 상태와 행동을 가진 시스템에서 문제를 수학적으로 모델링할 수 있다. MDP는 현재 상태가 과거 상태들에 무관하며 오직 직전 상태에만 의존한다는 마르코프 성질(Markov property)을 기반으로 하며, 다음과 같이 표현된다.

$$MDP \equiv (S, A, P, R, \gamma) \tag{4}$$

여기서 S는 상태, A는 행동, P는 전이 확률 행렬, R은 보상, 마지막으로 γ는 감쇠 인자를 나타낸다. 강화학습은 순차적 의사결정 문제를 효과적으로 해결할 수 있도록 학습된다. 본 논문에서는 심층강화학습(deep reinforcement learning, DRL)을 적용한다. DRL은 전통적인 Q-learning[12]이나 dynamic programming 알고리즘에 비해 고차원 상태 공간을 처리하고, 복잡한 환경에서도 효율적으로 학습할 수 있는 장점이 있다. Q-learning은 테이블 기반으로 상태-행동 가치 함수를 업데이트하기 때문에 상태 공간이 커지면 계산 비용이 급격히 증가하는 반면, DRL은 신경망을 활용해 이러한 문제를 극복하고 일반화 능력을 제공한다. 또한, dynamic programming은 환경의 전이 확률을 알고 있어야 하는 제약이 있지만, DRL은 전이 확률을 모르는 환경에서도 경험을 통해 학습할 수 있다. 따라서 DRL은 더 복잡하고 예측 불가능한 실제 환경에서도 효과적인 학습이 가능하다. 본 논문에서는 DRL 알고리즘을 활용하여, 동적이 환경에서도 유연한 대처가 가능한 DPRL 알고리즘을 제안한다.

3.2 상태

본 논문의 강화학습 에이전트인 AUV의 상태 정보는 시간 t에 따라 다음과 같이 정의된다.

$$State_t = \{s_t, s_{t-1}, s_{WP}, s_{PA}, s_{IA}, s_{tar}\} \quad (5)$$

$s_t = \{x_t, y_t\}$는 시간 t에서의 AUV의 위치를 나타내며, $s_{t-1} = \{x_{t-1}, y_{t-1}\}$는 시간 $t-1$, 즉 바로 이전의 AUV의 위치를 나타낸다. 또한 회전 각도가 달라질 때마다 시간 t까지의 생성된 변침점의 집합은 $s_{WP} = \{(x_1, y_1), \cdots, (x_N, y_N)\}$로 저장되며, 여기서 시간 N은 생성된 변침점의 개수이다. AUV의 기동이 제한되는 장애물은 반지름 r을 갖는 중심의 좌표 합인 $s_{PA} = \{(x_1, y_1, r_1), \cdots, (x_M, y_M, r_M)\}$으로 정의된다. M은 장애물의 개수이다. s_{IA}는 최종적으로 목표물에 도달할 때 반드시 만족해야 하는 목표물의 충격 각도 ϕ를 고려한 변침점으로, $s_{IA} = \{x_{IA}, y_{IA}\}$로 정의된다. 이때 충격각도를 고려한 변침점의 좌표는 $x_{IA} = x_{tar} + d\sin\phi$, $y_{IA} = y_{tar} + d\cos\phi$로 계산된다. AUV는 기동 시 반드시 충격각도를 만족하는 변침점을 지나야하며, 이 충격 각도를 통하여 실제 임무에서 목표물에게 AUV의 위치 노출의 위험을 최소화하는 제어 전략을 수립할 수 있다. 마지막으로 $s_{tar} = \{x_{tar}, y_{tar}\}$은 목표물의 위치를 나타낸다.

