A UAV Path Planning Method Based on Deep Reinforcement Learning

Yibing Li, Sitong Zhang, Fang Ye*, Tao Jiang, Yingsong Li
College of Information and Communication Engineering
Harbin Engineering University
Harbin, Heilongjiang, China
liyibing0920@126.com, zhangsitong0726@126.com, yefang0923@126.com, jiangtao@hrbeu.edu.cn,liyingsong@hrbeu.edu.cn

***Abstract*—The path planning of Unmanned Aerial Vehicle (UAV) is a critical component of rescue operation. As impacted by the continuity of the task space and the high dynamics of the aircraft, conventional approaches cannot find the optimal control strategy. Accordingly, in this study, a deep reinforcement learning (DRL)-based UAV path planning method is proposed, enabling the UAV to complete the path planning in a 3D continuous environment. The deep deterministic policy gradient (DDPG) algorithm is employed to enable UAV to autonomously make decisions. Besides, to avoid obstacles, the concepts of connected area and threat function are proposed and adopted in the reward shaping. Lastly, an environment with static obstacles is built, and the agent is trained using the proposed method. As has been proved by the experiments, the proposed algorithm can fit a range of scenarios.**

***Index Terms*—Unmanned aerial vehicle, path planning, deep reinforcement learning, reward shaping**

I. Introduction

UAVs have broad applications in terrain mapping, search and rescue. In different mission scenarios, autonomous path planning can ensure the mission to be completed. Due to the continuity of the task space and the high maneuverability of the UAV, conventional methods cannot serve as optimal control strategies. Nevertheless, reinforcement learning is capable of performing continuous decision-making tasks in complex environments. A global path planning method based on DQN was proposed in [1], enabling mobile robots to effectively obtain the optimal path in a dense environment. In [2], DQN was employed to track targets in a two-dimensional air combat environment. An optimized Double DQN method was proposed in [3] for UAV path planning in dynamic environment with potential enemy threats. However, the mentioned works assumed that the task environment is a two-dimensional and discrete space, so these methods could not satisfy engineering requirements. In brief, this study, with the fixed-wing UAV as the research background, proposes a path planning method based on DDPG in a 3D continuous environment with aerial obstacles.

II. Problem Formulation

In this section, our UAV path planning system is presented for simulation. First, the states and actions of this system are elucidated. Next, the goals of the UAV are analyzed.

A. State

The state of the UAV is defined by the position, velocity, as well as acceleration. The motion state of the UAV at time $t+1$ is defined as:

$$\mathbf{U}_{t+1} = \begin{bmatrix} \mathbf{p}_{uav}^{t+1} \\ \mathbf{v}_{uav}^{t+1} \\ \mathbf{a}_{uav}^{t+1} \end{bmatrix} = \begin{bmatrix} I & \tau I & \tau^2/2 \\ 0 & I & \tau I \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \mathbf{p}_{uav}^{t} \\ \mathbf{v}_{uav}^{t} \\ \mathbf{a}_{uav}^{t} \end{bmatrix} \tag{1}$$

where $\mathbf{p}_{uav}^t$ denotes the position of the UAV at time t; $\mathbf{v}_{uav}^t$ and $\mathbf{a}_{uav}^t$ refer to the velocity vector and the acceleration vector of the UAV during the time interval, respectively.

B. Action

The action is expressed as:

$$\mathbf{a}_t = [\varphi_1, \varphi_2] \tag{2}$$

where $\varphi_1 \in [-\pi, \pi]$ and $\varphi_2 \in [-2/\pi, 2/\pi]$ indicate the yaw and pitch angles of the UAV, respectively.

In terms of the dynamics of the UAV, the work of [4] is largely referenced, so the dynamics is not elucidated here.

C. Goal

The UAV aims to maintain a distance from the obstacles and reach the target area.

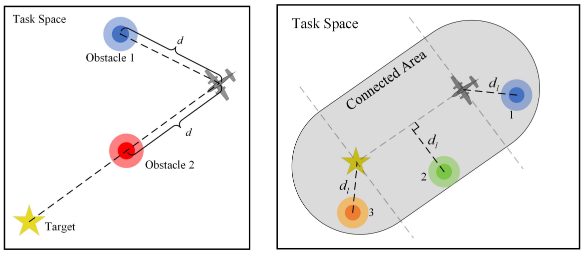
First, to guide the UAV to reach the target area, a reward function R_d related to the distance from the UAV to the target point is expressed as following,

$$R_d = -\left\| \oslash \left(\mathbf{p}_{uav}^t - \mathbf{p}_{target} \right) \right\| \tag{3}$$

where $\oslash(\cdot)$ denotes normalization; $\| \cdot \|$ indicates L2 norm; $\mathbf{p}_{target}$ refers to the position of the target point. Equation (3) suggests that the farther the UAV from the target area, the higher the negative reward it will get.

