대한조선학회논문집
Journal of the Society of Naval Architects of Korea
pISSN:1225-1143, Vol. 56, No. 1, pp. 58-65, February 2019
eISSN:2287-7355, https://doi.org/10.3744/SNAK.2019.56.1.058

심층 결정론적 정책 경사법을 이용한 선박 충돌 회피 경로 결정

김동함[1]·이성욱[2]·남종호[2,†]·요시타카 후루카와[3]
한국해양대학교 조선해양시스템공학과[1]
한국해양대학교 조선해양시스템공학부[2]
큐슈대학교 마린시스템공학과[3]

Determination of Ship Collision Avoidance Path using Deep Deterministic Policy Gradient Algorithm

Dong-Ham Kim[1]· Sung-Uk Lee[2]·Jong-Ho Nam[2,†]·Yoshitaka Furukawa[3]
Department of Naval Architecture and Ocean Systems Engineering, Graduate School, Korea Maritime and Ocean University, Korea[1]
Division of Naval Architecture and Ocean Systems Engineering, Korea Maritime and Ocean University, Korea[2]
Department of Marine Systems Engineering, Kyushu University, Fukuoka, Japan[3]

The stability, reliability and efficiency of a smart ship are important issues as the interest in an autonomous ship has recently been high. An automatic collision avoidance system is an essential function of an autonomous ship. This system detects the possibility of collision and automatically takes avoidance actions in consideration of economy and safety. In order to construct an automatic collision avoidance system using reinforcement learning, in this work, the sequential decision problem of ship collision is mathematically formulated through a Markov Decision Process (MDP). A reinforcement learning environment is constructed based on the ship maneuvering equations, and then the three key components (state, action, and reward) of MDP are defined. The state uses parameters of the relationship between own-ship and target-ship, the action is the vertical distance away from the target course, and the reward is defined as a function considering safety and economics. In order to solve the sequential decision problem, the Deep Deterministic Policy Gradient (DDPG) algorithm which can express continuous action space and search an optimal action policy is utilized. The collision avoidance system is then tested assuming the 90°intersection encounter situation and yields a satisfactory result.

Keywords : Collision avoidance(충돌 회피), Reinforcement learning(강화학습), Deep Deterministic Policy Gradient(DDPG, 심층결정론적정책경사법), Markov Decision Process(MDP, 마르코프결정과정), Autonomous ship(자율운항선박)

1. 서 론

최근 국제해사기구(IMO) 해사안전위원회(MSC)의 이내비게이션(e-navigation)에 대한 채택 승인과 관련하여 선박의 자동화를 위한 활동이 증가하고 있는 추세다. 해사안전위원회의 승인내용(제85차, 제94차)은 ICT와 융합을 통하여 차세대 해양안전 종합관리 체계를 마련하고자 하는 데 목적을 두고 있다(Shim et al., 2010; Jeong, 2015). 이와 더불어 선박을 건조하거나 관련 항해 장비를 개발하는 기업에서는 스마트선박(smart ship) 개념을 도입하여 이와 관련된 시스템 개발에 힘쓰고 있다. 스마트선박은 기존의 선박에 비해 운항 경제성 및 안전성을 대폭 향상시킨 선박이라 정의할 수 있다(Van, 2007).

선박 운항의 자동화에 있어 가장 비중 있게 고려되고 있는 사항 중 하나는 선박이 항내로 입항 또는 출항할 때 지켜져야 하는 안전성이다. 항내에서 발생할 수 있는 해양사고의 원인으로 여러 가지 요소가 있지만, 그 중 선박들 간의 충돌 및 좌초가 특히 중요한 요소로 고려된다. 실제로 최근 5년(2012~2017년) 사이 중앙해양안전심판원 통계자료(Korean maritime safety tribunal, 2017)에 따르면 충돌 및 좌초와 관련된 국내 해양사고의

Received : 16 July 2018 | Revised : 6 September 2018 | Accepted : 19 October 2018
† Corresponding author : Jong-Ho Nam, jhnam@kmou.ac.kr

96.5%는 항해자의 판단 착오 등의 운항 과실에 의해 발생하고 있으며, 이들 중 97%가 출항 후 직무별 과실 및 안전 수칙 미준수와 같은 직접적인 인적 과실에 기인하는 것으로 보고되었다(Kim & Kwak, 2011). 따라서 스마트선박 또는 이내비게이션(e-Navigation)을 도입하기 위해서는 이러한 충돌 방지 회피 시스템에 대한 연구가 필수적인 요소로 인식될 수 있다.

자동 충돌 회피 시스템 구축에서 가장 어려운 문제로 인식되고 있는 기술은 여러 가지 다양한 패턴의 문제에 대해 획기적으로 자동 충돌 회피를 할 수 있는 강건한 알고리즘의 구축이다. 이는 다른 운송 시스템과 마찬가지로, 일종의 무작위 현상에서 일어나는 문제에 대한 해결방안을 제시할 수 있는 일반적인 규칙을 찾는 문제로 인식 될 수 있다. 하지만 무작위 현상에 대한 일반적인 행동 규범을 찾는 것은 일반적으로 어려운 문제로 인식되어 왔는데, 최근 이에 대한 접근 이론의 하나로 기계학습(machine learning)을 적용한 연구가 수행되고 있다. 구글 딥마인드(DeepMind)가 개발한 알파고(AlphaGo)의 등장은 기계학습이 바둑과 같은 다양한 패턴의 문제해결에도 적용 가능함을 보여주었다(Silver et al., 2016).

