Journal of the KNST

2022; Vol.5, No.1; pp. 38-46

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2022.03.5.1.38

Received: 2022/01/03
Revised: 2022/02/08
Accepted: 2022/03/22
Published: 2022/03/31

*Corresponding Author:
Han-Seok Park
Department of Aerospace Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea
E-mail: intoto6@kaist.ac.kr

Abstract

본 논문에서는 해상에서 UAV의 함상착륙을 위한 접근 및 착륙 과정에서 강화학습에 기반한 경로계획 기법을 제안한다. 제안된 모델은 연속적인 행동영역에 적용이 가능한 대표적인 강화학습 알고리즘인 심층 결정론적 정책 기울기(DDPG) 방식을 이용하여, 이동 장애물에 대한 회피와 함정의 3축 운동을 고려한 경로추종문제를 다루었다. 본 연구의 결과는 시뮬레이션을 통해 보여진다.

This paper proposes a path planning model based on reinforcement learning in the process of approaching and landing for shipboard landing of UAV in maritime. The proposed model dealt with the path tracking problem considering avoidance of moving obstacles and triaxial motion of ships, using the deep deterministic policy gradient (DDPG) method, a representative reinforcement learning algorithm applicable to continuous behavioral areas. The results of this study are shown through simulation.

Keywords

무인 항공기(Unmanned Aerial Vehicles),
함상착륙(Shipboard Landing),
강화학습(Reinforcement Learning),
경로계획(Path Planning)

강화학습 기반 해상에서 UAV의 함상 자율착륙을 위한 경로계획

Reinforcement Learning Based Path Planning for Autonomous Shipboard Landing of UAV in Maritime

박한석[1*], 방효충[2]

[1]해군 대위/KAIST 항공우주공학과 석사과정
[2]KAIST 항공우주공학과 교수

Hanseok Park[1*], Hyochoong Bang[2]

[1]Lieutenant, ROK Navy/M.S. course student, Dept. of Aerospace Engineering, Korea Advanced Institute of Science and Technology
[2]Professor, Dept. of Aerospace Engineering, Korea Advanced Institute of Science and Technology

1. 서론

오늘날 무인항공기(UAV)는 산업, 재난 구조, 측량 및 지도 제작, 군사 정찰 작전과 같은 다양한 분야에 광범위하게 적용 중이며, 적용 목적 및 범위에 따라 민간용(civilian)과 군사용(military)이라는 두 개의 큰 범주로 구분할 수 있다. 군사용 목적의 UAV는 1차 세계대전 시 최초 도입된 이후 2차 세계대전, 베트남 전쟁, 코소보 전쟁 등에서 중추적인 역할을 수행하였으며, UAV가 지닌 유연성(flexibility), 적시성(timeliness), 저비용성(low cost), 저위험성(low risk), 정찰 능력 및 넓은 임무범위(strong monitor capability and widespread coverage) 등의 특징으로 인해 군사 작전 분야에서의 활용성이 높다[1]. 특히 육상 대비 해상의 관할 구역이 넓으나, 해상 전력이 부족해왔다는 점과 육상 대비 해상의 환경이 단순하다는 점 등 한반도의 지리적 특성을 고려하면, 해상 작전에서의 UAV의 활용은 매우 효과적일 것으로 판단된다. 따라서, 해상에서 UAV의 활용은 Table 1과 같이 해상초계, 긴급상황 대응 및 수색/구조 임무 등에서 작전적 이점이 있다[2].

최근 미국은 군사용 UAV로 정찰과 타격 등 다양한 용도로 활용이 가능한 저비용의 소형 UAV 투입을 확대하고 있다[3]. 대표적인 소형 UAV인 멀티 로터 형태는 소형/경량의 특징으로 인해 선박이라는 한정된 공간

에서 활용도가 매우 높을 것으로 판단되며, 함정 탑재하 운용을 통해 작전성과의 증대가 기대된다.

그러나 항공기의 중대 사고는 이/착륙단계에서 많이 발생하는 것으로 널리 알려져 있는데, 특히 해상에서의 함상 착륙은 파도에 의한 함정의 모션 등 많은 불확실성을 내재하고 있으므로, 사고가 야기될 가능성이 상대적으로 더 높다고 할 수 있다. 위와 같은 이유로 인해 함정에서 다수의 소형 UAV를 운용하기 위해서는 고도의 비행 기술을 갖춘 다수의 숙련된 조종사가 필요하나, 숙련된 조종사를 양성하기 위한 시간과 비용을 고려하면, 다수의 전력을 값싸게 운용할 수 있다는 이점에 부합하지 않게 된다.

이를 극복하면서 작전 효율성을 최대화하기 위해서는 UAV에 인간의 개입을 최소화할 수 있는 높은 자율성의 부여가 요구되며, 높은 자율성의 개념은 Table 2와 같다[4].

최근 무인전력의 자율성을 확보하기 위한 방안으로 강화학습 등의 머신러닝 기법을 활용한 많은 연구가 이루어지고 있다. 본 연구에서는 UAV의 함상착륙을 위한 접근 및 착륙 과정에서 강화학습 기반의 경로계획 기법을 제안하고자 하며, 기존 선행연구와 비교했을 때 크게 두 가지 차이점이 있다.