본 연구에서 사용된 환경은 AUV의 초기 위치를 (0,0)으로 설정하고, x축과 y축을 각각 -1에서 +1 범위로 매핑한 연속적인 값을 가질 수 있는 2×2 크기의 연속적인 공간으로 구성된다. 이러한 환경 설계는 AUV가 360도의 모든 방향으로 자유롭게 기동할 수 있는 유연성을 제공하며, 강화학습의 효율성을 극대화하기 위함이다. 연속적인 공간 표현은 격자 기반 표현보다 높은 해상도를 제공하며, 실제 해양 환경의 복잡성을 반영하여 장애물 회피 및 최적 경로 탐색을 지원한다. 또한, 좌표계를 표준화하여 다양한 시뮬레이션 및 실제 환경 간의 일관성을 보장하며, 강화학습 알고리즘의 상태 및 행동 처리 과정을 간소화하였다. 이와 같이 구성된 환경에서 상태 정보는 현재의 AUV가 어떤 상황에 있는지 나타내며 이를 기반으로 행동을 선택하게 된다. 이를 통해 최종적으로AUV는 목표물을 향해 특정 각도로 궤적을 조정하면서 변침점을 생성하여, 여러 개의 장애물을 피하고, 지정된 충격 각도를 준수하는 제어 전략을 수립할 수 있다.

3.3 행동

AUV의 기동 안정성을 유지하기 위해, AUV는 일정 시간마다 최대 90도만큼을 회전하며 일정 거리 d만큼

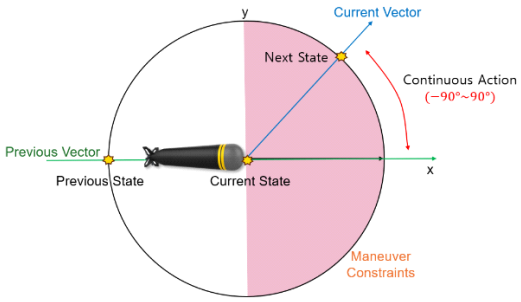


그림 3. 최대회전각도에 따른 AUV의 행동 집합
Fig. 3. Action space of AUV according to the maximum rotation angle

기동할 수 있다. 따라서 행동 집합은 그림 3에서의 붉은색 구 형태로 나타낼 수 있고, 다음과 같이 나타낼 수 있다.

$$A_t = \{\theta \mid -90^\circ \le \theta \le 90^\circ\} \quad (6)$$

행동에 따라 AUV는 s_{t-1}에서의 진행 방향 벡터를 기준으로 θ만큼 회전한 벡터로 방향을 전환하게 된다. 이는 그림 3에서의 붉은색의 구 위의 한 점으로 나타낼 수 있다. 따라서 다음 상태는 다음과 같이 $s_{t+1} = x_t + d * \cos(\theta)$, $y_t + d * \sin(\theta)$와 같이 업데이트 된다.

3.4 보상 함수

AUV의 보상 함수는 임무 경로의 안정성과 효율성을 보장하는 것을 목표로 하며, 동적인 환경에 대하여 안정적인 학습이 가능할 수 있도록 임무 성공에 대한 조건을 추가하여 다음과 같이 구성하였다. 특히, R_1과 R_2를 곱한 형태로 표현하여 에피소드 종료 상태에 따른 전체적인 학습 목표와 스텝 별 비용의 기여를 동시에 반영할 수 있도록 설계하였다.

$$R = R_1 + R_1 \cdot R_2 \quad (7)$$

이때 R_1은 에피소드 종료 여부에 따라 양수 또는 음수 보상을 부여하는 컨트롤 변수로, 성공적인 목표 도달 시 양의 보상을, 실패 시 음의 보상을 부여한다.

$$R_1 = \begin{cases} +100, & if \quad Done == True \\ -1, & Otherwise \end{cases} \quad (8)$$

이때 에피소드 종료 조건은 목표물과 충격각도를 고려한 변침점 그리고 새로 업데이트 된 AUV의 위치가 최대 회전각도를 만족하는 경우이다. 동적인 환경에서도

학습이 안정적으로 진행되어야 하므로 충격 각도를 고려한 변침점과 목표물까지의 도달이 아닌 새로 업데이트 된 현재 위치와 충격각도를 고려한 변침점 그리고 목표물까지, 총 세 점을 이었을 때의 기동 가능 여부로 설정하여 에피소드 종료의 성공률을 높였다. 이를 통하여 보다 다양한 동적인 환경에서도 안정적인 학습을 진행할 수 있다. R_2는 목표물에 도달했는지 못하였을 때의 부여되는 보상으로, 목표물과 현재 위치 사이의 거리와 현재 위치로 업데이트를 위해 AUV가 기동한 거리와 전환한 각도의 크기가 라디안 값으로 적용된다. 다음 수식을 통하여 AUV는 불필요한 각도 전환을 억제하고 효율적인 경로로 학습된다.

$$R_2 = \sqrt{(s_{tar} - s_t)^2} + d + |\theta| \tag{9}$$

최종적으로 위 보상함수를 통하여 AUV는 각 step별로 목표물까지의 주행이 성공하면 양수 보상을, 실패하면 거리에 따른 음수 보상이 부여된다. 이를 통해 AUV는 동적인 환경에서 충격 각도 및 최대회전각도를 만족하는 효율적인 임무 수행 전략을 수립할 수 있다.