본 연구에서는 기계학습을 구성하고 있는 요소기술 중 하나인 강화학습(reinforcement learning)을 이용하여 선박 충돌 문제를 해결하고자 한다. 강화학습의 목적은 환경(environment)으로부터 받는 보상(reward)들의 합이 최대가 되는 최적의 행동양식을 학습하는 것이다(Fig. 1). 선박의 조종운동 시뮬레이션을 통해 충돌 시나리오를 반복적으로 구현할 수 있기 때문에 강화학습은 충돌 회피를 위한 최적의 행동을 결정하기에 적합하다.

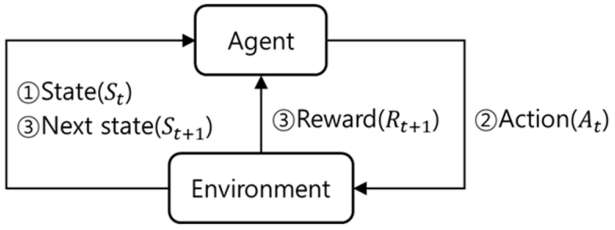

Fig. 1 Interaction between environment and agent

강화의 개념은 행동학자인 스키너(B. Skinner)가 처음 제시하였으며, 쥐가 보상을 통해 행동과 그 결과 사이의 관계를 학습하는 것을 스키너 상자를 이용하여 확인하였다. 이 개념을 컴퓨터 학습에 도입하여 순차적 행동 결정 문제를 푸는 것이 강화학습의 핵심이 된다. 순차적 행동 결정 문제는 MDP(Markov decision process)를 통해 수학적으로 정의될 수 있으며, 벨만(Bellman) 기대 방정식과 벨만 최적 방정식을 통해 MDP의 최적 가치함수와 최적 정책을 찾을 수 있다. 벨만 방정식들은 기본적으로 다이나믹 프로그래밍(dynamic programing)을 통해 풀 수 있으며 이는 살사(SALSA), 큐러닝(Q-learning)으로 발전하였다(Lee et al., 2017).

이후 딥마인드는 인공신경망을 큐러닝에 인공신경망을 적용한 오프폴리시(off-policy) 방법인 DQN(deep Q networks)을 소개하였다(Mnih et al., 2013). 하지만 DQN은 이산화된 행동 공간을 갖기 때문에 행동 공간의 수가 많아지거나 연속적인 행동 제어가 필요한 환경에서는 적용하기 힘들다. 이를 해결하기 위해 결정론적정책경사법(deterministic policy gradient, DPG) 알고리즘(Silver et al., 2014)을 기반으로 엑터-크리틱(actor-critic) 방법을 응용한 심층결정론적정책경사법(deep deterministic policy gradient, DDPG)이 제안되었다(Lillicrap et al., 2016).

충돌 회피 시스템은 자동 항해 시스템을 구성하는 주요 요소로 이와 관련된 다양한 연구가 진행되어 왔다. Kose et al.(1998)은 충돌 위험도에 의한 위험 분포를 지도에 나타내어 안전한 항로를 결정할 수 있도록 보조하는 시스템을 개발하였다. Lee and Rhee (2001), Kijima and Furukawa (2002), Ota et al.(2016)는 모두 퍼지(fuzzy)이론을 사용하여 충돌 위험도를 산출하였다. Lee and Rhee(2001)는 전문가 시스템과 A* 탐색법을 이용하여 충돌 회피 시스템을 개발하였고, Kijima and Furukawa(2002)는 충돌 위험도에 따라 회피 방향의 범위를 변화시키는 방법을 제안하였다. 위 두 방법은 회피 행동공간을 재구성 할 때 행동공간 사이의 변화를 고려하기 힘들다는 단점이 있는데, Son et al.(2009)은 가변공간 탐색법을 이용하여 이 단점을 극복하고자 하였다.

Ota et al.(2016)은 강화학습을 이용한 충돌 회피 시스템을 제안하였으며 자선(own ship)의 속도, 회피 시작 지점, 회피 종료 지점, 회피 경로를 충돌 회피 항로 결정을 위한 중요 요소로 두었다. 각 요소에서 이산적 행동들이 매개변수로 정의되었으며, 두 선박의 특정 조우상황에서 최적 회피 행동을 하는 매개변수들을 찾기 위해 큐러닝 방법을 활용하였다. 하지만 자선과 타선(target ship)의 상태에 따라 주기적으로 최적 행동이 결정되는 대신, 초기 조우 상황에 맞는 하나의 최적 행동이 결정되는데, 이러한 경우 타선의 목표 경로 변경과 같은 변화에 대응하기 힘들다는 단점을 수반한다. 이와 같은 단점을 극복하기 위해서 충돌 회피 문제를 순차적 행동 문제로 정의하는 것이 필요하다.

본 연구에서는 조종운동 방정식을 기반으로 강화학습의 환경을 구축하고, 선박의 충돌 회피 문제를 순차적 행동 문제로 다루기 위해 MDP를 정의한다. 연속된 행동 공간에서 MDP의 최적 정책과 행동을 찾기 위해 DDPG를 이용하며, 이를 통해 선박의 충돌을 회피 경로를 결정한다. 그리고 90도 횡단 상태의 조우 상황에서 MDP를 테스트하여 그 유효성을 검증하였다.

2. 조종운동 방정식 및 수학모델

2.1 조종운동 방정식

일반적으로 평수 중에서 선박의 조종운동은 종후동요(surge), 좌우동요(sway), 선수동요(yaw)에 대한 연성운동으로 표현된다. 선박 운동은 Fig. 2에 보이는 좌표계를 이용하여 나타낼 수 있으며, O-x_0y_0는 공간고정 좌표계, G-xy는 선체고정 좌표계를 나타낸다. 무차원화된 조종운동 방정식은 식 (1)과 같이 나타낼 수 있으며, 무차원화는 식 (2)와 같은 방식을 따른다.