첫째, 기존의 강화학습 기반의 이동표적에 대한 착륙경로 계획과 관련된 연구는 2차원 공간에서의 연구가 대부분을 차지하였다[5]. 3차원 공간상에서의 연구라 하더라도 수직축방향으로는 착륙대의 모션이 없음을 가정한 반면[6], 본 연구에서는 함상 착륙을 위해 수직 방향의 함정 heave motion을 고려한 연구를 수행하였다. 둘째, 해상에서 미식별 항공표적의 제3국 군함에 대한 근거리접근은 국제적 분쟁을 야기할 가능성이 있으며[7], 이를 고려하여 이동 장애물에 대한 회피문제를 다루었다. 따라서, 기존의 강화학습 기반의 경로계획 문제에서 장애물 회피를 고려함에 있어 고정된 장애물을 적용한 반면[8], 본 연구에서는 해상이라는 환경에서 회피가 요구되는 물체를 특정 선박으로 가정하여 이동 장애물 회피를 위한 연구를 수행하였다.

본 연구의 구성은 다음과 같다. 2장에서는 강화학습의 이론적 배경을 설명하고, 3장은 2장의 이론적 배경을 기반으로 접근단계와 착륙단계로 구분하여 연구 방법 및 시뮬레이션 결과를 서술하였다. 마지막으로 4장에서는 결론과 향후 과제를 제시한다.

Table 1. 해상에서 UAV 운용 시 이점

구분	함정 단독 운용시	UAV 복합 운용시
해상초계 및 긴급상황 대응	생존성 보장을 위해 미식별 구역/표적에 대한 근거리 접근 제한	함정의 접근이 제한되는 구역 및 표적에 대한 대응 가능
수색/구조	조난 위치에 도달까지 장시간 소요	UAV를 이용한 구조물품 투입을 통해 골든 타임이 지나기 전 선제적 작전 가능

Table 2. 자율화 수준

구분	적용개념
Remote control	Man in the loop, 인간의 개입 없이 운용 불가
Automatic control	Man on the loop, 사전 계획된 프로그램에 따라 작동하는 시스템으로, 인간은 모니터링 역할 수행
Autonomous control	Man off the loop, 스스로 상황을 판단하고 임무를 수행하는 시스템으로, 인간은 임무 할당 역할만 수행

2. 이론적 배경

2.1 강화학습 개요

강화학습(reinforcement learning)은 기계학습의 한 분야로 특정 환경에 대해 정의된 현재 상태를 기반으로 사전 정의된 보상을 최대화할 수 있는 행동을 선택하는 기법이다. 주요 구성요소로 환경(environment)과 에이전트(agent)가 있으며, 에이전트는 특정 환경에서 상태(state)에 따른 행동(action)을 결정하고 환경은 그 결정에 대한 보상(reward)을 내린다. 강화학습이 적용 가능한 문제는 MDP(Markov decision process)로 표현되며, MDP 문제에서의 상태 전이 확률(state transition probability) 및 보상은 확률적인(probabilistic) 값 또는 결정론적인(deterministic) 값 두 가지로 정의할 수 있다. 일반적인 강화학습의 구조는 Fig. 1과 같다.

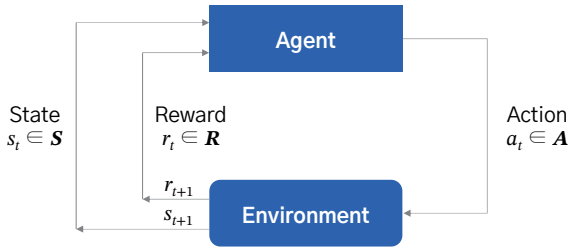


Fig. 1. 강화학습 구조

강화학습은 MDP에 학습의 개념을 추가한 것이며, MDP는 Markov property를 기반으로 한 의사결정모델로서 상태 집합 S, 행동 집합 A, 상태 전이 확률 P, 보상 함수 R, 할인요인(discount factor) γ로 구성된다. 이때, 상태 전이 확률 P는 특정 상태에서 행동을 수행하여 다른 상태로 이동할 확률을 뜻하며, 수식은 다음과 같다.

$$P_{ss'}^{a} = P(S_{t+1} = s' \mid S_t = s,\ A_t = a) \tag{1}$$

리턴(return)은 γ값을 적용하여 각 시점에서의 보상들을 현재가치로 환산하여 합한 누적 보상으로서 G_t로 표현되며, 강화학습의 목적은 리턴을 최대화하는 정책함수(policy, 특정 상태에서 어떤 행동을 할지를 정하는 매핑함수)를 찾는 것으로, 수식은 다음과 같다.

· Return

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{2}$$

· Policy

$$\pi(a \mid s) = P(A_t = a \mid S_t = s) \tag{3}$$

한편, 상태의 가치를 표현하는 함수로서 특정 상태에서 얻을 수 있는 리턴의 기댓값을 상태가치 함수(state value function) $V_\pi(s)$라 하며, 특정 상태에서 행동을 취했을 때 얻게 되는 리턴의 기댓값을 행동가치 함수(action value function) $q_\pi(s,a)$로 정의한다. 각각의 수식은 다음과 같다.

· State value function

$$V_\pi(s) = E_\pi(G_t \mid S_t = s) \tag{4}$$

· Action value function

$$q_\pi(s,a) = E_\pi(G_t \mid S_t = s,\ A_t = a) \tag{5}$$

2.2 벨만 방정식

상태가치 함수와 행동가치 함수들의 관계로 현재 상태/행동과 다음 상태/행동과의 관계식이 만들어지는데 이를 벨만 방정식이라 한다. 벨만 기대 방정식을 변형하고, 현재와 바로 다음 상태/행동 간의 관계가 드러나도록 정리를 하여 다음 식을 얻을 수 있다.

