http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0075

基于深度强化学习的固定翼无人机纵向控制

何海洋，赵振根*，孔飞
(南京航空航天大学 自动化学院，南京 210016)

摘 要：固定翼无人机(UAV)作为典型的非线性系统，其动态特性变得越来越复杂。传统的控制方法主要基于模型和经验设计，缺乏对复杂环境和任务的适应性。基于多维连续状态输入、多维连续动作输出的深度确定性策略梯度(DDPG)算法，设计了一种固定翼无人机的纵向飞行控制器，以多个时刻的速度、俯仰角跟踪误差及相关量作为控制器的输入，输出为升降舵舵偏角和发动机推力信号。为提高算法的学习效率，减轻稀疏奖励对算法学习的影响，奖励函数中除跟踪误差的密集惩罚项外，还引入了正值激励因子，当跟踪误差控制在一定范围内并快速跟踪目标时给予正值奖励。实现了从无人机状态到控制面的端到端控制，并使用比例-积分-微分(PID)控制器进行了变控制目标与模型参数摄动的飞行仿真对比，仿真结果表明，基于深度强化学习(DRL)算法构建的控制系统不仅能实现控制目标，还具备一定的泛化能力和鲁棒性，控制性能在部分情况下优于PID控制器。

关 键 词：深度确定性策略梯度；固定翼无人机；纵向控制；模型不确定性；稀疏奖励
中图分类号：V249.1
文献标志码：A **文章编号**：1001-5965(2026)04-1306-10

随着科技的飞速发展，无人机在军事侦察、环境监测、救灾物资输送等各个领域发挥着越来越重要的作用[1-2]。为了在复杂环境下完成各种任务，对无人机的控制系统提出了更高的要求。姿态控制是无人机控制系统中的一项关键技术，其性能直接影响到无人机的稳定性和可靠性[3]。为此，研究人员做了大量的研究工作，常用的控制方法有比例-积分-微分(proportional-integral-derivative, PID)控制[4]、H_∞控制[5]、反步控制[6]、滑模结构控制[7]和自适应控制[8]等。然而，传统的控制方法在处理无人机非线性、不确定性和外部干扰方面逐渐暴露出其局限性，为了使无人机能在极端的环境中保持稳定性，完成艰难而繁杂的工作，其控制系统需要具备强大的自适应能力，能够根据环境变化快速做出反应，并调整控制策略。

近年来，深度强化学习(deep reinforcement learning, DRL)理论不断发展，DRL算法也持续更新，因其在非线性系统建模和决策方面的优异性能，成为解决复杂控制问题的研究热点。通过结合深度学习与强化学习技术，DRL实现了策略空间的自动搜索和优化，具有较强的泛化能力和鲁棒性[9]。因此，许多科研人员在DRL与无人机的飞行控制器设计结合方面做了许多工作。文献[10-11]介绍了DRL的基本原理及其在无人机领域的控制和决策方面的应用，分析了将DRL方法应用于飞行器控制中存在的问题，并预测了未来发展方向。针对旋翼无人机，文献[12]提出了一种基于无模型强化学习的神经网络的低阶四旋翼控制算法，并探讨了该算法在四旋翼悬停和跟踪任务中的性能。文献[13]提出带积分补偿的改进深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法，仿真结果表明，积分补偿的加入能够有效消除四旋翼无

收稿日期：2024-02-01；**录用日期**：2024-03-29；**网络出版时间**：2024-04-28 10:27
网络出版地址：link.cnki.net/urlid/11.2625.V.20240426.1301.001
基金项目：国家自然科学基金(62233009,62003161)
*通信作者. E-mail：zhaozhengen@nuaa.edu.cn

引用格式：何海洋，赵振根，孔飞. 基于深度强化学习的固定翼无人机纵向控制[J]. 北京航空航天大学学报，2026，52(4)：1306-1315.
HE H Y，ZHAO Z G，KONG F. Longitudinal control of fixed-wing UAV based on deep reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52(4)：1306-1315 (in Chinese).