3.5 DPRL 알고리즘의 방향 정책 학습

본 논문에서는 AUV의 최적 방향 정책을 학습시키기 위하여 PPO기반의DPRL 알고리즘을 제안한다. DPRL 알고리즘은 PPO[13]와 동일한 학습 원리를 가지며, actor-critic 구조로 구성된다. Actor 네트워크는 관측 값을 바탕으로 행동을 생성하고, 이를 경험 재생 버퍼에 저장한다. 반면 critic 네트워크는 이득 함수를 사용하여 전략을 평가한다. DPRL알고리즘은 이 과정을 반복하여 AUV로 하여금 보상을 최대화하는 최적 방향 정책 π^*을 학습할 수 있다. 이러한 원리를 바탕으로 AUV가 주어진 상태에서 어떤 행동을 취할지 확률을 직접적으로 학습하는 방식을 채택하여 정책 함수의 파라미터를 업데이트하여 최적의 정책을 찾는다. 또한 정책 업데이트 시 클립핑을 활용하여 급격한 변화인 업데이트가 지나치게 큰 경우를 방지하여 수렴성을 높이고 동적 환경에서의 안정적으로 학습이 될 수 있도록 설계하였다.

Ⅳ. 성능 평가

동적 환경에서의 DPRL 알고리즘을 검증하기 위하여 제안한 알고리즘과의 동일한 환경 및 보상함수를 적용한 기존 DDPG 알고리즘과 비교하여 평가를 진행하였다. PPO를 기반으로 하는 DPRL과 actor-critic 구조를 기반으로 세밀한 탐색 능력을 제공하는 DDPG는 연속적인 환경에서 각각 학습 안정성과 탐색 성능의 상반된 특성을 보인다. 특히, DDPG는 경로 계획 문제에서 널리 활용되는 알고리즘으로, 이를 비교함으로써 DPRL의 학습 안정성, 탐색 성능 및 적용 가능성을 종합적으로 평가하고자 하였다. 본 알고리즘의 AUV 임무 수행 효율성을 판단하기 위해 보상 함수 수렴성, 변침점 생성 개수, AUV 기동 거리 및 성공률로 평가하고, 추론된 시각화 결과를 통해 검증한다.