· State value function

$$V_\pi(s) = \sum_{a \in \boldsymbol{A}} \pi(a \mid s)\left(R_s^a + \gamma \sum P_{ss'}^a V_\pi(s')\right) \tag{6}$$

· Action value function

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in \boldsymbol{S}} P_{ss'}^a \sum_{a' \in \boldsymbol{A}} \pi(a' \mid s')\, q_\pi(s',a') \tag{7}$$

한편, 위에서 언급한 바와 같이 강화학습에서 추구하고자 하는 목표는 최대의 보상을 얻는 정책함수를 찾는 것이며, 강화학습의 목표에 따라 찾아진 정책함수를 최적 정책(optimal policy)이라 부른다. 이러한 최적 정책을 따르는 벨만 방정식이 벨만 최적 방정식이며, 최적 가치에 대한 함수(optimal value function)는 최대의 보상을 갖는 가치함수로 볼 수 있으므로 다음과 같이 표현 가능하다.

$$\begin{aligned} V_*(s) &= \max_\pi V_\pi(s) \\ q_*(s,a) &= \max_\pi q_\pi(s,a) \end{aligned} \tag{8}$$

이 중 행동가치 함수에 대한 최적을 구하게 되면, 주어진 상태에서 행동가치가 가장 높은 행동을 선택할 수 있게 되며, 이를 통해 최적 정책을 구할 수 있게 된다. 이를 수식으로 표현하면 다음과 같다.

· Optimal policy

$$\pi_*(a \mid s) = \begin{cases} 1 & \text{if } a = \arg max_a\, q_*(s,a) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

· Optimal state value function

$$V_*(s) = \max_a \left[R_s^a + \gamma \sum_{s' \in \boldsymbol{S}} P_{ss'}^a V_*(s')\right] \tag{10}$$

· Optimal action value function

$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in \boldsymbol{S}} P_{ss'}^a \max_{a'} q_*(s',a') \tag{11}$$

2.3 DDPG(deep deterministic policy gradient)

강화학습을 구분하는 중요 요소 중 하나는 모델에 대한

기반 여부이다. 강화학습에서는 환경에 대한 가정을 모델이라 하며, 보상 함수와 상태 전이 확률을 포함하는 모든 MDP 정보를 아는 상태에서 학습하는 접근법을 모델 기반(model-based) 또는 플래닝(planning) 이라고 하며, 모르는 것으로 간주하여 학습하는 것을 모델-프리(model-free) 접근법이라 한다. 실제 세계에서는 MDP에 대한 정보를 모르는 상황이 더 많기 때문에 대부분의 강화학습은 모델-프리로 진행된다. 본 연구에서는 모델-프리 기법을 사용하였으며, 그중 Q-learning에 기반한 DDPG(deep deterministic policy gradient)를 다음과 같은 특징을 고려하여 활용하였다.

DDPG는 불연속적인 행동 영역에 대한 적용만이 가능하여 연속적이거나 매우 넓은 행동 영역을 가지는 문제에 대해서는 적용하지 못하는 DQN의 문제를 해결하고자 제안되었다[9]. DDPG 알고리즘은 행동가치 함수를 파라미터화하여 근사하는 크리틱(critic) 신경망과 정책함수를 파라미터화하여 근사하는 액터(actor) 신경망으로 구성된다. 이를 액터-크리틱 프레임워크라 하며, DDPG는 DQN의 replay memory 및 target network 개념과 액터-크리틱 프레임워크가 결합한 형태로, Fig. 2와 같은 구조를 갖는다[10]. 즉, DQN과 DPG의 개념을 함께 사용하여 액터 및 크리틱의 근사 함수를 신경망으로 대체하였으며, 이를 통해 앞서 언급한 바와 같이 DQN과 달리 연속적인 행동 영역에 대한 설계가 가능한 특징을 갖고 있다. 본 연구에서는 이러한 특징을 고려하여 DDPG를 이용한 연구를 수행하였다.

3. 연구 방법 및 시뮬레이션 결과

본 연구에서는 UAV가 임무 종료 후 원거리에서 모함으로 접근하는 접근단계(approaching phase), 함정에 착륙하는 착륙단계(landing phase)로 구분하며, 접근단계에서는 이동 장애물에 대한 회피 하에 사전 지정된 지점까지 접근, 착륙단계에서는 이동 중인 함정의 heave motion을 고려한 함상착륙을 목표로 한다.