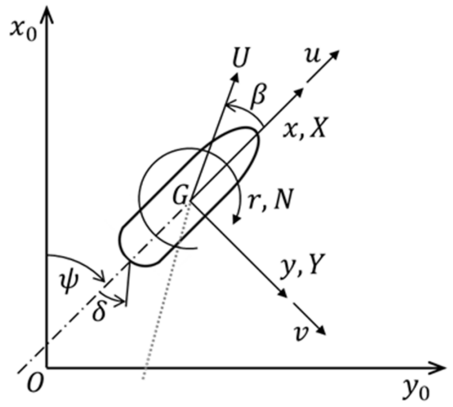

Fig. 2 Coordinate system

$$
\begin{aligned}
&(m' + m_x')\left(\frac{L}{U}\right)\left(\frac{\dot{U}}{U}\cos\beta - \dot{\beta}\sin\beta\right) \\
&\qquad + (m' + m_y')r'\sin\beta = X' \\
&-(m' + m_y')\left(\frac{L}{U}\right)\left(\frac{\dot{U}}{U}\sin\beta + \dot{\beta}\cos\beta\right) \\
&\qquad + (m' + m_x')r'\cos\beta = Y' \\
&(I_{zz}' + i_{zz}')\left(\frac{L}{U}\right)^2\left(\frac{\dot{U}}{L}r' + \frac{U}{L}\dot{r}\right) = N'
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&[m', m_x', m_y'] = [m, m_x, m_y]/\left(\frac{1}{2}\rho L^2 d\right) \\
&[I_{zz}', i_{zz}'] = [I_{zz}, i_{zz}]/\left(\frac{1}{2}\rho L^4 d\right) \\
&[X', Y'] = [X, Y]/\left(\frac{1}{2}\rho L d U^2\right) \\
&N' = N/\left(\frac{1}{2}\rho L^2 d U^2\right) \\
&r' = rL/U
\end{aligned} \tag{2}
$$

위 식에서 m', m_x', m_y'는 선체질량 및 X, Y 방향의 부가질량, I_{ZZ}', i_{ZZ}'는 선체의 관성모멘트 및 부가 관성모멘트, U는 선속, β는 편각, r는 선회각속도, L은 선체 길이, d는 흘수를 나타낸다. 기호 옆에 표기된 홑따옴표(')는 무차원화된 값을 의미한다. 식 (1) 우변에 있는 외력항의 경우 식 (3)과 같이 MMG(manoeuvring mathematical model group)형으로 표현할 수 있으며(Kijima & Nakiri, 2003), 아래첨자 H, P, R은 각각 선체, 프로펠러, 타를 의미한다. 프로펠러에 의한 외력 중, 횡력(Y_P')과 모멘트(N_P')는 다른 유체력에 비해 매우 작으므로 생략하였다.

$$
\begin{aligned}
X' &= X_H' + X_P' + X_R' \\
Y' &= Y_H' \qquad + Y_R' \\
N' &= N_H' \qquad + N_R'
\end{aligned} \tag{3}
$$

2.2 유체력 수학모델

선체에 작용하는 유체력은 식 (4)와 같으며, 프로펠러에 의해 발생하는 종방향 유체력은 식 (5), 방향타에 의한 유체력은 식 (6)과 같다. $X_{\beta r}'$, Y_{β}', N_{β}', Y_r', ... $N_{\beta rr}'$은 유체력 미계수로 Mori(1995)가 제안한 선미선형을 나타내는 파라미터를 사용하여 표현된다(Kijima & Nakiri, 2003).

$$
\begin{aligned}
X_H' &= X_{\beta r}' r' \sin\beta + X_{uu}' \cos^2\beta \\
Y_H' &= Y_{\beta}'\beta + Y_r' r' + Y_{\beta\beta}'\beta|\beta| \\
&\quad + Y_{rr}' r'|r'| + (Y_{\beta\beta r}'\beta + Y_{\beta rr}' r')\beta r' \\
N_H' &= N_{\beta}'\beta + N_r' r' + N_{\beta\beta}'\beta|\beta| \\
&\quad + N_{rr}' r'|r'| + (N_{\beta\beta r}'\beta + N_{\beta rr}' r')\beta r'
\end{aligned} \tag{4}
$$

$$
X_P' = (1 - t_P) n^2 D_P^4 K_T(J_P)/\left(\frac{1}{2} L d U^2\right) \tag{5}
$$

$$
\begin{aligned}
X_R' &= -(1 - t_R) F_N' \sin\delta \\
Y_R' &= -(1 + a_H) F_N' \cos\delta \\
N_R' &= -(x_R' + a_H \bullet x_H') F_N' \cos\delta
\end{aligned} \tag{6}
$$

여기서 t_P는 추력감소계수, n은 프로펠러의 초당 회전수, D_P는 프로펠러 직경, J_P는 전진계수, K_T는 추력 계수, t_R은 타에 의한 저항 증가 보정 계수, F_N'은 단독상태에서 타의 압력, x_R'은 무게중심에서 타축까지 거리, a_H는 선체와 타의 상호 간섭 계수, x_H'는 무게중심에서 a_H의 작용점까지 거리를 나타낸다.

두 선박의 조우상황을 강화학습 모델로 나타낼 때 자선과 타선의 상태(S)가 환경이며, 자선이 에이전트가 된다. 위와 같이 조종운동 방정식과 수학모델을 통해 선박의 운동을 나타냄으로써 반복 구현을 통한 학습이 가능하며, 두 선박의 행동에 따른 산술적 조우 상태를 산출할 수 있다.