Second, to keep a certain distance from obstacles, a threat function T is defined, representing the threat degree of the obstacles on the path planning. The schematic diagram of the task space is illustrated in Fig.1(b). As revealed from the picture, Obstacle 1 and 2 are at the identical distance from the UAV, whereas they impose threats to the path planning task with different extents. Obstacle 2 is noticeably more threatening to the task.

Accordingly, the connected area l is defined, namely, a function about the position of the UAV and the target point. Subsequently, the distance between the obstacle and the connected area is defined as d_l, as presented in Fig.1(b).



(a) (b)

Fig. 1. (a) The schematic diagram of the task space; (b) The definition of the distance between the obstacle and the connected area.

When the obstacle is located at position 1, d_l is the distance from the obstacle to the UAV; when it is located at position 2, d_l is the distance from the obstacle to the line segment formed by drone and target point; when it is located at position 3, d_l is equal to the distance from the obstacle to the target point. Accordingly, the threat function T is written as following.

$$T = \begin{cases} \exp\left(1 - \frac{d_{\min}^2}{d_{\min}^2 - d_l^2}\right) & d_l < d_{\min} \\ 0 & d_l > d_{\min} \end{cases} \tag{4}$$

The negative value of T is taken as a penalty for the UAV approaching obstacles and then employed in reward shaping.

III. Methodology

In this section, features for the network input are developed, and rewards are designed to solve the path planning problem.

A. DDPG

The path planning problem can be formulated as a Markov decision problem (MDP) in math. DDPG refers to a model-free algorithm to solve the MDP problem, i.e., decision-making process is independent of the system dynamics. For the process of DDPG, reference [5] is primarily referenced, so such process is not elucidated here.

B. Feature

Unlike inputting the system state directly into the network, we design the feature s_t which reflects the relative information of the goal. Besides, we normalize it to speed up DRL convergence.

$$\mathbf{s}_t = \left[\ \oslash(\mathbf{p}_{uav}^t), \oslash(\mathbf{p}_{uav}^t - \mathbf{p}_{target}), \oslash(\mathbf{v}_{uav}^t) \ \right]^T \tag{5}$$

C. Reward shaping

Reward refers to one of the critical points of reinforcement learning, which largely determines the performance of the algorithm. In this study, a reward function is designed to guide the UAV to reach the target area, while ensuring its safety. The setting of the reward function is listed in Table 1. R_d and T have already been mentioned above.

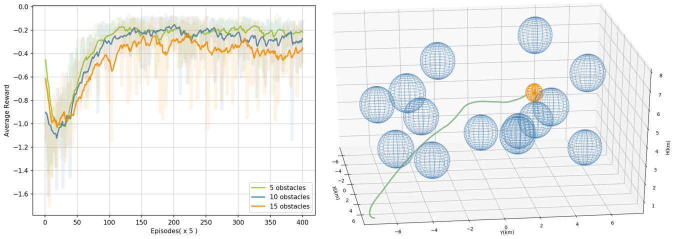
TABLE I
The setting of reward function.

State	Value
UAV arrives within 1km of the target	10
UAV moves away from or close to the target	R_d
Connected area is away from or close to obstacles	$-T$
UAV collides with obstacles	-2

IV. Experiments

The proposed method is trained in the environment with static obstacles. The task space for the UAV is a cuboid, 15 km in length (from -7.5 km to 7.5 km), 15 km in width (from -7.5 km to 7.5 km), as well as 7.5 km in height (from 0.5 km to 8 km). Obstacles and threat areas are simplified to spheres. At each episode, the target point and obstacles exhibit random locations.

The smoothed curves of the cumulative rewards in the different training environments are plotted in Fig.2. The figure reveals that the network converges at the number of obstacles as 5, 10, and 15. The optimal trajectory planned by the proposed method is illustrated in Fig.3. As revealed from the figure, the proposed method is capable of guiding the UAV to avoid obstacles successfully, and the path is short.



(a) (b)

Fig. 2. (a) Smoothed curves of cumulative rewards obtained from a range of static situations with the proposed method; (b) Diagram of optimal trajectory planned by the proposed method (the orange sphere, blue spheres, and the green line represent target area, threat area and UAV path).

Acknowledgement

This paper is funded by the National Natural Science Foundation of China (Grant No. 61701134) and the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation.

References

[1] Zhou S, Liu X, Xu Y, et al. A Deep Q-network (DQN) Based Path Planning Method for Mobile Robots[C], 2018.

[2] Ma X, Xia L, Zhao Q. Air-Combat Strategy Using Deep Q-Learning[C], 2018.

[3] SHEN Yangyang, YANG Zhong, XU Hao, et al. Trajectory control for a small quad tilt-rotor aircraft[J]. Applied science and technology,2018,45(03):71-75.[doi:10.11991/yykj.201706014]

[4] You S, Diao M, Gao L. Deep Reinforcement Learning for Target Searching in Cognitive Electronic Warfare[J]. IEEE Access, 2019, 7: 37432-37447.

[5] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.