3.1 UAV 모델

강화학습에 기반한 경로계획 기법 연구를 우선으로 하여 본 연구에서는 UAV 모델을 점 질량으로 가정하며, 3차

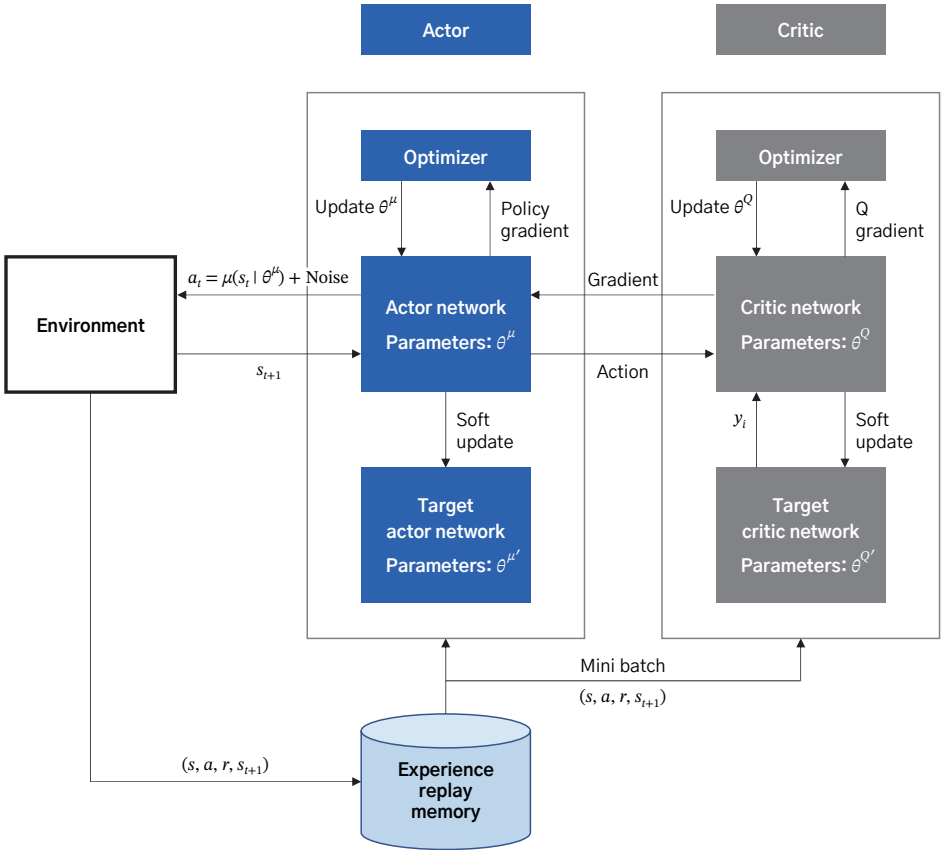


Fig. 2. DDPG 알고리즘 구조

원 관성 프레임에 대한 UAV의 위치, 속도 및 가속도를 다음과 같이 표현한다.

$$\begin{aligned}\boldsymbol{p}(t) &= [x(t) \quad y(t) \quad z(t)]^T \\ \boldsymbol{v}(t) &= [\dot{x}(t) \quad \dot{y}(t) \quad \dot{z}(t)]^T \\ \boldsymbol{a}(t) &= [\ddot{x}(t) \quad \ddot{y}(t) \quad \ddot{z}(t)]^T\end{aligned} \tag{12}$$

접근단계에서는 속도, 착륙단계에서는 가속도를 제어입력으로 사용하며, UAV의 성능상 제한치를 고려하여 다음과 같이 제약조건을 설정한다.

$$\begin{aligned}&\boldsymbol{v}_{\min} \le \boldsymbol{v}(t) \le \boldsymbol{v}_{\max} \\ &\text{where} \quad \boldsymbol{v}_{\min} = [v_{x,\min} \quad v_{y,\min} \quad v_{z,\min}]^T, \\ &\qquad\quad \boldsymbol{v}_{\max} = [v_{x,\max} \quad v_{y,\max} \quad v_{z,\max}]^T \\ &\boldsymbol{a}_{\min} \le \boldsymbol{a}(t) \le \boldsymbol{a}_{\max} \\ &\text{where} \quad \boldsymbol{a}_{\min} = [a_{x,\min} \quad a_{y,\min} \quad a_{z,\min}]^T, \\ &\qquad\quad \boldsymbol{a}_{\max} = [a_{x,\max} \quad a_{y,\max} \quad a_{z,\max}]^T\end{aligned} \tag{13}$$

또한, 오일러 기법을 이용하여 $\triangle t$에 대한 UAV의 위치 및 속도는 다음과 같이 구할 수 있다.

$$\begin{aligned}\boldsymbol{p}(t+1) &= \boldsymbol{p}(t) + \boldsymbol{v}(t)\Delta t \\ \boldsymbol{v}(t+1) &= \boldsymbol{v}(t) + \boldsymbol{a}(t)\Delta t\end{aligned} \tag{14}$$

3.2 함정 Heave Motion

해상에서 선박의 운동은 수평축에서의 운동 외에 파도에 의한 수직 방향의 상하 운동이 발생한다. 이를 heave motion이라 하며, heave motion에 의해 착륙지점은 상하 방향으로 지속 변경된다. 본 연구에서는 heave motion을 고려한 시뮬레이션을 수행하기 위하여, 다음과 같은 사인파의 합으로 운동을 모사하였다[11].

$$\begin{aligned}h(t) = {} & 0.2171\sin(0.4t) + 0.4714\sin(0.5t) + \\ & 0.3592\sin(0.6t) + 0.2227\sin(0.7t)\end{aligned} \tag{15}$$

본 연구에서는 $\dot{h}(t)$을 함정의 수직축 속도로 활용하였으며, 과적합(Overfitting)을 방지하기 위해 가우시안 분포에서 추출한 난수를 노이즈($\boldsymbol{Noise} \sim N(\mu = 0, \sigma^2 = 1)$, Max/Min = $\pm 1.5m$)로 더하여 학습을 진행하였다. 또한, 시뮬레이션 시간 내 heave motion 및 그에 따른 UAV의 행동을 가시적으로 관측하기 위하여, 위 사인파의 주기를 1/10배로 단축시켜 적용하였다.

