doi:10.3772/j.issn.1002-0470.2022.10.006

基于 DDPG 三维无人机路径规划①

司鹏搏② 吴 兵 杨睿哲③ 李 萌 孙艳华

(北京工业大学信息学部信息与通信工程学院 北京 100124)

摘 要 无人机(UAV)以其移动性高、环境适应能力强等优点,引起了广泛的关注,并被应用于军事和民用领域。本文研究了在复杂环境中,使用深度确定性策略梯度算法(DDPG)实现无人机路径规划。首先,基于三维场景模型,将无人机任务过程划分为飞行、等待、通信阶段;其次,提出三维偏离度来表示无人机与障碍物及目标用户的相对位置,以提高无人机飞行和避障的有效性;最后,采用深度确定性策略梯度算法规划无人机的连续飞行动作,实现减少能量消耗、提高服务质量(QoS),同时避开障碍、完成对用户数据传输的目的。通过仿真验证所提方案在各参数配置下的有效性,且优于现存算法。

关键词 无人机(UAV);三维场景;路径规划;深度确定性策略梯度算法(DDPG);避障

0 引 言

无人机(unmanned aerial vehicle, UAV)由于其体积小、成本低、环境适应力强等优点,获得了广泛关注,已被应用在目标追踪[1]、通信[2]、监测[3]、农业[4]、灾难管理[5]等方面。无人机在完成任务时,自主导航是实现对无人机控制的关键部分,因此,无人机路径规划是实现无人机自主飞行的重要因素。路径规划是确定无人机从起始点到目标点的路径,其目的不仅在于寻找最佳和最短的路径,而且还为无人机提供无碰撞的环境,并在运动动力学约束下优化给定的成本函数[6]。

近年来,对无人机路径规划的研究越来越多。无人机飞行路径规划是一个复杂的优化问题,需要考虑路径长度、时间消耗、能量消耗、障碍规避、鲁棒性等多个问题,文献[7]提出一种基于多宇宙优化器(multi-verse optimizer,MVO)的 2D 无人机路径规划方案,将服务质量(quality of service, QoS)作为衡量路径优劣的指标,考虑多个无人机的协同工作与碰撞,同时也将最短路径与最短时间作为约束条件。文献[8]研究一种城市环境中无人机导航覆盖路径规划算法,考虑障碍物环境下无人机无障碍最短路径的路径规划,并探索不同障碍物形状对路径的影响。在实现无人机路径规划优化问题的探索中,研究学者提出了很多无人机路径规划算法,如 A* 算法[9]、人工势场[10]、线性规划[11]、随机树[12]等算法,但是,当无人机路径规划具有多个约束条件时,这些方法中的大多数都具有较高的时间复杂度和局部极小陷阱[13],且如果在大范围的环境下,计算压力也会急剧增加。

为了解决这些问题,将深度强化学习(deep reinforcement learning, DRL)算法引入无人机路径规划研究中。深度强化学习是将具有感知能力的深度学习与具有决策能力的强化学习相结合,所形成的一种端对端的感知与控制系统,使用函数拟合的方法对 Q 表逼近,使其在高维环境下也有很好的效果,具有很强的通用性[14]。文献[15]研究搜索和救援场景中的无人机导航,提出扩展双深度 Q 网络(double deep Q-network, DDQN)算法用于基于无人机捕获的图像来提高无人机对环境的理解,大幅减

① 国家自然科学基金(61901011)和北京市教委科技计划项目(KM202010005017,KM202110005021)资助。
② 男,1983 年生,教授;研究方向:区块链技术,深度强化学习,无线通信网络,无线资源管理;E-mail: sipengbo@bjut.edu.cn。
③ 通信作者,E-mail: yangruizhe@bjut.edu.cn。
(收稿日期:2021-09-03)

少了每个任务期间处理的数据量。文献[13]将环境建模为有障碍的三维环境,提出将强化学习算法与灰狼优化算法(grey wolf optimizer, GWO)结合的算法,并将路径规划分为搜索、几何调整和最佳调整三部分,解决局部优化中陷入困局和无人机路径规划不平稳的问题。文献[16]提出了一种快速态势评估模型,能够将全球环境状况转换为顺序的态势图,采用了决斗双深度 Q 网络(dueling double deep Q-network, D3QN)算法,并将 ε 贪心策略与启发式搜索规则结合选择动作,使用网格方法将动作划分为 8 个离散的值。文献[17]用 Q 学习算法,并将 Q 值基于表的近似和神经网络(neural network, NN)近似进行对比,而对于无人机的动作值同样需要离散化。以上深度强化学习算法的应用虽然都取得了良好的效果,但大多数算法都需要将动作空间离散化,这样就限定了无人机只能在特定几个方向进行转角与飞行,而在实际中无人机的飞行方向需是全方位的,且由于需不断躲避障碍物,其高度也不断变化,此时,再将动作值离散化会大幅增加计算负担。

本文研究复杂环境、连续空间状态下,无人机无碰撞的路径规划问题。首先,建立一种复杂 3D 场景模型,将无人机任务过程划分为飞行、等待、通信 3 个阶段;其次,提出一种无人机高度避障方法,引入偏离度 δ 表示无人机与障碍物及目标用户的相对位置;最后,采用深度确定性策略梯度算法[18](deep deterministic policy gradient, DDPG)实现无人机路径规划,并与现有算法比较以验证提出方法的有效性。

1　系统模型

假设在一定区域的城市空间中,分布着如手机、电脑等智能用户,由于自然灾害、距离等原因,用户不能直接与基站通信,为保障灾后救援,满足用户需求,使用体积小、对环境要求低的无人机作为中继通信。无人机的飞行任务需满足以下约束。