3. 마르코프 결정 과정

순차적 행동 결정 문제는 MDP를 통해 수학적으로 나타낼 수 있다. MDP의 구성요소는 상태(states, S), 행동(actions, A), 보상(reward, R), 상태 변환 확률(state transition probability, $P_a(S,S')$), 그리고 감가율(discount factor, γ)이다. 상태 변환 확률은 현재 상태(S)에서 어떤 행동(A)을 할 경우 다음 상태(S')에 도달할 수 있는 확률이다. 자유모델 기반 강화학습 방법에서 상태 변환 확률은 신경망(neural network) 학습에 반영되나(Li, 2017), 본 연구에서 사용하는 DDPG는 자유모델 기반 강화학습 방법이기 때문에 상태 변환 확률은 생략한다. 감가율은 현재의 보상이 나중에 받을 보상보다 얼마나 더 중요한가를 나타내는 상수이며, 상태, 행동, 보상에 대한 정의는 다음 절에서 상세히 설명된다.

3.1 상태

상태(S)는 MDP의 한 요소로 두 선박의 조우 상황에서 에이

전트가 처해 있는 상태를 수치적으로 표현한 집합이다(Fig. 3). Lillicrap et al.(2016)은 DDPG를 제안하면서 아타리(Atari) 게임의 연속된 이미지를 상태 집합의 예시로 활용한다. 하지만 두 선박의 조우상황을 이미지로 타나내면, 각 선박은 윤곽 수준의 단순한 이미지로 나타낼 수 있으며 윤곽은 픽셀 단위로 표현된다. 이미지 크기가 작으면 선박의 현재 운동 상태를 명확히 나타내기 어렵게 되며, 이미지 크기를 키우게 되면 운동 상태가 보다 명확해지지만 상태 집합의 크기도 커지기 때문에 학습속도가 저하된다. 본 논문에서는 이미지를 활용하는 것 보다 상태 요소를 파라미터로 하여 활용하는 것이 유리하다 판단하였으며, 식 (7)과 같이 상태 집합을 정의하였다.

$$S=\{v_{i1}, v_{j1}, r_1, l, \psi_s, \delta, i_{rel}, j_{rel}, v_{i\_rel}, v_{j\_rel}\} \tag{7}$$

여기서 아래첨자 i와 j는 목표 경로 방향과 목표 경로에 수직한 방향을 의미한다. v_{i1}, v_{j1}은 자선의 i, j 방향 속도, r_1은 자선의 선회 각속도, l 은 목표 경로에서 자선까지 수직하게 떨어진 거리, ψ_s는 목표 경로와 선수방향(heading)이 이루는 각도, δ는 타각, i_{rel}, j_{rel}은 i와 j 방향으로 타선의 상대 거리, v_{i_rel}, v_{j_rel}은 i와 j 방향으로 타선의 상대속도($v_{i2}-v_{i1}$, $v_{j2}-v_{j1}$)이다(Fig 3). 타선의 정보는 선박자동식별장치(automation information system, AIS)로부터 획득하는 것으로 가정한다.

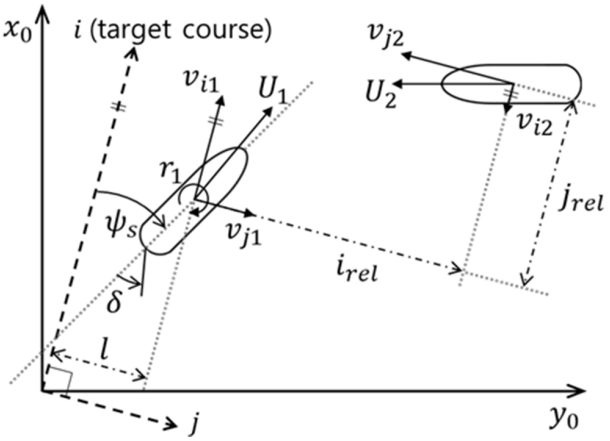

Fig. 3 State of agent at encounter situation

3.2 행동

에이전트가 할 수 있는 행동(A)은 DDPG 이전 강화학습 방법들과 달리 이산적으로 표현하지 않고 연속적 행동 공간으로 정의한다. 본 연구에서는 충돌 가능성이 있는 선박이 나타났을 때 Fig. 4와 같이 목표 경로에 평행한 새로운 회피 경로(avoidance course)를 추가하고, 식 (8)과 같이 목표 경로에서 회피 경로까지 수직한 방향의 거리(a)를 행동 집합의 원소로 정의하였다. 범위인 $a_{\max}$는 목표 경로에서 회피 경로까지의 최대 거리로 20L로 설정한다.

$$A=\{a\}, \quad -a_{\max} \le a \le a_{\max} \tag{8}$$

자동 경로 추종 제어기는 목표 경로를 설정하였을 때 배가 목표 경로를 따라 운항하도록 타를 조종한다. 자동 경로 추종 제어기 구축에는 PID(proportional-integral-derivative)제어와 퍼지제어 등을 이용한 방법이 연구되어 왔으며, 본 연구에서는 Furukawa et al.(2004)의 퍼지를 이용한 자동 경로 추종 제어기를 활용하였다.

Fig. 4와 같이 타선이 나타날 경우 목표 경로를 회피 경로로 대체하고, 자동 경로 추종 제어기를 통해 자선이 회피 경로를 향하도록 한다. 자선과 타선 사이에 충돌 가능성이 없을 때에는 회피 경로를 목표 경로와 일치시킴으로써 자선이 목표 경로를 유지할 수 있다. 강화학습을 통해 타선을 회피하는 것이 충돌하는 것보다 높은 보상을 받는다는 것을 학습하면, 충돌 가능성이 있는 상태(S)일 경우 회피 경로를 이동하여 목표 경로를 벗어나도록 변침한다. 회피 후, 회피 경로를 목표 경로에 가깝게 이동하는 것이 높은 보상을 받는다는 것을 학습하면 자선은 목표 경로로 복귀하게 된다.