3.3 강화학습 프레임워크

3.3.1 문제 정의

본 연구에서 강화학습 환경의 기본 구조는 Fig. 3과 같으며, 프레임워크 구성을 위해 상태 집합 $\boldsymbol{S}$, 행동 집합 $\boldsymbol{A}$, 보상 함수 $\boldsymbol{R}$에 대한 정의가 필요하다. 먼저 접근단계에서의 $\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}$을 다음과 같이 정의한다.

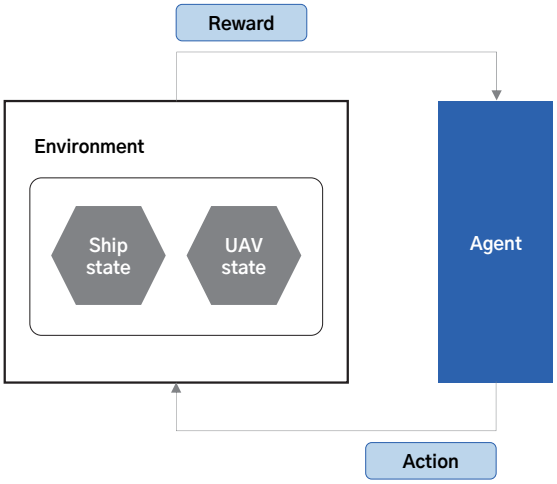


Fig. 3. 강화학습 환경 구성

$$\boldsymbol{S} = \begin{Bmatrix} x, y, z, \dot{z}, x_{\text{ship}}, y_{\text{ship}}, \dot{x}_{\text{ship}}, \dot{y}_{\text{ship}}, \\ d_{\text{prev}}, x_{\text{obs}\,(i)}, y_{\text{obs}\,(i)}, \dot{x}_{\text{obs}\,(i)}, \dot{y}_{\text{obs}\,(i)} \end{Bmatrix} \tag{16}$$

상태 집합 $\boldsymbol{S}$는 UAV의 위치 및 수직축의 속도, 모함의 위치 및 속도, 이전 시점에서의 UAV와 모함 간 상대거리인 d_{prev}가 포함되며, 접근단계에서는 heave motion을 고려하지 않기 때문에 모함의 수직축에 대한 위치는 포함시키지 않는다. 또한, obs(i)는 1척 이상 회피가 요구되는 이동표적의 위치 및 속도를 의미하며, 모든 정보는 센서 및 데이터링크 등의 양방향 통신장비를 이용하여 획득됨을 가정한다.

$$\boldsymbol{A} = \{\dot{x}, \dot{y}\} \tag{17}$$

행동 집합 $\boldsymbol{A}$는 UAV의 속도를 제어 인자로 사용하며, 수직축에 대한 속도는 UAV의 고도가 모함에 접근함에 따라 일정한 강하율을 적용하여, UAV의 행동이 아닌 상태로 고려한다. 한편, 강화학습에서 보상 함수 $\boldsymbol{R}$의 설계는 전반적인 학습 성능에 가장 큰 영향을 주는 것으로 널리 알려져 있다. 따라서, 명확한 보상 함수 설계가 요구되며, 접근단계에서는 식 (18)과 같이 6개 보상 함수의 합으로 구

성된다. 이때, $d_{\text{UAV-ship}}$은 UAV와 모함 간 상대거리, d_i는 UAV와 회피가 요구되는 이동표적 간 상대거리, d_{safety}는 UAV와 이동표적 간 안전을 위해 요구되는 최소 이격거리를 의미한다. 또한, 아래 식에서 c_i는 가중치(weight)로서 각각 상이한 특정 상수를 나타낸다.

$$\boldsymbol{R} = \begin{cases} r_1 = -c_1|\boldsymbol{v}| \\ r_2 = -c_2 \quad \text{if } d_{\text{UAV-ship}} \geq d_{\max} \\ r_3 = c_3 \quad \text{if } d_{\text{UAV-ship}} \leq d_{\text{aim}} \\ r_4 = c_4\left[d_{\text{UAV-ship}}(t-1) - d_{\text{UAV-ship}}(t)\right] \\ r_5 = -c_5 \sum_{i=1}^{n}\left(\frac{1}{d_i} - \frac{1}{d_{\text{safety}}}\right)^2 \\ r_6 = -c_6 \quad \text{if } d_i \leq d_{\text{safety}} \end{cases} \tag{18}$$

위의 보상 함수 중 r_1, r_2, r_6는 페널티로, UAV가 급격하게 속도를 변화하거나, 모함으로부터 일정거리 이상 벗어나거나, 회피가 요구되는 표적에 일정거리 내로 접근하게 되면 음수의 보상을 받는다. 반대로, UAV가 모함까지 일정거리 내(착륙단계로의 전환 조건)로 접근하게 되면 r_3보상을 받게 되며, r_4는 UAV와 모함 간 과거시점의 상대거리와 현재시점의 상대거리 차로 이를 통해 UAV가 모함과의 거리를 감소시키기 위한 행동을 하게 된다. 마지막으로 r_5를 통해 장애물 회피 시 지나친 회피 등의 비효율적인 기동을 지양한 가운데 안전거리를 유지할 수 있게 유도한다. UAV와 모함 간의 상대거리가 사전 지정된 거리인 d_{aim} 내로 접근하게 되면 착륙단계로 전환됨을 가정하며, 착륙단계의 $\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}$은 다음과 같이 정의한다.