(1) 用户(UEs)随机分布,且 UEs 之间互联互通,每个 UE 都能接收来自邻近 UEs 的消息。

(2) UAV 从结束收集 UE 数据到结束收集下一个 UE 数据为一个飞行任务。

(3) UAV 在一个任务中能量充足,不考虑由于能量耗尽导致任务终止。

如图 1 所示,包括 1 个 UAV 以及随机分布的 N 个 UEs 和 M 个障碍物 OBs 。当 UE 有数据传输请求时,会向全网广播其位置信息。而位于 UAV 通信范围内的 UE 则会将其获得的具有数据传输请求的 UE 位置信息传递给 UAV 。UAV 获得数据请求信息后,利用深度确定性算法规划路径、规避障碍,向目标 UE 移动并为其提供服务。UAV 服务完毕后,若无新的 UE 数据上传请求,UAV 将悬停在此处,等待新的目标 UE。实际情况中,UAV 由于体积小,搭载能量有限,UAV 需要在有限的能量限制下服务更多 UE;同时,为满足用户的服务质量,需要在最短时间内完成飞行任务,并且避免与障碍物的碰撞。

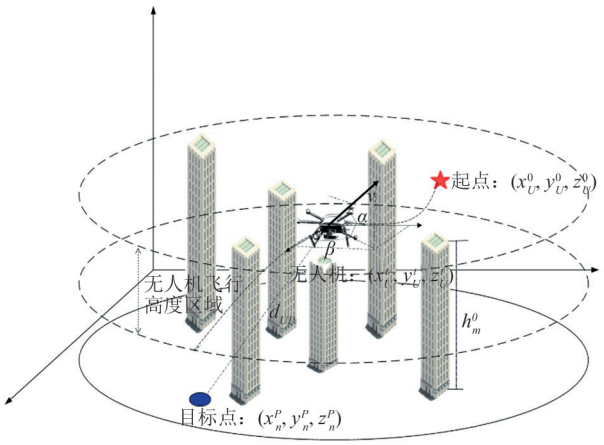

图 1　无人机路径规划系统模型

为简化问题,在三维坐标系中对环境建模,如图 1所示,将地面建模为 oxy 平面, z 轴满足右手法则,有以下假设:UAV 在 t 时刻的位置坐标为 (x_U^t, y_U^t, z_U^t), r_U 为 UAV 半径, $h_{\min}$、$h_{\max}$ 为 UAV 飞行的最小及最大高度, $E_{\max}$ 为 UAV 搭载的最大能量, v 为 UAV 的速度, b 为加速度, $v_{\max}$ 表示 UAV 飞行时能够达到的最大线速度, α 为俯仰角, β 为偏航角, ρ_α、ρ_β 为相应的角速度;障碍物集合 $\boldsymbol{O} = \{O_1, O_2, \cdots, O_m, \cdots, O_M\}$ 随机分布在 UAV 服务区域, (x_m^O, y_m^O, z_m^O) 为障碍物 O_m 的中心坐标, h_m^O 为 O_m 的高度, r_m^O 为 O_m 的半径;UEs 集合 $\boldsymbol{P} = \{P_1, P_2, \cdots, P_N\}$ 随机分布在 UAV 服务区域, (x_n^P, y_n^P, z_n^P) 为 UE P_n 的位置坐标, $\dot{X}$ 表示对 X 求导。规定 UAV 基于 DDPG 训练

一步为一个时隙 t, 在三维空间中,UAV 飞行的运动学模型可建模为[19]

$$\dot{x}_U = v\cos\alpha\cos\beta \tag{1a}$$

$$\dot{y}_U = v\cos\alpha\sin\beta \tag{1b}$$

$$\dot{z}_U = v\sin\vartheta \tag{1c}$$

$$\dot{\alpha} = \rho_\alpha \tag{1d}$$

$$\dot{\beta} = \rho_\beta \tag{1e}$$

$$\dot{v} = b \tag{1f}$$

2 无人机任务建模与分析

假设 UAV 已完成 UEs 中 $P_{n\text{-}1}$ 的数据收集,正在等待或直接前往 P_n 收集数据,则无人机从 $P_{n\text{-}1}$ 飞往 P_n。

2.1 任务建模

2.1.1 飞行距离

t 时刻 UAV 到 UE 的位移 d_{UP} 为

$$d_{UP} = \sqrt{(x_U^t - x_n^P)^2 + (y_U^t - y_n^P)^2 + (z_U^t - z_n^P)^2} \tag{2}$$

则,UAV 飞行的最短距离 $d_{\min}$ 为

$$d_{\min} = \sqrt{(x_U^t - x_n^P)^2 + (y_U^t - y_n^P)^2} \tag{3}$$

实际情况中,UAV 需躲避障碍,避免碰撞,UAV 实际飞行距离 d_U 满足: $d_U \geqslant d_{\min}$。

2.1.2 俯仰角 α 与 偏航角 β

UAV 的速度 v 与 z 轴的夹角为俯仰角 α; UAV 与 P_n 投影在 xoy 平面,UAV 的速度 v 与 x 轴的夹角为偏航角 β, 则:

$$\alpha_t = \arccos\frac{|\ z_U^t - z_n^P\ |}{d_{UP}} \tag{4a}$$

$$\beta_t = \arccos\frac{|\ x_U^t - x_n^P\ |}{\sqrt{(x_U^t - x_n^P)^2 + (y_U^t - y_n^P)^2}} \tag{4b}$$