人机位置跟踪静差, 提高控制的准确性, 增强系统的稳定性。文献 [14]将经典控制器与强化学习策略集成, 使其线性组合, 仿真和实验结果表明, 该算法具有更快的收敛速度和更好的控制性能。文献 [15] 在小角度约束下, 将四旋翼动力学分解为 6 个子系统, 提出一种基于 DRL 的级联四旋翼飞行控制器。文献 [16] 在软演员-评论家 (soft actor-critic, SAC) 算法的基础上引入专家信息, 提出进化式软演员-评论家(evolution-based soft actor-critic, ESAC)算法, 以增强控制策略的易用性和扩展性。文献 [17] 利用专家经验对强化学习算法进行改进, 提出基于示教知识辅助的无人机强化学习控制算法, 用来解决强化学习在无人机控制应用中学习率低的问题。相比于上述引入专家信息或将 DRL 算法与传统方法相结合的做法, 文献 [18] 选择对奖励函数进行更细致的构造, 仿真结果表明, 该控制器能够引导平流层浮空器较好地实现变高度的跟踪控制。

针对固定翼无人机, 文献 [19] 表明, DRL 可以成功学习直接在原始非线性动力学上操作的固定翼无人机姿态控制, 只需要 3 min 的飞行数据, 并展示了与最先进的 ArduPlane PID 姿态控制器相当的性能。之后, 该团队使用近端策略优化 (proximal policy optimization, PPO) 算法设计了一种 DRL 控制器来处理非线性姿态控制问题, 使固定翼无人机能够从大量初始条件稳定到参考滚转、俯仰和空速值扩展飞行包线[20]。针对全尺寸固定翼飞行器, 文献 [21] 使用 DDPG 算法设计了从飞行状态到舵面/推力控制的端到端六自由度一体化智能控制器, 并通过引入偏航角误差作为控制器输入, 实现近零侧滑的稳定巡航飞行。从上述文献不难看出, DRL 的理论已经发展得较为成熟,且已经应用到许多无人机研究过程中, 但大部分研究对象为四旋翼无人机, 对于固定翼无人机的研究较少, 且 DRL 需要不断试错, 学习效率较低。为此, 本文提出基于 DDPG 算法的固定翼无人机 DRL 控制方法, 同时, 为了提高智能体的学习效率, 对智能体的状态空间、动作空间、奖励函数与神经网络结构进行设计并训练。通过仿真对比验证该控制器与 PID 控制器的控制效果、泛化能力和鲁棒性。

1　固定翼无人机建模

仿真对象为小型固定翼无人机 Aerosonde[22], 飞行器的坐标系和受力如图 1 所示。定义飞行器机体坐标系 $O_b x_b y_b z_b$ 及地面坐标系 $O_g x_g y_g z_g$。原点 O_b 位于无人机质心, 机体坐标系与机体固连, $O_b x_b$ 轴在无人机对称面内指向头部, $O_b y_b$ 轴垂直于飞行器对称面指向机身右方, $O_b z_b$ 轴在飞行器对称面内与 $O_b x_b$ 垂直指向机身下方。地面坐标系固定于地面, 原点 O_g 位于地面某点, $O_g x_g$ 轴在地平面内指向某一方向, $O_g y_g$ 轴垂直于地面并指向地心, $O_g z_g$ 轴位于地平面内并垂直于 $O_g x_g$ 轴, 其方向通过右手定则确定。