$$\boldsymbol{S} = \begin{Bmatrix} x, y, z, \dot{x}, \dot{y}, \dot{z}, \\ x_{\text{ship}}, y_{\text{ship}}, z_{\text{ship}}, \dot{x}_{\text{ship}}, \dot{y}_{\text{ship}}, \dot{z}_{\text{ship}}, \\ d_{\text{prev}}, \dot{x}_{\text{prev}}, \dot{y}_{\text{prev}}, \dot{z}_{\text{prev}} \end{Bmatrix} \tag{19}$$

상태 집합 $\boldsymbol{S}$는 UAV의 위치/속도, 모함의 위치/속도, 이전 시점에서의 UAV와 모함 간 상대거리인 d_{prev} 및 UAV의 이전 시점 속도가 포함되며, 함정의 heave motion을 고려하기 위해 UAV 및 함정 모두 수직축에 대한 위치/속도를 포함시킨다. 모든 정보는 센서 및 데이터링크 등의 양방향 통신장비를 이용하여 획득됨을 가정한다.

$$\boldsymbol{A} = \{\ddot{x}, \ddot{y}, \ddot{z}\} \tag{20}$$

행동 집합 $\boldsymbol{A}$는 UAV의 가속도이며, 수평/수직축을 제어한다.

$$\boldsymbol{R} = \begin{cases} r_1 = -c_1|\boldsymbol{v}| \\ r_2 = -c_2 \quad \text{if } d_{\text{UAV-ship}} \geq d_{\max} \text{ or} \\ \qquad\qquad z_{\text{UAV}} < z_{\text{ship}} \\ r_3 = c_3 \quad \text{if } d_{\text{UAV-ship}} \leq d_{\text{aim}} \text{ and} \\ \qquad\qquad z_{\text{UAV}} \geq z_{\text{ship}} \\ r_4 = pushing(t-1) - pushing(t), \\ \quad \text{where } pushing(t) = \left(c_4 d_{\text{UAV-ship}} + c_5|\boldsymbol{v}|\right) \end{cases} \tag{21}$$

착륙단계의 보상 함수 $\boldsymbol{R}$은 식 (21)과 같은 4개 보상 함수의 합으로 구성된다. r_1, r_2는 페널티로, UAV가 급격하게 속도를 변화하거나, 모함으로부터 일정거리 이상 벗어나거나, UAV의 고도가 모함보다 낮아지면 음수의 보상을 받는다. 반대로 UAV가 모함까지 d_{aim} 내(착륙 조건) 도달하게 되면 r_3 보상을 받게 되며, r_4는 UAV와 모함 간 과거시점의 상대거리 및 속도와 현재시점의 상대거리 및 속도의 차로, 이를 통해 UAV가 모함과의 거리를 감소시키며 점진적인 감속을 위한 행동을 하게 된다. 이때, 접근단계에서 UAV와 모함의 상대거리는 2차원 공간상의 상대거리, 착륙단계에서는 3차원 공간상의 상대 거리로 다음과 같이 유클리드 거리(Euclidean distance)로 산출한다.

$$\begin{aligned} d_{\text{approaching}} &= \|x_{\text{UAV}} - x_{\text{ship}},\quad y_{\text{UAV}} - y_{\text{ship}}\|^2 \\ d_{\text{landing}} &= \left\| \begin{matrix} x_{\text{UAV}} - x_{\text{ship}}, & y_{\text{UAV}} - y_{\text{ship}}, \\ z_{\text{UAV}} - z_{\text{ship}} \end{matrix} \right\|^2 \end{aligned} \tag{22}$$

3.3.2 시스템 및 네트워크 아키텍처

DDPG의 네트워크는 위에서 언급한 바와 같이 크리틱과 액터 신경망으로 구성되며, Fig. 4와 Fig. 5가 각각의 구조를 나타낸다. 크리틱 신경망의 경우 상태와 행동의 2개 입력 경로로 구분되며, 상태 경로와 행동 경로는 각각 2개, 1개의 은닉층을 갖고, 이후 공통 경로로 1개의 은닉층을 지난다. 액터 신경망은 3개의 은닉층으로 구성되며, 크리틱 및 액터 신경망 모두 활성화 함수로는 rectified linear unit(ReLU) 및 tanh function(출력층 전)을 이용한다.

이때, 각 층의 노드는 은닉층의 경우 128개, 크리틱 신경망의 출력층은 1개, 액터 신경망의 출력층은 행동의 차원과 동일한 3개(접근단계 시 2개)로 설정한다. 그 외 학습을 위한 주요 hyper parameter는 Table 3와 같다.

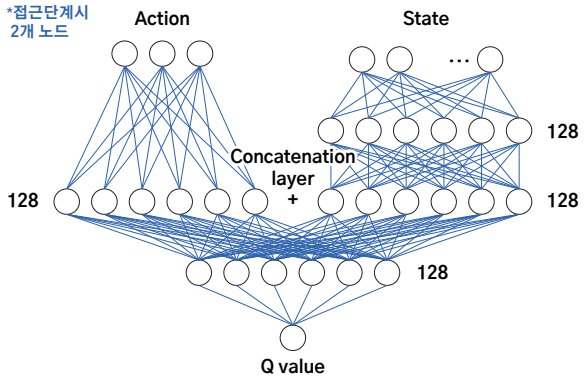


Fig. 4. 크리틱 신경망 구조

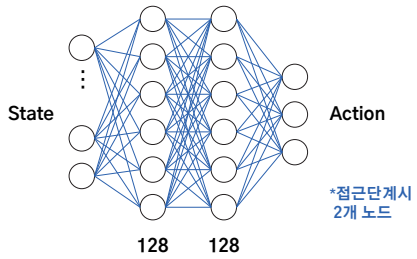


Fig. 5. 액터 신경망 구조

Table 3. 하이퍼파라미터

매개 변수	적용값
경험 재생 버퍼	10^6
미니 배치	128
액터 학습률	0.0001
크리틱 학습률	0.001
할인요인	0.99
최적화 알고리즘	Adam