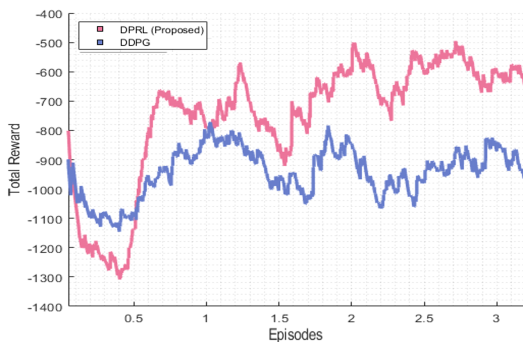


그림 4. 에피소드 진행에 따른 누적 보상
Fig. 4. Total Reward according to Episodes

4.1 에피소드 진행에 따른 보상 함수 수렴성

그림 4를 통하여 강화학습 알고리즘인 DPRL과 DDPG의 학습에서의 안정성을 확인할 수 있다[14]. 본 그래프에서는 X축에서의 총 300000번의 에피소드가 진행됨에 따라 Y축에서의 AUV가 학습 과정에서 얻은 누적 보상을 나타낸다. 제안된 DPRL과 DDPG 알고리즘 모두 학습을 진행할수록 누적 보상이 증가하며, 일정 에피소드 이후 수렴하는 것을 확인할 수 있다. 하지만 제안된 알고리즘이 기존 DDPG 알고리즘보다 더 빠르게 누적 보상을 얻고 있으며, 더 높은 보상 값에 도달하는 것을 확인할 수 있다. 또한, 보상의 변동폭이 기존 DDPG알고리즘에 비해 적은 것을 통하여 안정적인 학습을 진행함을 알 수 있다. 따라서, 제안된 DPRL 알고리즘은 기존의 DDPG보다 빠르게 학습이 진행되며, 더 높은 보상을 안정적으로 얻는 성능 우위를 확인할 수 있다. 이는 제안된 알고리즘이 강화학습 문제에서 더욱 효과적이고 안정적인 학습 방법임을 확인할 수 있다.

4.2 환경 복잡도에 따른 변침점 생성

그림 5는 환경 복잡도에 따른 변침점 생성 개수를 나타낸 결과이다. X축은 환경 복잡도(장애물 개수를 환경 크기로 나눈 값)를 나타내며, Y축은 생성된 변침점의 개수를 나타낸다. 제안된 DPRL과 기존 DDPG의

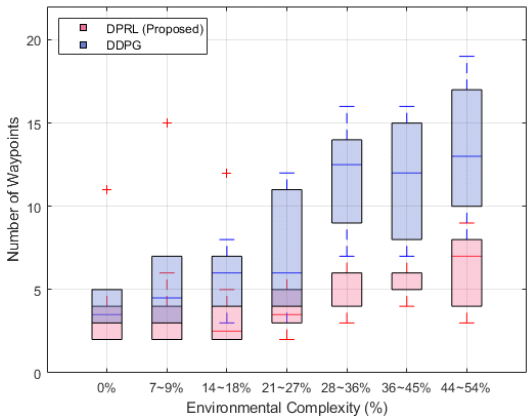


그림 5. 환경 복잡도에 따른 변침점 생성 결과
Fig. 5. Number of generated waypoint according to environment complexity

성능 차이를 비교하기 위한 그래프로, 환경 복잡도가 증가함에 따라 변침점이 어떻게 변화하는지를 비교할 수 있다. 제안된 DPRL은 그래프에서 붉은색 상자로 표시되며, 제안된 DPRL은 환경 복잡도가 낮을 때와 높을 때 모두 DDPG와 비교하여 변침점 개수가 더 적게 생성되는 것을 확인할 수 있다. 이는 제안된 알고리즘이 목표 지점에 도달하는 AUV의 임무 수행에 있어서 더욱 효율적인 변침점 생성 알고리즘임을 확인할 수 있다. 또한, 이는 환경 복잡도가 높아질수록 두 알고리즘 간의 차이가 뚜렷해진다. 기존 알고리즘은 복잡한 환경에서 변침점 개수가 증가하는 경향을 보이며, 변침점 생성의 변동성 또한 크다. 반면, 제안된 알고리즘은 환경 복잡도가 낮을 때와 높을 때 모두 DDPG와 비교하여 변침점 개수가 더 적게 생성되는 것을 확인할 수 있다. 이는 제안된 알고리즘이 목표 지점에 도달하는 AUV의 임무 수행에 있어서 더욱 효율적인 변침점 생성 알고리즘임을 확인할 수 있다. 또한, 이는 환경 복잡도가 높아질수록 두 알고리즘 간의 차이가 뚜렷해진다. 기존 알고리즘은 복잡한 환경에서 변침점 개수가 증가하는 경향을 보이며, 변침점 생성의 변동성 또한 크다. 반면, 제안된 알고리즘은 환경 복잡도가 높은 동적 환경에서도 더욱 적은 변침점을 생성하며, 이를 기반으로 보다 AUV가 효율적으로 임무를 수행할 수 있음을 확인할 수 있다.