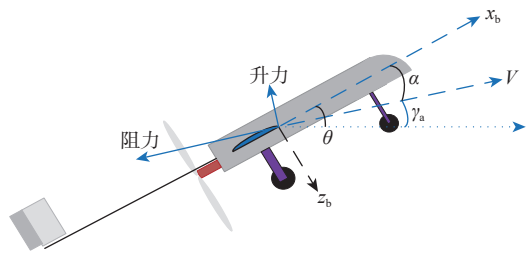


图 1　无人机系统坐标、角度与空气动力学示意图

Fig. 1　Illustration of UAV system coordinates, angles, and aerodynamics forces

针对水平无侧滑条件下的固定翼无人机, 对其连续状态下的非线性动力方程进行建模, 得到连续状态下的纵向非线性模型。无人机纵向状态主要有 4 个变量, 分别为速度 V、航迹角 γ_a、攻角 α 和俯仰角速度 q, 俯仰角 $\theta=\alpha+\gamma_a$。

$$\begin{cases}\dot{V}=\dfrac{T\cos\alpha-D}{m}-g\sin\gamma_a\\ \dot{\gamma}_a=\dfrac{T\sin\alpha+L}{mV}-\dfrac{g\cos\gamma_a}{V}\\ \dot{\alpha}=q-\dot{\gamma}_a\\ \dot{q}=\dfrac{M}{I_y}\end{cases}\tag{1}$$

式中: $I_y=1.135\ \text{kg}\cdot\text{m}^2$ 为转动惯量; $m=13.5\ \text{kg}$ 为无人机质量; $g=9.8\ \text{kg/m}^2$ 为重力加速度; L 为升力; T 为推力; D 为阻力; M 为俯仰力矩。

升力 L、推力 T、阻力 D 和俯仰力矩 M 都可以表示为无人机参数、状态和控制输入的相关函数, 计算公式如下:

$$\begin{cases}L=\dfrac{1}{2}\rho V^2 S C_L\\ D=\dfrac{1}{2}\rho V^2 S C_D\\ T=\dfrac{1}{2}\rho S_{\text{prop}} C_{\text{prop}}((K_{\text{motor}}\delta_T)^2-V^2)\\ M=\dfrac{1}{2}\rho V^2 S c C_M\end{cases}\tag{2}$$

式中: $\rho=1.268\,2\ \text{kg/m}^3$ 为空气密度; $S=0.55\ \text{m}^2$ 为机翼面积; $S_{\text{prop}}=0.202\,7\ \text{m}^2$ 为螺旋桨旋转面积; $c=0.189\,94\ \text{m}$ 为平均气动弦长; $C_{\text{prop}}=1.0$ 为螺旋桨平均弦长; $K_{\text{motor}}=80$ 为发动机常数; δ_T 为发动机推力信号。

气动参数为

$$\begin{cases} C_L = C_{L_0} + C_{L_\alpha}\alpha + C_{L_{\delta_e}}\delta_e \\ C_D = C_{D_p} + \dfrac{(C_{L_0} + C_{L_\alpha}\alpha)^2}{\pi e R_A} + C_{D_{\delta_e}}\delta_e \\ C_M = C_{M_0} + C_{M_\alpha}\alpha + C_{M_{\delta_e}}\delta_e \end{cases} \tag{3}$$

式中：C_L为升力系数；C_D为阻力系数；C_M为俯仰力矩系数；$C_{L_0}=0.28$为零迎角引起的升力系数；$C_{L_\alpha}=3.45$为迎角引起的升力系数；$C_{L_{\delta_e}}=-0.36$为升降舵偏转引起的升力系数；$C_{D_p}=0.0437$为寄生阻力引起的阻力系数；$C_{D_{\delta_e}}=0$为升降舵偏转引起的阻力系数；$C_{M_0}=-0.023\,38$为零状态俯仰力矩系数，$C_{M_\alpha}=-0.38$为迎角引起的力矩系数，$C_{M_{\delta_e}}=-0.5$为升降舵偏转引起的俯仰力矩系数；$e=0.9$为奥斯瓦尔德效率因子，用来修正校正理想情况下的机翼最小诱导阻力系数与实际诱导阻力系数之间的差异；$R_A=0.152$为机翼展弦比；δ_e为升降舵舵偏角信号$(-0.4\leqslant\delta_e\leqslant0.4)$，rad。