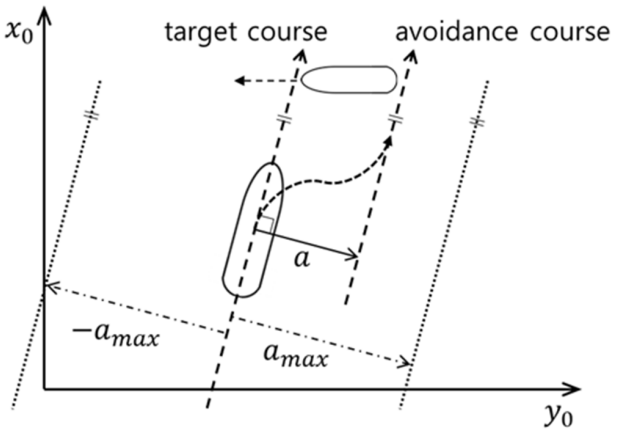

Fig. 4 Action of agent

3.3 보상

Fig. 1에 도시되었듯이 에이전트는 현재 처한 상태(S_t)에서 어떤 행동(A_t)을 했을 때, 다음 상태(S_{t+1})가 되면서 환경으로부터 보상(R_{t+1})을 받는다. 보상을 최대로 받거나 받게 될 행동을 학습하는 것이 강화학습의 목표이기 때문에 보상은 강화학습 적용에서 가장 중요한 요소이다. 본 연구에서는 두 선박의 조우 상황에서 경제성과 안전성을 고려한 최적의 회피 경로를 찾도록 한다.

$$R=R_e+R_s \tag{9}$$

$$R_e=-|l|/a_{\max}+c_e\times l/a_{\max}+U \qquad \begin{cases}\text{if}\,(l<0), & c_e=1\\ \text{else}, & c_e=0\end{cases} \tag{10}$$

$$R_s=-w_s\times c_s \qquad \begin{cases}\text{if}\,(collision), & c_s=1\\ \text{else}, & c_s=0\end{cases} \tag{11}$$

식 (9)와 같이 보상(R)은 경제성과 안정성에 관련된 두 가지 보상(R_e, R_s)으로 나눌 수 있다. 운항 경제성은 항해 거리를 줄

이고 항해 속도를 유지하는 것과 밀접한 관계가 있다. 충돌 회피 상황에서 자선이 목표 경로에 수직한 거리(l)만큼 운항 거리는 늘어나며, 변침에 의해 속도(U)가 줄어든 만큼 운항 시간이 늘어나기 때문에 경제성과 관련된 보상(R_e)은 목표 경로에 가까이 위치하고 속도를 유지하면 높은 보상을 받을 수 있도록 한다.

식 (10)에서 첫 번째 항은 목표 경로에서 벗어날수록 목표 경로에 수직한 거리(l)에 비례하여 음의 보상을 받게 하며, 자선이 목표 경로에 가까워지도록 학습한다. 두 번째 항은 c_e를 조건 변수로 두어 자선이 좌현 변침을 통해 목표 경로의 좌측에 위치할 경우($l < 0$) 음의 보상 받게 하며, 타선을 좌현에 두고 자선이 우현으로 선회할 수 있도록 한다. 세 번째 항은 자선의 속도(U)를 양의 보상으로 주어 선회 시에 자선의 속도가 느려지지 않는 방향으로 학습한다.

안전성과 관련된 보상(R_s)은 식 (11)과 같다. w_s는 충돌에 대한 가중치 상수로 그 값을 5로 두었으며, c_s는 충돌 유무에 대한 조건 변수로 두 선박이 충돌할 경우 1, 그렇지 않으면 0 이다. 실제 충돌은 두 선박이 부딪힌 상태지만 본 논문에서는 블록영역(blocking area)을 설정하고 회피 거리의 안전성을 확보하기 위해 블록영역 안에 타선의 선체 윤곽이 들어오면 충돌 하였다고 간주한다. 블록영역은 타선의 감시영역(watching area)이 블록영역을 침범했을 때 자선이 현재 경로를 변경 또는 유지 할 것인지 여부를 결정하는 영역으로 Kijima and Furukawa(2003)에서 식 (12)와 같이 사용하였다.

$$
\begin{aligned}
R_{bf} &= L + (1+s)\,T_{90}\,U \\
R_{ba} &= L + T_{90}\,U \\
S_b &= B + (1+t)\,D_T
\end{aligned}
\tag{12}
$$

여기서 T_{90}은 선수방향이 0도에서 90도까지 회전하는데 걸리는 시간, D_T는 선회지름(tactical diameter), s, t는 조우 상황을 고려한 계수이다. 조우 각도 및 상대 속도에 따라 s와 t 계수의 값이 변하기 때문에 블록영역의 크기 역시 달라진다(Kijima & Furukawa, 2003). Fig. 5는 90도 각도 교차 조우 상황에서 블록영역을 나타내고 있다. 이때 블록영역은 R_{bf}, R_{ba}, S_b 파라미터에 의해 생성된 두 타원의 조합으로 만들어진다.

강화학습 과정 중 두 선박이 충돌 할 경우, 안전성 보상(R_s)은 경제성 보상(R_c)에 비해 상당히 큰 음의 값을 갖기 때문에 자선은 타선과 충돌하지 않는 범위에서 경제성 보상을 최대화하는 회피 동작을 하게 된다. 즉, 자선은 타선이 블록영역에 들어오지 않으면서 최대한 블록영역에 가깝게 거리를 유지하며 둘러가도록 학습한다.