4.3 환경 복잡도에 따른 AUV 임무 수행을 위한 평균 거리

그림 6은 환경 복잡도에 따라 구성된 동적 환경에서의 총 20번의 AUV 임무 수행을 위하여 생성된 변침점 기반으로 구축된 거리의 평균을 구하여 비교한 결과이다. X축은 환경 복잡도를, Y축은 AUV가 임무를 수행

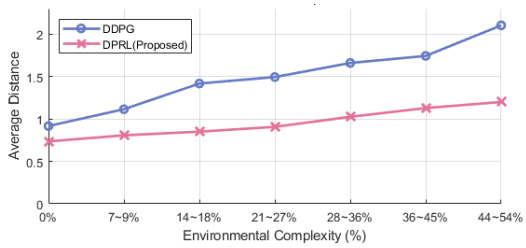


그림 6. 환경 복잡도에 따른 AUV 평균 거리
Fig. 6. Average distance for AUV to environment complexity

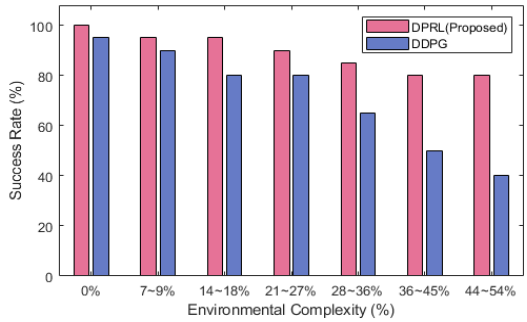


그림 7. 동적 환경에 따른 AUV 임무 수행 성공률
Fig. 7. AUV mission success rate according to dynamic environment

하는 동안 이동한 평균 거리를 나타낸다. 제안된 DPRL은 붉은색 선으로, 기존 DDPG는 파란색 선으로 표시된다. 그래프에서 볼 수 있듯이, 환경 복잡도가 증가함에 따라 기존 DDPG의 평균 거리는 점차적으로 증가하는 반면, 제안된 DPRL의 경우 DDPG와 비교하였을 때 상대적으로 짧고 일정하게 유지되는 것을 확인할 수 있다. 이는 환경 복잡도가 증가할수록 기존 알고리즘이 비효율적으로 더 긴 경로를 선택하게 되는 반면, 제안된 알고리즘은 환경의 복잡성에 크게 영향을 받지 않고 효율적인 경로를 선택할 수 있음을 의미한다. 제안된 알고리즘은 평균 거리를 짧게 유지하면서도 안정적인 성능을 바탕으로, AUV가 복잡한 환경에서도 더 적은 거리를 이동하면서 임무를 수행할 수 있음을 검증할 수 있다. 따라서 제안된 DPRL 알고리즘은 기존 DDPG 알고리즘과 비교하였을 때 보다 효율적인 임무 수행이 가능함을 확인할 수 있다.

4.4 동적 환경에서의 AUV 임무 수행 성공률

그림 7은 동적 환경에서 AUV의 임무 수행 성공률을 제안된 DPRL과 기존DDPG에서의 결과를 비교한 그래프이다. 이때 임무 수행 성공률은 변침점 개수가 10개를 넘지 않으며, 충격 각도를 만족하여 임무를 수행한