α_t、β_t 分别为 t 时刻的俯仰角和偏航角。

UAV 在飞行中, α 与 β 随 UAV 速度的变化而不断变化,则 α 与 β 的变化有以下规律:

$$\alpha_{t+1} \leftarrow \alpha_t + \rho_\alpha t \tag{5a}$$

$$\beta_{t+1} \leftarrow \beta_t + \rho_\beta t \tag{5b}$$

2.2 成本函数

2.2.1 障碍规避与目标抵达

UAV 在接收到 P_n 的位置信息后,在向 P_n 飞行的过程中,需要避开障碍,尽可能到达 P_n 上方接收数据,因此,引入偏离向量集 $\boldsymbol{\Phi} = \{\boldsymbol{\sigma}_1^o, \boldsymbol{\sigma}_2^o, \cdots, \boldsymbol{\sigma}_m^o, \cdots, \boldsymbol{\sigma}_M^o\}$, 其中, $\boldsymbol{\sigma}_m^o = (\sigma_{mx}^o, \sigma_{my}^o)$ 辅助判断 UAV 与障碍物是否碰撞,其中:

$$\begin{cases} \sigma_{mx}^o = |\ x_m^O - x_U^t\ | \\ \sigma_{my}^o = |\ y_m^O - y_U^t\ | \end{cases} \tag{6}$$

为避免 UAV 与障碍物的碰撞,偏离向量 $\boldsymbol{\sigma}_m^o = (\sigma_{mx}^o, \sigma_{my}^o)$ 需满足以下约束条件:

如果 $h_m^O \geqslant h_{\max}$, 则:

$$\begin{cases} \sigma_{mx}^o > r_U + r_m^O + \varepsilon \\ \sigma_{my}^o > r_U + r_m^O + \varepsilon \end{cases} \tag{7}$$

如果 $h_{\min} < h_m^O < h_{\max}$, 则:

$$\begin{cases} \text{无约束} & z_U > h_m^O \\ \sigma_{mx}^o > r_U + r_m^O + \varepsilon & z_U \leqslant h_m^O \\ \sigma_{my}^o > r_U + r_m^O + \varepsilon & \end{cases} \tag{8}$$

如果 $h_m^O \leqslant h_{\min}$, 则:

无约束。

以上 $0 < \varepsilon < 1$。 (9)

对于 UAV 悬停位置的判断,引入目标偏离向量 $\boldsymbol{\sigma}_P = (\sigma_{Px}, \sigma_{Py})$, 其中:

$$\begin{aligned} \sigma_{Px} &= |\ x_n^P - x_U^t\ | \\ \sigma_{Py} &= |\ y_n^P - y_U^t\ | \end{aligned} \tag{10}$$

当 UAV 到达目标 UE 附近,为提高数据传输效率,则存在极小值 $\epsilon(0 < \epsilon < 1)$, 使得 UAV 悬停位置满足以下约束条件:

$$\begin{aligned} \sigma_{Px} &< \epsilon \\ \sigma_{Py} &< \epsilon \end{aligned} \tag{11}$$

2.2.2 任务时间

UAV 完成一个任务过程所需时间包括 3 部分:飞行时间 T_f、等待时间 T_w、通信时间 T_{com}。

飞行时间 T_f: UAV 从 P_{n-1} 出发至到达 P_n 耗费的时间,当 UAV 以最大速度飞行最小距离时,耗费最短飞行时间为

$$T_{f\min} = \frac{d_{\min}}{v_{\max}} \tag{12}$$

UAV 在飞行中,为躲避障碍,需不断改变飞行方向及飞行高度,则飞行时间 T_f 满足:

$$T_f > T_{f\min} \tag{13}$$

等待时间 T_w: UAV 等待下一个具有数据传输

请求 UE 出现的时间。

通信时间 T_{com}：UE 将数据传输到 UAV 耗费的时间，在该过程中，数据接收率为 R，则传输 D_n 数据量耗时为

$$T_{\text{com}} = D_n / R \tag{14}$$

综上，UAV 完成从 P_{n-1} 到 P_n 的数据收集任务耗费的总时间为

$$T_{\text{total}} = T_f + T_w + T_{\text{com}} \tag{15}$$

2.2.3 能量消耗

在一个数据收集过程中，耗能分为 3 种，分别为飞行、等待、通信，各阶段耗能情况如下。

飞行能耗：每时隙耗能 e_f，耗时 T_f，则总耗能为

$$E_f = e_f \times T_f \tag{16}$$

等待能耗：UAV 悬停在 UE 上方每时隙耗能 e_w，耗时 T_w，则悬停总耗能为

$$E_w = e_w \times T_w \tag{17}$$

通信能耗：每时隙耗能 e_{com}，耗时 T_{com}，则通信总耗能为

$$E_{\text{com}} = e_{\text{com}} \times T_{\text{com}} \tag{18}$$

综上，UAV 在一个任务中总耗能为

$$E_{\text{total}} = E_f + E_w + E_{\text{com}}, E_{\text{total}} < E_{\max} \tag{19}$$

3 基于 DDPG 的无人机路径规划

3.1 深度确定性策略梯度算法

深度确定性策略梯度算法适用于连续动作空间，包括 Actor 网络和 Critic 网络两部分，二者利用深度神经网络分别实现对策略和 Q 函数的逼近[20-21]。DDPG 的训练过程如下。

(1) Actor 网络在状态 s_t 下给出动作 $a_t = \pi(s_t)$，为了增加样本的随机性，会对 Actor 网络给出的动作 $a_t = \pi(s_t)$ 增加一个随机噪声（使用 Uhlenbeck-Ornstein 随机过程，作为引入的随机噪声）$\mathscr{N}$，即行为动作 $\varphi_t = \pi(s_t) + \mathscr{N}$。