2 基于DRL的无人机纵向控制方案设计

2.1 DDPG算法原理与更新策略

强化学习本质上是一种试错学习算法，通过与环境的持续交互和信息互换来改进策略，其框架由智能体与马尔可夫决策过程(Markov decision process, MDP)组成：智能体在当前状态$\boldsymbol{s}_t$根据策略$\pi(\boldsymbol{s}_t)$输出动作$\boldsymbol{a}_t$，环境接收到该动作，通过奖励函数向智能体反馈奖励信号r_t，并进入下一个状态$\boldsymbol{s}_{t+1}$，智能体结合反馈信息调整策略生成下一个动作$\boldsymbol{a}_{t+1}$，持续学习直到获得使累计奖励最大的动作策略。MDP由状态空间集合、动作空间集合、状态转移函数和奖励函数组成。R表示为学习周期中奖励函数r的折扣求和，即

$$R=\sum_{i=1}^{T}\gamma^i r_i \tag{4}$$

式中：$\gamma\in(0,1]$为折扣系数；T为周期长度。

DDPG算法由深度Q网络(deep Q network, DQN)算法发展而来，用于解决“连续动作”问题，是一种采用Actor-Critic结构的离线策略DRL算法。该算法使用神经网络表示Actor-Critic结构。其中，Actor网络用来拟合确定性策略$\mu(\boldsymbol{s};\boldsymbol{\theta}^\mu)$，输入观察到的状态，并输出计算得到的动作，$\boldsymbol{\theta}^\mu$为Actor网络参数，而Critic网络拟合动作价值函数$Q(\boldsymbol{s},\mu(\boldsymbol{s};\boldsymbol{\theta}^\mu)|\boldsymbol{\theta}^Q)$，用以评价确定性策略$\mu(\boldsymbol{s};\boldsymbol{\theta}^\mu)$的优劣，$\boldsymbol{\theta}^Q$为Critic网络参数。由于DDPG算法训练的目的是找到使式(4)中的累积奖励最大的策略，其使用在策略$\mu(\boldsymbol{s};\boldsymbol{\theta}^\mu)$下执行动作得到的动作价值函数来近似给定$\boldsymbol{s}_t$、$\boldsymbol{a}_t$时$R_t$的期望值。

$$Q^\mu(\boldsymbol{s}_t,\boldsymbol{a}_t)=E^\mu[R_t|\boldsymbol{s}_t,\boldsymbol{a}_t] \tag{5}$$

式中：E为期望。

故Actor网络参数沿着目标函数$J(\mu_\theta)=E^\mu[R_t|\boldsymbol{s}_t,\boldsymbol{a}_t]$的梯度方向进行更新：

$$\nabla_{\theta^\mu}J(\mu_\theta)=E[\nabla_{\theta^\mu}\mu(\boldsymbol{s}|\boldsymbol{\theta}^\mu)_{\boldsymbol{s}=\boldsymbol{s}_t}\quad\nabla_{\boldsymbol{a}}Q^\mu(\boldsymbol{s},\boldsymbol{a}|\boldsymbol{\theta}^Q)_{\boldsymbol{s}=\boldsymbol{s}_t,\boldsymbol{a}=\mu(\boldsymbol{s}_t)}] \tag{6}$$

$$\boldsymbol{\theta}^\mu_{t+1}=\boldsymbol{\theta}^\mu_t+\alpha_\mu\nabla_{\theta^\mu}J(\mu_\theta) \tag{7}$$

式中：α_μ为Actor网络的学习率。

Critic网络通过采用时间差分算法来设定误差函数，通过最小化动作价值函数的估计值与预测值之间的误差构建损失函数$L(\boldsymbol{\theta}^Q)$来更新网络参数$\boldsymbol{\theta}^Q$：

$$y_t=r_{t+1}+\gamma Q'\left(\boldsymbol{s}_{t+1},\mu'\left(\boldsymbol{s}_{t+1}\mid\boldsymbol{\theta}^{\mu'}\right)\mid\boldsymbol{\theta}^{Q'}\right) \tag{8}$$

$$L(\boldsymbol{\theta}^Q)=E[(y_t-Q(\boldsymbol{s}_t,\boldsymbol{a}_t|\boldsymbol{\theta}^Q))^2] \tag{9}$$

$$\boldsymbol{\theta}^Q_{t+1}=\boldsymbol{\theta}^Q_t+\beta_Q(y_t-Q(\boldsymbol{s}_t,\boldsymbol{a}_t)\nabla_{\theta^Q}Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t|\boldsymbol{\theta}^Q\right) \tag{10}$$