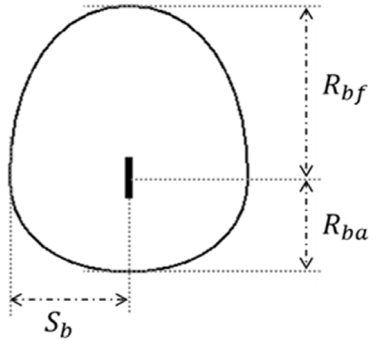

Fig. 5 Blocking area

4. 심층 결정론적 정책 경사법 적용

두 선박의 충돌 조우상황에 대한 순차적 행동문제 정의를 DDPG 방법에 적용한다. DDPG는 엑터-크리틱에서 엑터 네트워크를 업데이트 할 때 DPG 방법을 사용하며 재생 버퍼를 활용하는 특징을 갖고 있다.

4.1 탐험

보상을 최대화 할 수 있는 행동을 찾기 위해 무작위 행동을 수행하는 탐험(exploration)이 DDPG 과정에 필요하다. 본 연구에서는 Lillicrap et al.(2016)이 제안한 Ornstein-Uhlenbeck 방법을 사용하여 무작위 노이즈(noise) 값을 생성하고, 식 (13)과 같이 엑터 네트워크에서 산출되는 행동(a)에 노이즈를 더해줌으로써 탐험이 가능하도록 한다. 또한 감소하는 엡실론-탐욕(epsilon-greedy)정책(Lee et al., 2017)을 활용하여 에피소드를 거듭할수록 노이즈의 영향을 줄인다. 엡실론(ε)은 한 에피소드를 수행할 때마다 식 (14)와 같이 업데이트된다. 엡실론이 ε_{min}보다 클 경우 ε_{decay}만큼 감소하고, 그렇지 않으면 ε_{min} 값이 된다. 초기 엡실론과 ε_{decay}, ε_{min}은 각각 1.0, 5.0E-4, 2.0E-4로 설정되며, 이는 2,000에피소드가 지나면 탐험을 거의 하지 않도록 하는 값이다.

$$
a = a + N\epsilon \tag{13}
$$

$$
\epsilon = \begin{cases} \epsilon - \epsilon_{decay} & \text{if}, \epsilon > \epsilon_{\min} \\ \epsilon_{\min} & otherwise, \end{cases} \tag{14}
$$

4.2 시나리오

DDPG 시나리오는 15노트 속도의 두 선박이 90도 각도에서 교차 조우하는 상황이며, 시나리오의 시작 조건은 Table 1과 같다. 본 연구에서는 선박해양플랜트연구소 공개 선형인 KVLCC의 주요 제원을 Table 2와 같이 축척하여 활용하였으며, 유체력 미계수는 Kijima and Nakiri(2003)의 추정식을 통해 산출된 Table 3의 값을 사용하였다.

Table 1 Initial conditions for encounter scenario

Item	Own ship	Target ship
x_0[m]	-40.0	0.0
y_0[m]	0.0	40.0
ψ_0[deg]	0.0	-90.0
U_0[m/s]	0.6821	0.6821

Table 2 Principal dimensions of KVLCC

Item	Value
Length[m]	2.5
Breadth[m]	0.4531
Draft[m]	0.1625
Block Coefficient	0.8101

Table 3 Hydrodynamic derivative values

Item	Value	Item	Value
X'_{uu}	-0.02050873	$X'_{\beta r}$	-0.032517508
Y'_{β}	0.36208424	N'_{β}	0.125635713
Y'_{r}	0.113776326	N'_{r}	-0.0533008836
$Y'_{\beta\beta}$	0.843024015	$N'_{\beta\beta}$	0.0141946375
Y'_{rr}	0.0761068165	N'_{rr}	-0.0286193416
$Y'_{\beta\beta r}$	-0.362524033	$N'_{\beta\beta r}$	-0.191315025
$Y'_{\beta rr}$	0.423484325	$N'_{\beta rr}$	-0.0610068738

4.3 하이퍼파라미터 최적화

심층 결정론적 정책 경사법은 신경망을 활용하고 있기 때문에 하이퍼파라미터(hyper parameter) 최적화가 중요하다. 최적화에 활용되는 기법 중 그리드탐색법(grid search)과 같은 규칙적인 탐색 보다는 무작위 샘플링을 통해 탐색하는 것이 좋은 결과를 낸다고 알려져 있다(Bergstra & Bengio, 2012). 본 연구에서는 무작위 탐색법을 활용하여 Table 4와 같이 하이퍼파라미터 값을 최적화하였다.

하이퍼파라미터에서 아래첨자 a와 c는 각각 엑터 네트워크와 크리틱 네트워크를 의미한다. L_a, L_c는 각 네트워크에서 은닉층(hidden layer)의 개수, n_a, n_c는 각 네트워크에서 각 은닉층의 유닛(unit) 개수, α_a, α_c는 각 네트워크의 학습률, γ는 감가율, τ는 각 네트워크를 천천히 갱신하기 위한 상수로 학습의 안전성을 높여준다. B는 재생 버퍼의 크기, r은 배치 크기로 한번 학습할 때 사용되는 샘플의 수이며, w_s는 충돌 가중치로 식 (11)의 보상 식에 사용되는 상수이다.