(2) 动作 φ_t 作用于环境，DDPG 得到奖赏 r_t 和下一个状态 s_{t+1}，DDPG 将集合 $(s_t, \varphi_t, r_t, s_{t+1})$ 存储到经验缓冲区 H。

(3) DDPG 从经验缓冲区随机选取大小为 K 的小批量数据集作为 Actor 网络和 Critic 网络的输入。

(4) 在 Critic 网络，目标 Critic 网络利用式(20)根据小批量数据集计算累计奖赏更新：

$$y_t = r_t + \gamma Q_{\omega'}(s_{t+1}, \pi_{\theta'}(s_{t+1})) \tag{20}$$

在线 Critic 网络利用动作 φ_t 逼近目标 Q 值 $Q_w(s_t, \varphi_t)$，并使用最小化损失函数式(21)进行在线 Critic 网络的更新。

$$L_\omega = \frac{1}{K}\sum_t (y_t - Q_w(s_t, \varphi_t))^2 \tag{21}$$

其中，ω 为在线 Critic 网络的参数，ω' 为目标 Critic 网络的参数，$\pi_{\theta'}(s_{t+1})$ 为目标 Actor 网络根据小批量数据集得出的下一状态的动作。

(5) 在 Actor 网络中，使用式(22)：

$$\nabla_\theta L = \frac{1}{K}\sum_t [\nabla_\varphi Q_w(s_t, \varphi_t) \nabla_\theta \pi(s) |_{s=s_t}] \tag{22}$$

对网络进行更新，其中，θ 为在线 Actor 网络的参数，θ' 为目标 Actor 网络的参数。

(6) 通过步骤(4)、(5)分别对在线 Critic 网络及 Actor 网络参数更新，而目标网络的参数以一定的频率从在线网络复制更新，更新规则分别为式(23a)与(23b)。

$$\omega' \leftarrow \tau\omega + (1-\tau)\omega \tag{23a}$$

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta \tag{23b}$$

3.2 基于 DDPG 的无人机路径规划设计

本文针对连续空间内无人机路径规划，将适用于连续空间问题的 DDPG 算法引入，以寻求满足优化目标的最优路径。

状态空间：包括无人机坐标 (x_U^t, y_U^t, z_U^t)，无人机速度：x 轴方向速度 v_x^t、y 轴方向速度 v_y^t 以及 z 轴方向速度 v_z^t，则 t 时刻的状态空间为

$$s_t = \{(x_U^t, y_U^t, z_U^t), v_x^t, v_y^t, v_z^t\}$$

动作空间：t 时刻分别在 x 轴、y 轴、z 轴方向的加速度，则 t 时刻的动作值为

$$a_t = \{b_x^t, b_y^t, b_z^t\}$$

奖赏：合理的奖赏设置能够更加快速地训练出最优的策略。为使 UAV 用最短时间、最小能量消耗到达目的点，同时避开障碍，以及更加接近目的点，则将奖赏划分以下几个部分。

障碍物奖赏 r_{obs}：如果 UAV 与障碍物的位置关系满足式(7)、(8)、(9)，则 $r_{\text{obs}} = 0$，否则 $r_{\text{obs}} = -r_{\text{obs}}^0$，并且结束游戏。

路径奖赏 r_{TE}:主要包括对路径中的时间及能耗的衡量。

$$r_{\mathrm{TE}} = -(T_{\mathrm{total}} + E_{\mathrm{total}}) \times \mu, 0 < \mu < 1$$

目的点奖赏 r_{des}:衡量 UAV 是否到达目的点完成任务,$r_{\mathrm{des}} = r_{\mathrm{des}}^{0}$。

区域奖赏 r_{b}:将 UAV 限定在一定区域内,当 UAV 飞出该区域,$r_{\mathrm{b}} = -r_{\mathrm{b}}^{0}$。

综上,则 t 时刻总奖赏为式(24)。

$$r_t = r_{\mathrm{obs}\,t} + r_{\mathrm{TE}\,t} + r_{\mathrm{des}\,t} + r_{\mathrm{b}\,t} \tag{24}$$

则基于 DDPG 无人机路径规划算法(deep deterministic policy gradient algorithm UAV path planning, DDPG-UPP)具体内容如算法 1 所示。

算法 1 基于 DDPG 无人机路径规划算法

初始化 UAVs 与 UEs 的位置;
初始化经验缓冲区 H;
初始化目标网络和在线网络;
for each episode do
  初始化环境并获得初始状态 s_1;
  初始化随机噪声 $\mathscr{N}_t$;
  for each step t of episode do
    基于状态 s_t 得到动作 $\varphi_t = \pi(s_t) + \mathscr{N}$;
    执行动作 φ_t,得到新状态 s_{t+1};
    if UAV fly beyond the border then
      限制 UAV 在研究区域内,并给予奖赏 $-r_{\mathrm{b}}^{0}$
    end if
    if UAV collides with obstacles then
      结束任务,并获得奖赏 $-r_{\mathrm{obs}}^{0}$;
    end if
    if UAV arrives at the destination then
      结束任务,并获得奖赏 r_{des}^{0};
    end if
    获得奖赏 r_t;
    将元组($(s_t, \varphi_t, r_t, s_{t+1})$)存储到经验缓冲区 H;
    从经验缓冲区 H 中采样小批量数据 K 计算当前目标网络 Q 值:

$$y_t = r_t + \gamma Q_{\omega'}(s_{t+1}, \pi_{\theta'}(s_{t+1})).$$

    使用均方差损失函数更新 Critic 在线网络参数:

$$L_\omega = \frac{1}{K}\sum_t (y_t - Q_w(s_t, \varphi_t))^2$$

    更新 Actor 在线网络参数:

$$\nabla_\theta L = \frac{1}{K}\sum_t [\nabla_\varphi Q_w(s_t, \varphi_t) \nabla_\theta \pi(s) \mid_{s=s_t}]$$