式中：y_t为动作价值函数的估计值；β_Q为Critic网络的学习率。

为避免单个Critic网络在更新参数的同时也用于计算目标值，最终导致Critic网络容易发散的问题，DDPG算法借鉴DQN引入目标Actor网络$\mu'(\boldsymbol{s}|\boldsymbol{\theta}^{\mu'})$和目标Critic网络$Q'(\boldsymbol{s},\boldsymbol{a}|\boldsymbol{\theta}^{Q'})$，通过降低目标动作价值函数的更新速度来提高神经网络稳定性。目标网络的结构与对应的神经网络相同，对应的网络参数分别为$\boldsymbol{\theta}^{\mu'}$和$\boldsymbol{\theta}^{Q'}$，更新方式如下：

$$\begin{aligned}\boldsymbol{\theta}^{\mu'}&=\tau\boldsymbol{\theta}^{\mu}+(1-\tau)\boldsymbol{\theta}^{\mu'}\\ \boldsymbol{\theta}^{Q'}&=\tau\boldsymbol{\theta}^{Q}+(1-\tau)\boldsymbol{\theta}^{Q'}\end{aligned} \tag{11}$$

式中：τ为惯性更新率。

DDPG还引入经验回放，将经验数据$(\boldsymbol{s}_t,\boldsymbol{a}_t,r_t,\boldsymbol{s}_{t+1})$存储在经验回放池中，并通过mini-batch采样经验数据来消除连续样本数据的相关性，使数据满足独立同分布，从而减小参数更新的方差，提高收敛速度，且能够提高数据利用率。则式(6)～式(10)调整为

$$y_i=r_{i+1}+\gamma Q'\left(\boldsymbol{s}_{i+1},\mu'\left(\boldsymbol{s}_{i+1}\mid\boldsymbol{\theta}^{\mu'}\right)\mid\boldsymbol{\theta}^{Q'}\right) \tag{12}$$

$$L\left(\boldsymbol{\theta}^Q\right)=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-Q\left(\boldsymbol{s}_i,\boldsymbol{a}_i\mid\boldsymbol{\theta}^Q\right)\right)^2 \tag{13}$$

$$\nabla_{\theta^Q}L\left(\boldsymbol{\theta}^Q\right)=\frac{1}{N}\sum_{i=1}^{N}(y_i-Q(\boldsymbol{s}_i,\boldsymbol{a}_i))\nabla_{\theta^Q}Q(\boldsymbol{s}_i,\boldsymbol{a}_i) \tag{14}$$

$$\boldsymbol{\theta}^Q_{t+1}=\boldsymbol{\theta}^Q_t+\beta_Q\nabla_{\theta^Q}L\left(\boldsymbol{\theta}^Q\right) \tag{15}$$

$$\nabla_{\theta^\mu} J(\theta^\mu) \approx \frac{1}{N}\sum_i \nabla_a Q\left(s,a|\theta^Q\right)|_{s=s_i,a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i} \tag{16}$$

$$\theta^\mu_{t+1} = \theta^\mu_t + \alpha_\mu \nabla_{\theta^\mu} J(\theta^\mu) \tag{17}$$

式中: N为 mini-batch 样本大小; i为抽取的样本序号。

由于 DDPG 算法是确定性策略, 为确保其对环境的探索与利用, 在与环境交互时, Actor 网络的输出动作与 OU(Ornstein-Uhlenbeck) 噪声$\mathcal{N}$叠加后作为最后的控制信号。

$$a_t = \mu(s_t|\theta^\mu) + \mathcal{N} \tag{18}$$

2.2　基于 DDPG 的无人机纵向控制设计

本文设计的固定翼无人机飞行控制器飞行任务为跟踪目标俯仰角θ_c和目标速度V_c。为使 DDPG 算法能够训练并学习到符合期望的无人机控制策略, 结合控制对象与控制目标对 MDP 过程进行设计, 算法结构如图 2 所示。