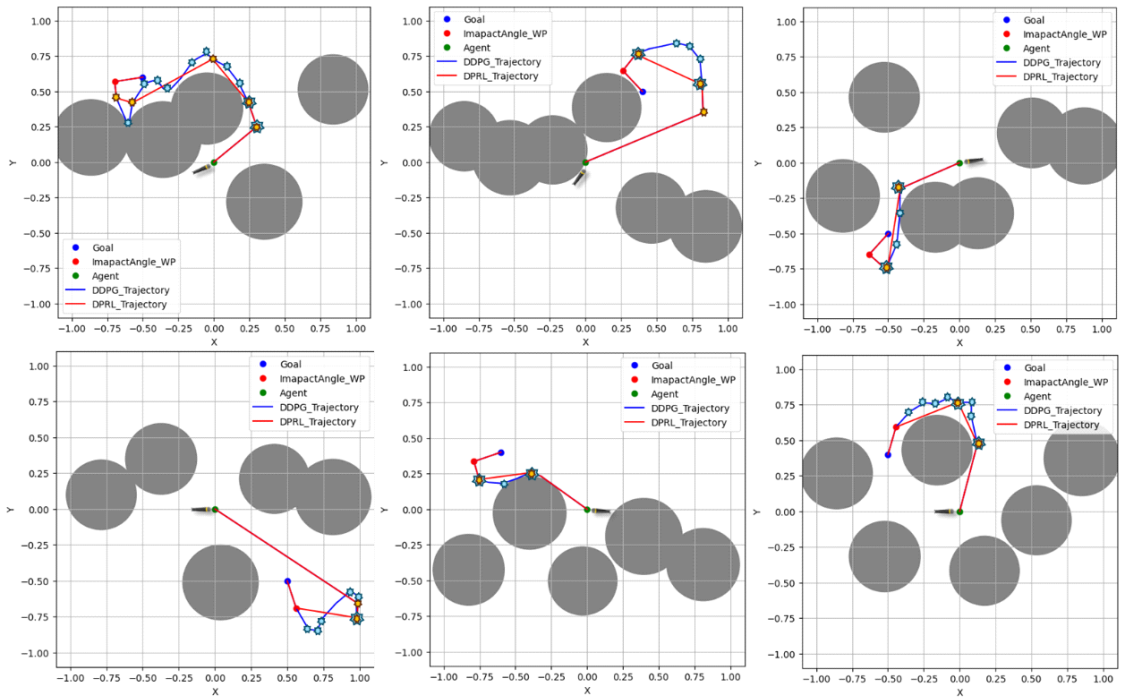


그림 8. 제안한 DPRL과 DDPG 알고리즘기반 생성된 변침점 기반 AUV 임무 수행 추론 결과
Fig. 8. AUV mission performance result based on generated waypoint via proposed DPRL, DDPG algorithm

경우를 의미한다. 이때 다양한 환경 복잡도를 갖는 50개의 동적 환경에서의 추론 결과를 바탕으로 검증을 진행하였다. 그래프의 X축은 환경 복잡도를, Y축은 임무 성공률(%)을 나타내며, 이를 통해 두 알고리즘의 환경 복잡도가 증가할 때 임무 성공률이 어떻게 변화하는지를 확인할 수 있다. 제안된 DPRL은 붉은색 막대로, 기존 알고리즘 DDPG은 파란색 막대로 표시하였다. 그래프를 통하여 제안된 알고리즘은 모든 환경 복잡도에서 매우 높은 성공률을 유지하며, 특히 복잡도가 높아질수록 기존 알고리즘인 DDPG와의 차이가 커짐을 확인할 수 있다. 반면, 기존 알고리즘 DDPG는 환경 복잡도가 증가함에 따라 성공률이 급격히 감소하는 경향을 볼 수 있다. 특히 환경 복잡도가 36% 이상일 때, 기존 알고리즘의 성공률이 크게 떨어지며, 가장 복잡한 환경(44~55%)에서는 성공률이 매우 낮아지는 반면, 제안된 알고리즘은 여전히 높은 성공률을 유지함을 확인할 수 있다. 따라서 제안된 DPRL 알고리즘이 동적이고 복잡한 환경에서도 AUV 임무를 성공적으로 수행할 수 있으며, 기존 DDPG 알고리즘보다 더 높은 성능을 보여줌을 통하여 우수성과 안정성을 확인할 수 있다.

4.5 변침점 기반 AUV 임무 수행 추론 결과

그림 8은 제안된 DPRL 알고리즘과 기존 DDPG 알고리즘의 AUV 임무 수행 변침점 생성 추론 결과를 비교한 시각화한 것이다. 장애물의 사이즈와 목표물의 위치, 충격 각도가 랜덤으로 조성되는 동적 환경에서의 임무 수행을 위한 변침점 생성 결과를 시각적으로 표현하고 있다. 장애물은 흑백 원으로 나타나 있고, AUV의 목표물은 파란색 점, 충격 각도를 만족하는 변침점은 빨간색 점, AUV 출발 지점은 (0,0)으로 초록색 점으로 환경이 구성되어 있다. 시각화 결과에서 붉은색 선과 노란색 별표는 제안된 DPRL 알고리즘의 변침점 생성 결과를, 파란색 선과 파란색 별표는 기존 DDPG 알고리즘에서의 변침점 생성 결과를 나타낸다. 이렇게 구성된 추론 결과는 AUV가 장애물을 피하며 목표 물까지 도달하는 과정을 나타내며, 제안된 알고리즘과 기존 알고리즘 간의 변침점 생성의 차이를 비교할 수 있다. 제안된 DPRL 알고리즘은 DDPG 알고리즘에 비해 더 짧고 효율적인 변침점 생성 결과를 보여주며, 변침점 기반으로 생성된 경로가 장애물에 부딪히지 않고 안정적으로 목표 지점에 도달하고 있음을 확인할 수 있다. 반면, DDPG 알고리즘은 비교적 더 많은 변침점을 생성하고 복잡한 경로를 따라가며 목표 지점에 도달하게 된다.