Table 4 Hyper parameters

Item	Value
L_a (number of actor hidden layers)	2
n_a (number of actor hidden units)	500
α_a (learning rate of actor)	1.4917255E-5
L_c (number of critic hidden layers)	4
n_c (number of critic hidden units)	500
α_c (learning rate of critic)	1.7392787E-4
γ (discount factor)	1 - 2.0815591E-7
τ (tau)	8.2819554E-5
B (buffer size)	16,400
r (batch size)	32
w_s(collision weight)	5.0

4.4 학습

최적화된 하이퍼파라미터로 시나리오를 학습하였을 때, 8,000 에피소드 동안 받은 보상은 Fig. 6과 같다. 보상을 최대화하는 방향으로 학습되었으며, 약 4,000 에피소드부터 보상 값이 수렴하는 것을 확인할 수 있다.

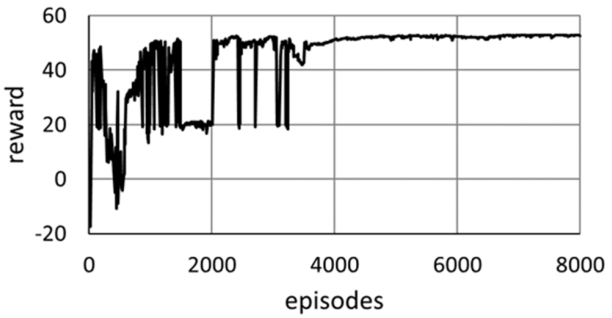

Fig. 6 History of achieved rewards through training

5. 조우상황 조종운동 예제

DDPG 방법을 이용하여 최적 정책을 찾는 학습 결과는 Fig. 7과 같다. Fig. 7의 위 그림에서 굵은 실선은 자선의 경로이며 점선은 타선의 경로를 나타낸다. 특정 시간에 자선의 블록영역들을 도식화였는데, 십자 형상의 표식은 같은 시간대에서 자선과 타선의 위치를 나타낸다. 전체 결과를 보면 약 100초까지 목표 경로를 유지하다 타선을 회피하기 위한 행동을 시작한다. 회피 행동에 의해 선수 방향이 우현으로 향하도록 조타를 시작하며, 그로 인해 자선의 속도가 줄어들기 시작한다. 약 115초에 타가 복원되면서 자선은 선수 방향을 유지하며 속도가 상승되지만, 타선을 회피하였다고 판단한 시점인 약 130초부터 다시 목표 경로로 향하는 행동을 시작한다. 조타에 의해 선수는 좌현 방향으로 향하고 속도는 다시 줄어들며, 약 150초부터 목표 경로에 진입하기 위해 타가 복원되면서 속도가 상승하지만 회두모멘트의 영향이 남아 있기 때문에 선수 방향은 여전히 좌현 쪽으로 향한다. 약 170초 구간은 목표 경로를 안정적으로 추종하기 위해 타를 사용하고 오버슈트 없이 약 200초 구간부터 목표 경로와 선수 방향을 일치시켜 15노트까지 속도를 회복하게 된다. 이 결과는 횡단상태에서 타선의 홍등을 바라보고 있는 선박이 우현측로 진로를 피하여야 한다는 규칙을 준수하고 있다.