    更新目标网络参数:

$$\omega' \leftarrow \tau\omega + (1-\tau)\omega$$

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta$$

  end for
end for

4 仿真分析

本部分将通过仿真评估算法 DDPG-UPP 的性能,仿真环境使用 Python 3.6、TensorFlow 1.12。本实验将模拟 500 m×500 m×500 m 区域内无人机使用 DDPG-UPP 算法从起点到目标点的路径规划情况,其中障碍物随机分布在该区域内。本文测试 DDPG-UPP 算法的性能通过不同学习率性能比较、不同算法及不同维度路径规划的性能比较,从而获得最优学习率并验证 DDPG-UPP 算法的最优性。仿真使用的各参数设置如表 1 所示。

算法 1[22] 采用演员评论家(Actor-Critic, AC)算法,并融合指针网络(pointer network-A*, Ptr-A*)进行无人机路径规划探索,将 Ptr-A* 的参数在小规模聚类问题实例上进行训练,以便在 Actor-Critic 算法中进行更快的训练。

算法 2[16] 采用决斗双深度 Q 网络 D3QN 算法,同时使用 ε-greedy 策略与启发式搜索结合选择动作,实现离散环境下无人机自主路径规划。

算法 3 采用了策略梯度(policy gradient, PG),将策略表示为连续函数,并用梯度上升等连续函数优化方法寻找最优策略,有效弥补了基于值函数算法(DQN 等)适用场景的不足。

表 1 仿真参数

参数	定义	值
$v_{\max}$	UAV 飞行的最大速度	20 m/s
$E_{\max}$	UAV 电池容量	100 kJ
γ	折扣因子	0.95
τ	软更新参数	0.001
B	经验缓冲集	100 000
b	小批量数据集	128
episodes	训练的周期数	10 000
step	每周期迭代的最大步数	500

图 2、图 3 分别为在二维与三维环境下对无人机路径规划的效果采样图。图 3 对三维环境无人机路径规划仿真实验中设置无人机与目标点的阈值为 20，即当无人机在以目标点为中心、20 为半径的球形区域内时，可认为无人机到达目标位置。通过对比，在将环境从二维拓展到三维并不断增加障碍物数量的过程中，使用本文算法训练的无人机都能准确到达目标点，同时精准避开障碍物。