3.4 시뮬레이션 결과

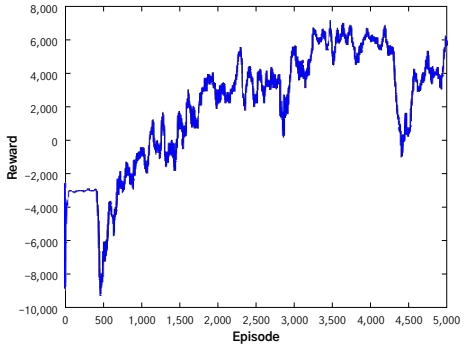
학습 시와 동일 환경에서 접근단계 및 착륙단계에 대하여 각각의 시뮬레이션을 수행하였으며, UAV의 초기위치는 접근단계의 경우 모함 외곽 약 100 m, 착륙단계에서는 약 10 m를 설정하였다. 과적합을 방지하기 위해 매 에피소드마다 UAV 및 모함의 위치는 임의의 값으로 초기화하였으며, 모함의 수평축 운동은 매 에피소드마다 다른 속도(접근단계: 약 10 m/s 내, 착륙단계: 약 3 m/s 내)로 등속운동을 하는 것으로 가정하였다. 한편, 접근단계에서 이동 장애물은 3척의 선박으로 설정하였으며, 매 에피소드마다 UAV와 모함 사이의 임의 위치를 시작점으로 하여, 매번 다른 속도로(약 7 m/s 내) 등속 운동을 하는 것으로 가정하였다. 각 단계별로 아래와 같은 성공조건 하에 산출된 시뮬레이션 결과는 Table 4와 같으며, 학습 간에 각 단계별/에피소드별 평균 보상은 Fig. 6와 같다.

· Approaching phase

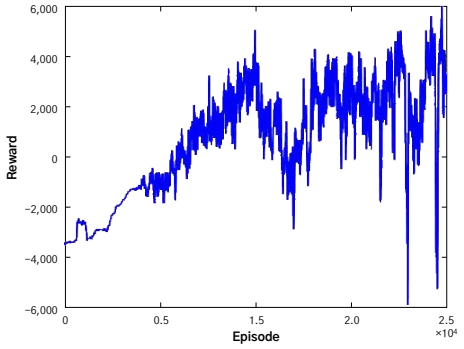
$$\text{Success condition} = d_{\text{approaching}} \leq 10 \text{ and } d_i \geq d_{\text{safety}} \quad (23)$$

· Landing phase

$$\text{Success condition} = d_{\text{landing}} \leq 0.7 \text{ and } z_{\text{UAV}} \geq z_{\text{ship}} \quad (24)$$



(a) 접근단계(approaching phase)



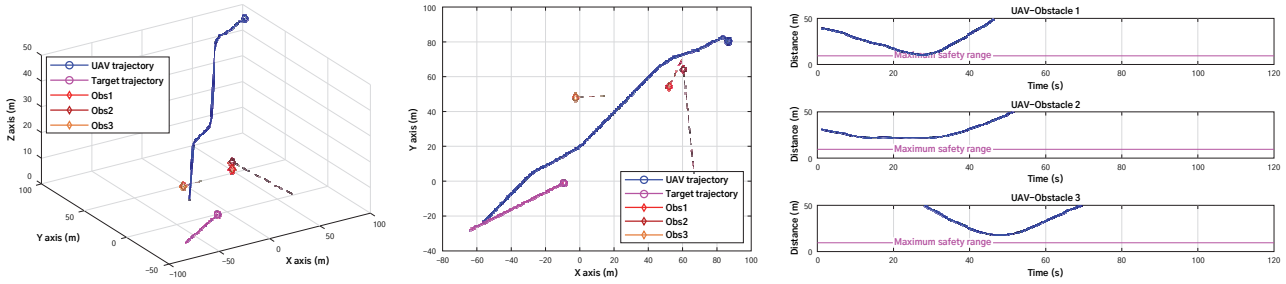
(b) 착륙단계(landing phase)

Fig. 6. 평균 보상

Table 4. 시뮬레이션 결과 (1,000회 시도)

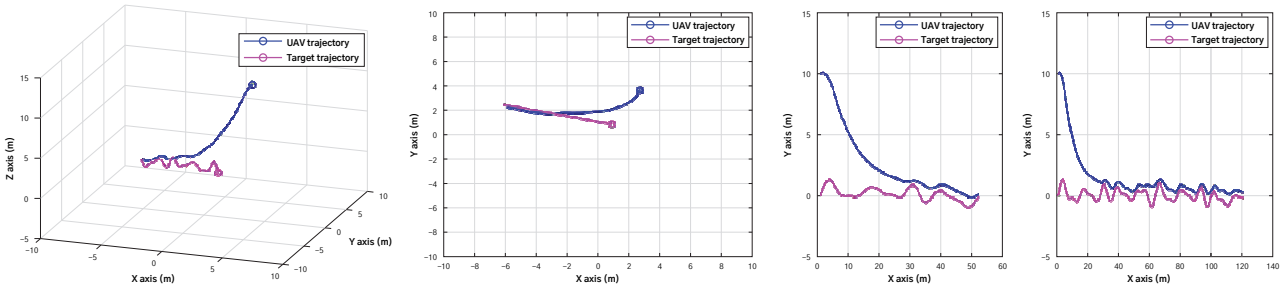
구분	성공 횟수	실패 횟수	성공률
접근단계	757	243	76 %
착륙단계	903	97	90 %