이러한 추론 시각화 결과를 통하여 제안된 DPRL 알고리즘이 기존 DDPG 알고리즘과 비교하여 우수성을 검증할 수 있다.

V. 결 론

본 논문은 불확실한 동적인 해양 환경에서의 방향 정책 학습을 통하여 AUV의 목표물 도달을 위한 강화학습 기반 AUV 제어 DPRL 알고리즘을 제안한다. 이때 실제 해양 환경에서의 동적인 환경을 고려하여 AUV의 효율적인 기동 전략을 위한 MDP기반의 강화학습 알고리즘을 설계하였다. 성능 평가를 통하여 여러 동적인 환경에서도 DPRL 알고리즘은 타 알고리즘과 비교하여 더 효율적으로 변침점을 생성하는 우수한 성능과 안정적인 성공률을 보임을 통하여 효과적임을 증명하였다. 이를 통하여 본 알고리즘을 통해 실제 해양에서도 AUV의 자율성 및 효율성을 증대하는 전략을 토대로 적용할 수 있을 것으로 기대된다.

References

[1] G. Ioannou, N. Forti, L. M. Millefiori, S. Carniel, A. Renga, G. Tomasicchio, S. Binda, and P. Barca, “Underwater inspection and monitoring: Technologies for autonomous operations,” *IEEE Aerospace and Electr. Syst. Mag.*, vol. 39, no. 5, pp. 4-16, May 2024, (https://doi.org/10.1109/MAES.2024.3366144)

[2] Y. R. Petillot, G. Antonelli, G. Casalino and F. Ferreira, “Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles,” *IEEE Robotics & Automat. Mag.,* vol. 26, no. 2, pp. 94-101, Jun. 2019. (https://doi.org/10.1109/MRA.2019.2908063)

[3] E. J. Roh, H. Lee, S. Park, J. Kim, K. Kim, and S. Kim, “Directional autonomous torpedo maneuver control using reinforcement learning,” *J. KICS,* vol. 49, no. 5, pp. 752-761, May 2024. (https://doi.org/10.7840/kics.2024.49.5.752)

[4] N. Dai, P. Qin, X. Xu, Y. Zhang, Y. Shen, and B. He, “An AUV collision avoidance algorithm in unknown environment with multiple constraints,” *Ocean Eng.,* vol. 294, p. 116846, Feb. 2024. (https://doi.org/10.1016/j.oceaneng.2024.116846)

[5] R. Bellman, “A Markovian decision process,” *J. Mathematics and Mechanics*, vol. 6, no. 5, pp. 679-684, Apr. 1957. (https://doi.org/10.1512/iumj.1957.6.56038)

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in *Proc. 4th ICLR*, Caribe Hilton, San Juan, Puerto Rico, May 2016.

[7] T. I. Fossen, “*Handbook of Marine Craft Hydrodynamics and Motion Control*,” Wiley, Apr. 2011. (https://doi.org/10.1002/9781119994138)

[8] H. Niu, Z. Ji, A. Savvaris, and A. Tsourdos, “Energy efficient path planning for unmanned surface vehicle in spatially-temporally variant environment,” *Ocean Eng.,* vol. 196, p. 106766, Jan. 2020. (https://doi.org/10.1016/j.oceaneng.2019.106766)

[9] J. Yang, J. Huo, M. Xi, J. He, Z. Li, and H. H. ong, “A time-saving path planning scheme for autonomous underwater vehicles with complex underwater conditions,” *IEEE Internet of Things J.,* vol. 10, no. 2, pp. 1001-1013, Jan. 2023. (https://doi.org/10.1109/JIOT.2022.3205685)

[10] W. J. Yun, D. Kwon, M. Choi, J. Kim, G. Caire, and A. F. Molisch, “Quality-aware deep reinforcement learning for streaming in infrastructure-assisted connected vehicles,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2002-2017, Feb. 2022 (https://doi.org/10.1109/TVT.2021.3134457)