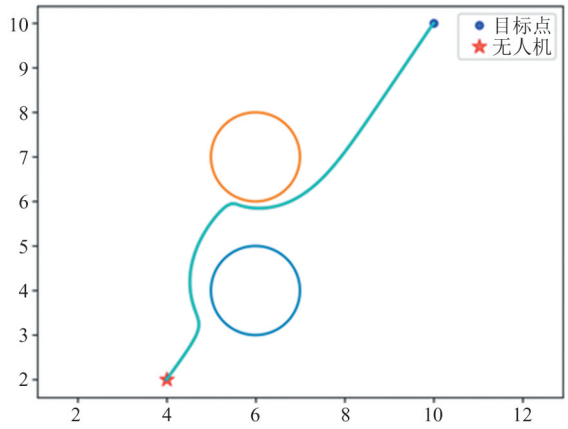

图 2　二维场景路径仿真图

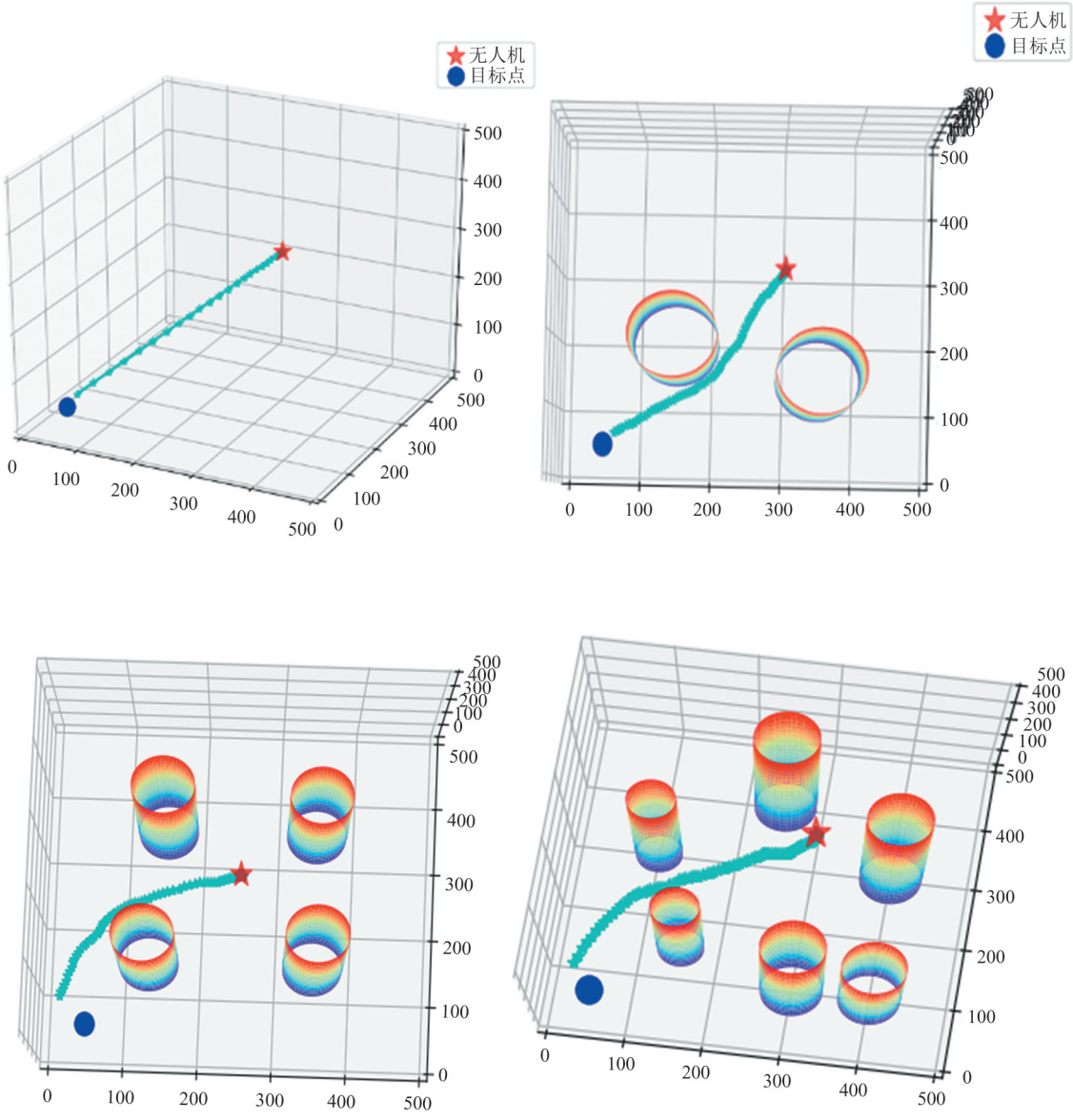

图 3　三维场景路径仿真图

（无人机起点：(300，300，300)，目标点：(50，50，50)）

图4 展示了算法 DDPG-UPP 在不同学习率下的性能评估。学习率决定着目标函数能否收敛到局部最小值以及何时收敛到最小值,合适的学习率能够使目标函数在合适的时间内收敛到局部最小值。从图 4 可以看出,当 Actor 网络学习率为 0.005、0.001,Critic 网络学习率为 0.01、0.002 时,随着训练次数的增多,UAV 在不断试错过程中获得的奖赏会逐渐稳定,这表明 UAV 学会到达目标点并满足约束条件的最优路径。同时,如图 5 所示,UAV 到达相同的目标点所需要的步数也逐渐减小,并稳定到固定值,UAV 随着学习次数的增多,能够更加准确地到达目标点。而对于 Actor 网络学习率为0.0005、0.0001,Critic 网络学习率为 0.001、0.0002 时,奖赏值及到达相同目标所需的步数虽然也收敛到定值,但相较于$a=0.005$、$c=0.01$与$a=0.001$、$c=0.002$的学习率,此时算法的性能并未达到最优,无人机学习到的路径并不是最优路径。另外,当学习率为 Actor = 0.01、Critic = 0.02 时,算法不收敛,无人机并不能学会到达目标的最优路径。因此,学习率的大小对算法 DDPG-UPP 的性能至关重要,能指导 UAV 在合适的时间找到最优路径。

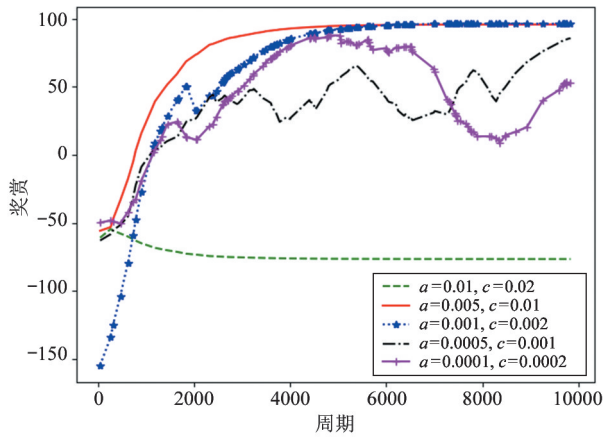

图 4　不同学习率下算法 DDPG-UPP 的性能对比图(Reward)

图 6、图 7 分别为不同算法下无人机路径规划奖赏以及到达相同目标所需步数的对比图。将本文提出的 DDPG-UPP 算法与算法 1、算法 2、算法 3 的性能比较,如图 6 所示,DDPG-UPP 算法用于 UAV 路径规划相较于算法 1、算法 2、算法 3 收敛较快且获得的奖赏值也明显高于其他 3 种算法,表明使用 DDPG-UPP 算法获得的路径在能耗及时间都是最少的。这是因为算法 2、算法 3 适用于离散动作空间,UAV 在进行训练前需将动作空间离散化,而对于 UAV 路径规划的动作空间,要想实现 UAV 更加自主、高效动作,其离散动作空间复杂化,且在每一次训练中,无人机只能在特定的几个方向中选择,大幅降低了无人机的灵活性;其次,对于算法 1,虽然 Actor-Critic 算法可用于连续动作空间,但由于 Actor 的行为取决于 Critic 的值,Critic 难收敛导致 Actor-Critic 算法很难收敛,尽管算法 1 融入了指针网络 Ptr-A* 以加快 Actor-Critic 算法的收敛,但相较于本文算法仍有很大差距。本文算法也采用 Actor-Critic 结构,但融入了深度 Q 网络(deep Q-network, DQN)的优势,既解决了算法 2 的空间离散问题,又区别于算法1、算法3中Actor的概率分布输出,而是以确