접근단계에서 성공조건을 충족한 시뮬레이션 결과는 Fig. 7과 같으며, 착륙단계에서 성공조건을 충족한 결과는 Fig. 8(a) - Fig. 8(c)와 같다. Fig. 8(d)는 성공조건을 충족하지 못한 채 최대 시뮬레이션 스텝에 도달하여 시뮬레이션이 종료된 결과이며, 기준을 충족시키진 못했으나 함정의 경로를 지속적으로 추종함을 확인할 수 있다.



(a) 3차원 경로 (b) 2차원 경로 (c) 상대거리 (UAV – 장애물)

Fig. 7. 접근 단계 시뮬레이션 결과



(a) 3차원 경로 (b) 2차원 경로 (c) 수직 경로 (d) 실패 케이스

Fig. 8. 착륙 단계 시뮬레이션 결과

4. 결론 및 향후과제

본 논문에서는 UAV의 함상착륙을 위한 접근 및 착륙 과정에서 강화학습 기반의 경로계획 기법을 제안하였다. 학습된 모델을 이용 몬테카를로 시뮬레이션을 수행하여 연구결과를 확인하였으며, 이를 통해 함 탑재 UAV의 자율화를 시도하고 함상착륙 문제에서 강화학습 기법의 활용 가능성을 확인하였다는 점에서 연구의 의의가 있다. 향후 지속적인 연구를 통해 연구 내용을 구체화 및 발전시켜 다른 경로계획 기법과 동일한 조건에서 성능 비교를 수행할 계획이며, 주요 향후 과제는 다음과 같다.

(1) 시뮬레이션 수행 간 학습 모델의 과적합 문제를 확인하였으며, 함정의 모션 등 많은 불확실성을 극복하기 위해 학습 중 외란을 모사하는 다양한 형태의 노이즈 패턴을 적용하여 모델을 학습시킬 방안이 요구된다.

(2) 본 논문에서는 경로계획 기법 구현에 집중하기 위해 UAV를 점 질량으로 가정하였으나, 향후 비행실험 및 실제 환경에서 적용을 위해서는 UAV의 동역학을 고려한 시스템 모델링이 요구된다.

(3) 행동의 변화가 최대/최소치 내에서 과도하게 발생됨을 확인하여, 향후 안정적인 행동을 위한 보상 함수 설계가 필요하며, 마지막으로 원거리 임무 수행 후 복귀를 가정하여 확장된 스케일에서의 후속 연구가 요구된다.

참고문헌

[1] G. J. Duan and P. F. Zhang, “Research on Application of UAV for Maritime Supervision,” Journal of Shipping and Ocean Engineering, No. 4, pp. 322-326, 2014.

[2] Dan Gettinger, “Summary of Drone spending in the FY 2019 Budget Request,” Center for the Study of the Drone at Bard College, 2018.

[3] Kim, J. G. and Lee, S. H., “Study on Possible Use of Navy’s Future Military Drone,” Proceedings of the Korean Society of Computer Information Conference, Vol. 28, No. 1, pp. 83-86, 2020.

[4] T. Zhang et al., “Current trends in the development of intelligent unmanned autonomous systems,” Frontiers Inf. Technol. Electron. Eng., Vol. 18, No. 1, pp. 68-85, 2017.

[5] Z Gan et al., “UAV Maneuvering Target Tracking based on Deep Reinforcement Learning,” Journal of Physics: Conference Series. 1958 012015, 2021.

[6] Alejandro Rodriguez-Ramos, Carlos Sampedro, Hriday Bavle, Ignacio Gil Moreno and Pascual Campoy, “A Deep Reinforcement Learning Technique for Vision- Based Autonomous Multirotor Landing on a Moving Platform,” Proceeding of the 2018 IEEE/RSJ International Conference,

pp. 1010-1017, 2018.
[7] Bae, Y. K., "Strategic Response against Drones at Sea," Journal of the KNST, Vol. 4, No. 1, pp. 36-41, 2021.
[8] Li, B., Yang, Z.P., Chen, D.Q., Liang, S.Y. and Ma, H., "Maneuvering target tracking of UAV based on MN-DDPG and transfer learning", Def. Technol, Vol. 17, pp. 457-466, 2020.
[9] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning", arXiv preprint arXiv:1509.02971, 2015.
[10] Cho, Y. W., Lee, J. S. and Lee, K. Y., "CNN based Reinforcement Learning for Driving Behavior of Simulated Self-Driving Car", The transactions of The Korean Institute of Electrical Engineers, Vol. 69, No.11, pp. 1740-1749, 2020.
[11] Hanjie Hu., Yu Wu., Jinfa Xu. and Qingyun Sun.,"Path Planning for Autonomous Landing of Helicopter on the Aircraft Carrier," Mathematics, 6 (10), 178, 2018.
[12] Rho, S. G., "RL from basics," Youngjin.com Inc., Seoul, 2020.