[11] W. J. Yun, S. Park, J. Kim, M. Shin, S. Jung, D. A. Mohaisen, and J-H. Kim, “Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control,” *IEEE Trans. Industrial Inf.,* vol. 18, no. 10, pp. 7086-7096, Oct. 2022 (https://doi.org/10.1109/TII.2022.3143175)

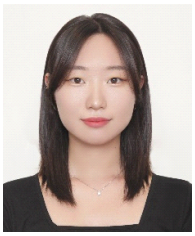
[12] C. Watkins and P. Dayan, “Q-learning,”

Mach. Learn., vol. 8, pp. 279-292, May 1992. (https://doi.org/10.1007/BF00992698)

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, Jul. 2017.

[14] C. Park, G. S. Kim, S. Park, S. Jung, and J. Kim, "Multi-agent reinforcement learning for cooperative air transportation services in city-wide autonomous urban air mobility," *IEEE Trans. Intell. Veh.*, vol. 8, no. 8, pp. 4016-4030, Aug. 2023. (https://doi.org/10.1109/TIV.2023.3283235)

노 지 민 (Emily Jimin Roh)



2024년 2월 : 세종대학교 지능기전공학부 무인이동체공학전공 졸업 (공학사)
2024년 3월~현재 : 고려대학교 전기전자공학과 석박사통합과정

<관심분야> Reinforcement Learning, Autonomous Mobility, Quantum Machine Learning
[ORCID:0009-0008-0013-6342]

이 현 수 (Hyunsoo Lee)

2021년 2월 : 숭실대학교 전자정보공학부 졸업 (공학사)
2021년 3월~현재 : 고려대학교 전기전자공학과 석박사통합과정
<관심분야> Reinforcement Learning, Electronic Engineering, Communication Engineering
[ORCID:0000-0003-1113-9019]

송 일 석 (Ilseok Song)

2008년 2월 : 건국대학교 컴퓨터시스템 졸업 (공학사)
2024년 9월~현재 : 고려대학교 SW·AI융합대학원 석사
2007년 12월~현재 : LIG넥스원
<관심분야> 수중체계, 유도무기, 유무인복합체계, 강화학습
[ORCID:0009-0009-9056-526X]

김 승 환 (Seunghwan Kim)

2009년 2월 : 숭실대학교 정보통신전자공학 졸업 (공학사)
2011년 2월 : 숭실대학교 정보통신공학 석사 졸업
2011년1월~2014년 6월 : 산엔지니어링
2015년 1월~현재 : LIG넥스원
<관심분야> 수중체계, 어뢰기동, 해양
[ORCID: 0000-0002-5841-1879]

김 영 대 (Youngdae Kim)

2002년 2월 : 아주대학교 전자공학 졸업 (공학사)
2002년 4월~현재 : LIG넥스원
<관심분야> 수중감시, 유도무기, 유무인복합체계, 강화학습
[ORCID:0009-0006-5927-0349]

박 수 현 (Soohyun Park)

2019년 2월 : 중앙대학교 컴퓨터공학과 졸업 (공학사)
2023년 8월 : 고려대학교 전기전자공학과 졸업 (공학박사)
2023년 9월~2024년 2월 : 고려대학교 정보통신기술연구소 박사후연구원
2024년 3월~현재 : 숙명여자대학교 소프트웨어학과 조교수
<관심분야> Deep Learning Theory, Network/Mobility Applications, Quantum Machine Learning, AI-based Autonomous Control
[ORCID:0000-0002-6556-9746]

김 중 헌 (Joongheon Kim)

2004년 2월 : 고려대학교 컴퓨터학과 졸업 (이학사)
2006년 2월 : 고려대학교 컴퓨터학과 석사
2014년 8월 : University of Southern California Computer Science 박사
2016년 3월~2019년 8월 : 중앙대학교 소프트웨어대학 조교수
2019년 9월~현재 : 고려대학교 전기전자공학부 부교수
<관심분야> Stochastic Optimization, Mobility, Reinforcement Learning, Quantum
[ORCID: 0000-0003-2126-768X]