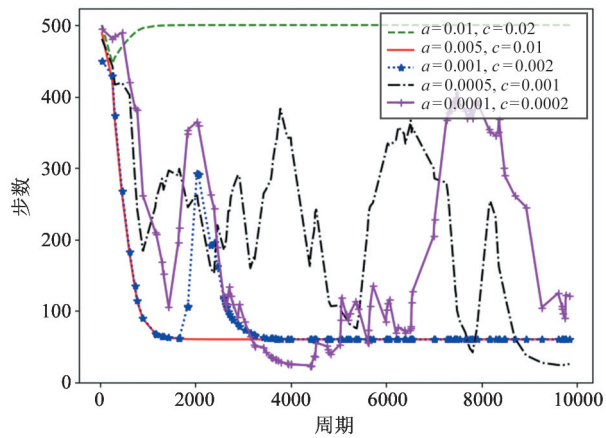

图 5　不同学习率下算法 DDPG-UPP 的性能对比图(Step)

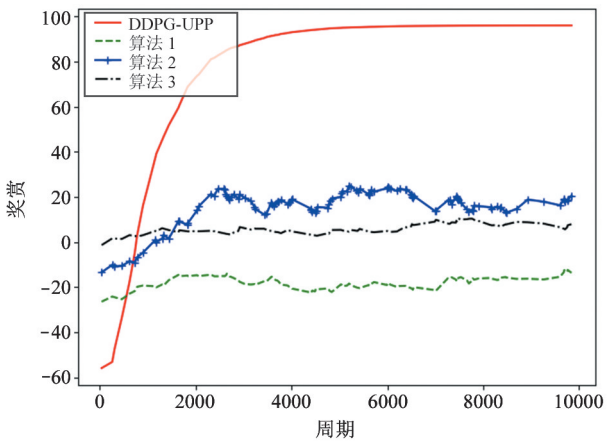

图 6　不同算法下无人机路径规划性能对比图(Reward)

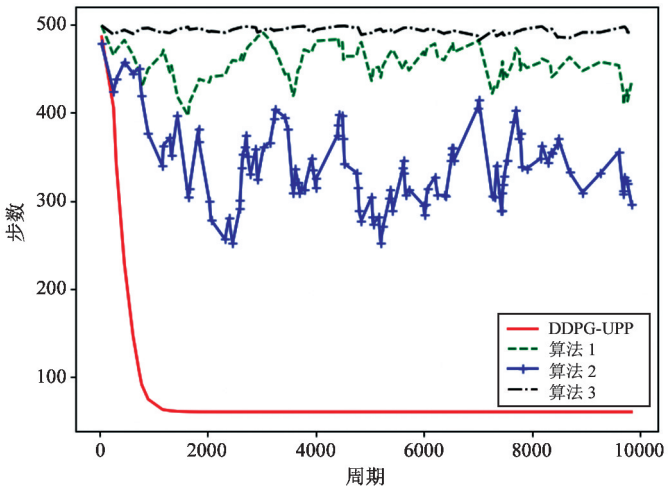

图 7 不同算法下无人机路径规划性能对比图(Step)

定性的策略输出加快了算法的收敛。因此,如图 6、图 7 所示,本文算法不仅能够使 UAV 更快获得到达目标的最优路径,而且使得无人机能耗及时间都是最小的,同时能在到达相同目标时使用更少步数。

图 8 为二维环境与三维环境下分别使用 DDPG-UPP 算法与算法 2 的性能对比图。首先,图 8 显示无论是二维环境还是三维环境,使用 DDPG-UPP 算法的性能都要优于使用算法 2。这是由于算法 2 虽然改变了 DQN 的模型结构,但仍需将动作空间离散化,而针对本文无人机飞行环境,则至少需要将动作空间离散为 6 个维度,在每一次试错中,相较于本文算法,算法 2 都增加了试错成本,同时也增加了计算复杂度,从而增加了无人机探索最佳路径的难度;其次,DDPG-UPP 算法在无障碍环境中的奖赏值要高于有障碍环境,且较有障碍环境更快收敛,这是因为

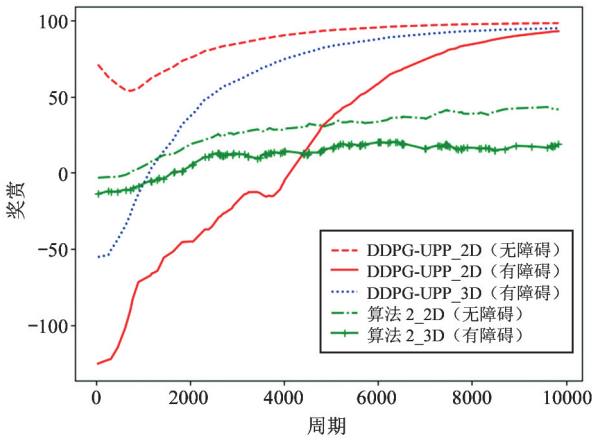

图 8 不同维度下无人机路径规划性能对比图(2D 与 3D)

环境中的障碍会在一定程度上阻碍无人机的探索,无人机需进行更多次尝试才能学习到最优路径;此外,对于本文算法,在同时考虑障碍物的环境下,在三维环境中的性能也要明显优于二维环境。综上,本文算法在三维环境避障路径选择中相较于其他算法具有更优的性能。

5 结 论

本文研究了一种三维复杂环境下无人机路径规划方法,提出一种无人机高度避障方法,引入偏度 δ 表示无人机与障碍物及目标用户的相对位置,使 UAV 能够更加自主、灵活地避开障碍,更加适应 UAV 实际工作环境。另外,考虑 UAV 动作空间的连续性,采用深度确定性策略梯度算法进行无人机路径规划。实验结果表明,本文算法能够克服传统算法需将动作离散化的弊端,增加了环境适应性。