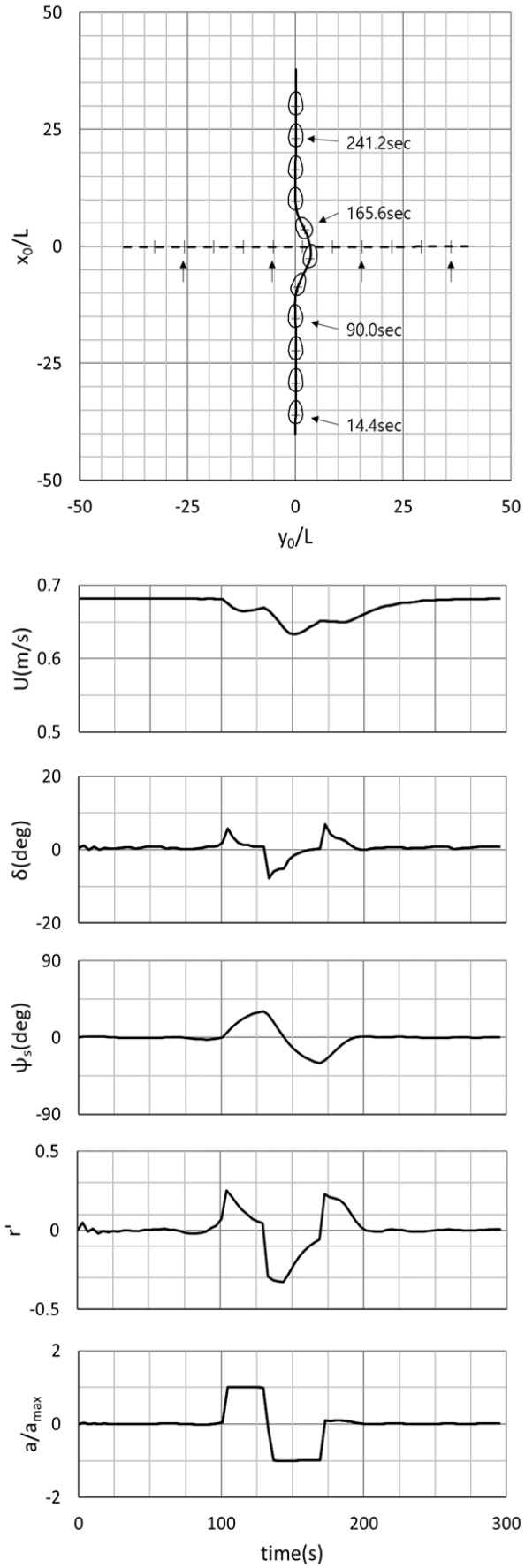

Fig. 7 Results of reinforcement learning

6. 결 론

두 선박의 조우상황에서 경제성과 안전성을 고려하여 회피 동작을 최적화하는 강화학습 방법을 개발하였다. 우선 조종운동 수학모델을 기반으로 강화학습 환경을 구축하고, 조종운동수학 모델은 MMG에서 제안한 모델을 활용하였다. 선박의 조우상황에서 충돌 회피를 위한 동작은 순차적 행동 문제로 간주하고, MDP를 이용해 수학적으로 표현하였다. 목표 경로에 수평한 새로운 회피 경로를 생성하고 목표 경로와 회피 경로 사이의 거리를 행동으로 설정한 후, 두 선박의 조우 상태에 따른 최적의 행동 정책, 즉 최적의 회피 경로를 찾았다. 보상함수는 경제성과 안전성을 동시에 고려하여 실제 선박 운항 조건과 가깝게 설정하였다. 마지막으로 두 선박의 90도 교차 조우상황을 테스트 하여 제안된 방법의 유효성을 검증하였다.

본 연구에서는 90도 교차 조우상황만을 가정하여 테스트하였으나 보다 포괄적인 조우 상황을 고려하기 위하여 향후 마주침(head-on) 및 추월(overtaking) 상황에 대한 테스트를 진행할 계획이다. 또한 조타만을 이용하여 피항이 가능한 상황에 대해 테스트를 하였는데 향후 행동 집합에 추력 조절 요소를 추가함으로써 속도에 대한 영향을 포함하는 연구가 진행되어야 할 것으로 사료된다.

후 기

본 연구는 2018년도 한국연구재단의 이공학 개인기초연구지원사업 (NRF-2017R1D1A3B03030423)의 지원으로 수행된 연구임을 밝히며, 연구비 지원에 감사드립니다. 아울러 산업통상자원부 조선해양산업핵심기술개발사업(조선소 생산 관리 정밀도 향상을 위한 리드타임 기준 정보 체계 개발)의 재정지원에도 감사드립니다.

References

Bergstra, J. & Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, pp.281-305.

Furukawa, Y., Kijima, K. & Ibaragi, H., 2004. Development of automatic course modification system using fuzzy inference. *International Federation of Automatic Control Proceedings*, 37(10), pp.77-82.

Jeong, J.S., 2015. Korean e-Navigation goal and government plan. *Telecommunications Technology Association Journal*, 159, pp.20-27.

Kijima, K. & Furukawa, Y., 2002. Development of collision avoidance algorithm using fuzzy inference. *Proceedings of ISOPE Pacific/Asia Offshore Mechanics Symposium*,

pp.123-130.

Kijima, K. & Furukawa, Y., 2003. Automatic collision avoidance system using the concept of blocking area. *International Federation of Automatic Control Proceedings*, 36(21), pp.223-228.

Kijima, K. & Nakiri, Y., 2003. On the practical prediction method for ship manoeuvring characteristics. *Transaction of the West-Japan Society of Naval Architects*, 105, pp.21-31.

Kim, D.J. & Kwak, S.Y., 2011. Evaluation of human factors in ship accidents in the domestic sea. *Journal of the Ergonomics Society of Korea*, 30(1), pp.87-98.

Korean Maritime Safety Tribunal, 2017, *Current situation of causes of maritime accidents by type of accident* [online] Available at: https://www.kmst.go.kr/kmst/statistics/annualReport/selectAnnualReportList.do [Accessed 21 May 2018].

Kose, K., Hirono, K., Sugano, K. & Sato, I., 1998. A new collision-avoidance-supporting-system and its application to coastal-cargo-ship "SHOYO MARU". *IFAC Proceeding*, 31, 263-268.

Lee, H.J. & Rhee, K.P., 2001. Development of collision avoidance system by using expert system and search algorithm. *International Shipbuilding Progress*, 48, pp.197-212.

Lee, W.W., Yang, H.R., Kim, K.W., Lee, Y.M. & Lee, U.R., 2017. *Reinforcement Learning with Python and Keras*. Wikibook.

Li, Y., 2017. Deep Reinforcement Learning: An Overview. arXiv preprint arXiv:1701.07274.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. & Wierstra, D., 2016. Continuous control with deep reinforcement learning. *International Conference on Learning Representations*, 1509.02971.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M., 2013. Playing Atari with Deep Reinforcement Learning. *Neural Information Processing Systems*, Lake Tahoe, USA, 9 December 2013.

Mori, S., 1995. Note of Ship Form Design(24). *FUNE-NO-KAGAKU*, 48, pp.40-49.

Ota, D., Masuyama, T., Furukawa, Y. & Ibaragi, H., 2016. Development of automatic collision avoidance system for ships using reinforcement learning. *Proceedings of 7th PAAMES and AMEC2016*, Hong Kong, 13-14 October 2016.

Shim, W.S., Park, J.W. & Lim, Y.K., 2010. The study on the trend of international standards and the domestic plan to cope with e-navigation. *Journal of the Korea Institute of Information and Communication Engineering*, 14(5), pp.1057-1063.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Graepel, T., Lillicrap T., Leach, M., Kavukcuoglu, K. & Hassabis, D., 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), pp.484-489.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. & Riedmiller, M., 2014. Deterministic policy gradient algorithms. *International Conference on Machine Learning*, 32, pp.387-395.

Son, N.S., Furukawa, Y., Kim, S.Y. & Kijima, K., 2009. Study on the collision avoidance algorithm against multiple traffic ships using changeable action space searching method. *Journal of the Korean Society for Marine Environmental Engineering*, 12(1), pp.15-22.

Van, S.H., 2007. *Planning research for development of core technologies for smart ship*. KORDI Report No. UCPM0147A-42-7.


김 동 함


이 성 욱


남 종 호


Furukawa, Y.