参考文献

[1] YU H, LI G, ZHANG W, et al. The unmanned aerial vehicle benchmark: object detection, tracking and baseline [J]. *International Journal of Computer Vision*, 2020 (128): 1141-1159

[2] PEREIRA D S, MORAIS M R D, NASCIMENTO L B P, et al. Zigbee protocol-based communication network for multi-unmanned aerial vehicle networks [J]. *IEEE Access*, 2020, 8:57762-57771

[3] BASILICO N, CARPIN S. Deploying teams of heterogeneous UAVs in cooperative two-level surveillance missions [C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, Hamburg, Germany, 2015: 610-615

[4] KIM J, KIM S, JU C, et al. Unmanned aerial vehicles in agriculture: a review of perspective of platform, control, and applications [J]. *IEEE Access*, 2019, 7: 105100-105115

[5] LUO C, WANG M, ULLAH H, et al. Unmanned Aerial Vehicles for Disaster Management [M]. Singapore: Springer, 2019: 83-107

[6] AGGARWAL S, KUMAR N. Path planning techniques for unmanned aerial vehicles: a review, solutions, and challenges [J]. *Computer Communications*, 2020, 149: 270-299

[7] KUMAR P, GARG S, SINGH A, et al. MVO-based two-dimensional path planning scheme for providing quality of service in UAV environment [J]. *IEEE Internet of Things Journal*, 2018,5(3): 1698-1707

[8] MAJEED A, LEE S. A new coverage flight path planning

algorithm based on footprint sweep fitting for unmanned aerial vehicle navigation in urban environments [J]. *Applied Sciences*, 2019, 9(7):1470

[9] ALSHAWI I S, YAN L, WEI P, et al. Lifetime enhancement in wireless sensor networks using fuzzy approach and A-Star algorithm [J]. *IEEE Sensors Journal*, 2012, 12(10):3010-3018

[10] CHEN Y B, LUO G C, MEI Y S, et al. UAV path planning using artificial potential field method updated by optimal control theory[J]. *International Journal of Systems Science*, 2016, 47(6):1407-1420

[11] RADMANESH M, KUMAR M. Flight formation of UAVs in presence of moving obstacles using fast-dynamic mixed integer linear programming[J]. *Aerospace Science and Technology*, 2016, 50:149-160

[12] KOTHARI M, POSTLETHWAITE I. A probabilistically robust path planning algorithm for UAVs using rapidly-exploring random trees[J]. *Journal of Intelligent and Robotic Systems*, 2013, 71(2):231-253

[13] QU C, GAI W, ZHONG M, et al. A novel reinforcement learning based grey wolf optimizer algorithm for unmanned aerial vehicles (UAVs) path planning[J]. *Applied Soft Computing*, 2020, 89:106099

[14] 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1):1-27

[15] MACIEL-PEARSON B G, MARCHEGIANI L, AKCAY S, et al. Online deep reinforcement learning for autonomous UAV navigation and exploration of outdoor environments[EB/OL]. https://arxiv.org/pdf/1912.05684.pdf:arXiv,(2019-12-11),[2021-09-03]

[16] YAN C, XIANG X, WANG C. Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments[J]. *Journal of Intelligent and Robotic Systems*, 2020, 98(3-4):297-309

[17] BAYERIEIN H, KERRET P D, GESBERT D. Trajectory optimization for autonomous flying base station via reinforcement learning[C]//2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications, Kalamata, Greece, 2018:1-5

[18] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. https://arxiv.org/pdf/1509.02971.pdf:arXiv,(2019-07-05),[2021-09-03]

[19] SHARMA R K, GHOSE D. Collision avoidance between UAV clusters using swarm intelligence techniques[J]. *International Journal of System Science*, 2009, 40(5):521-538

[20] MATHERON G, PERRIN N, SIGAUD O. The problem with DDPG: understanding failures in deterministic environments with sparse rewards[EB/OL]. https://arxiv.org/pdf/1911.11679v1.pdf: arXiv, (2019-11-26), [2021-09-03]

[21] ZHANG M, ZHANG Y, GAO Z, et al. An improved DDPG and its application based on the double-layer BP neural network [J]. *IEEE Access*, 2020, 8: 177734-177744

[22] BOTAO Z, EBRAHIM B, HA H N, et al. UAV trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning [J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(9): 9540-9554

A DDPG algorithm to UAV path planning in 3D

SI Pengbo, WU Bing, YANG Ruizhe, LI Meng, SUN Yanhua
(College of Information and Communications Engineering, Faculty of Information Technology, Beijing University of Technology, Beijing 100124)

Abstract

Unmanned aerial vehicles (UAVs) have attracted widespread attention due to their high mobility and strong environmental adaptability, and have been used in military and civilian fields. This work studies the method of using the deep deterministic policy gradient (DDPG) algorithm to achieve UAV path planning in complicated environment. Firstly, establish a three-dimensional scene model and divide the drone mission process into three stages: flight, waiting, and communication. Secondly, a three-dimensional deviation degree is proposed to indicate the relative position of the drone, obstacles and target users, so as to improve the flight performance of the drone and effectiveness of obstacle avoidance. Finally, the deep deterministic policy gradient algorithm is used to plan the continuous flight movements of the UAV to reduce energy consumption and improve the quality of service (QoS), while avoiding obstacles and completing data transmission to users. The simulation experimental results show that the proposed scheme is effective under various parameter configurations, and it is better than existing algorithms.

Key words: unmanned aerial vehicle (UAV), three-dimensional scene, path planning, deep deterministic policy gradient (DDPG), avoid obstacles