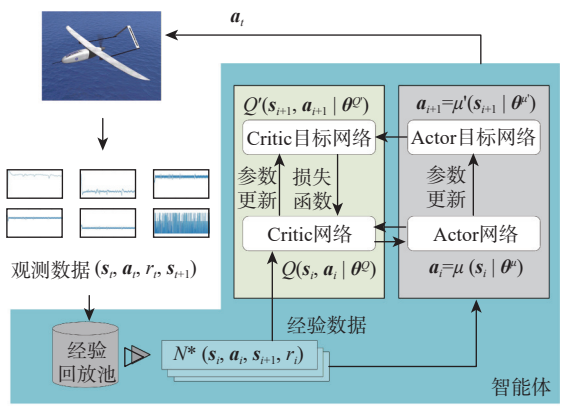


图 2　DDPG 算法结构

Fig. 2　DDPG algorithm structure

对于智能体所能观察到的状态空间, 若只使用当前时刻的无人机状态作为智能体的观测量, 由于缺乏控制目标的针对性, 以及神经网络在拟合动作价值函数时会存在误差, 会导致训练出的智能体在进行飞行控制时出现稳态误差及训练时间过长的问题。针对存在稳态误差的问题, 有研究引入 PID 中积分器的概念, 在观测状态中加入历史误差信息来提升控制器性能, 将过去时刻的误差进行累积作为补偿后的状态误差, 则输入策略网络中的状态包含了当前的状态误差和过去的累积误差, 如果稳态误差存在, 策略网络就会持续输出动作, 以减小稳态误差, 直到降为零[13]。

$$s_c^t = s_e^t + \beta\sum_{i=1}^{t} \lambda^{t-i} s_e^i \tag{19}$$

式中: s_c^t为 t 时刻补偿后的状态误差; s_e^t为 t 时刻的状态误差; s_e^i为 i 时刻补偿后的状态误差; β为系数; λ为权重因子, 用于调节过去时刻误差在累积误差中的比重。

故本文参考 PID 控制器中使用历史跟踪误差和误差变化量来减小稳态误差的理念, 将当前时刻与上一时刻的速度跟踪误差ΔV、速度误差变化量$\Delta\dot{V}$、俯仰角误差$\Delta\theta$、俯仰角误差变化量$\Delta\dot{\theta}$、俯仰角速度q共计 10 维作为智能体当前时刻的观测量:

$$\begin{aligned} s_t = [&\Delta V_t, \Delta\dot{V}_t, \Delta\theta_t, \Delta\dot{\theta}_t, q_t,\\ &\Delta V_{t-1}, \Delta\dot{V}_{t-1}, \Delta\theta_{t-1}, \Delta\dot{\theta}_{t-1}, q_{t-1}]^{\mathrm{T}} \end{aligned} \tag{20}$$

对于固定翼无人机的纵向平面, 控制输入为升降舵舵偏角和发动机推力信号, 因此, 智能体的动作输出为升降舵舵偏角δ_e和发动机推力信号δ_T:

$$a_t = [\delta_e, \delta_T]^{\mathrm{T}} \tag{21}$$

通过更细致地设置奖励函数, 可以提高控制器的性能, 减小跟踪时的稳态误差。针对无人机特性和控制目标, 奖励函数根据跟踪误差及跟踪速度进行设置和调整, 同时考虑控制信号的影响:

$$\begin{cases} r_1 = -\omega_V|\Delta V_t| - \omega_\theta|\Delta\theta_t| - \omega_q|q_t| \\ r_2 = \omega_1(|\Delta V_t| < |\Delta V_{t-1}|) + \\ \qquad \omega_2(|\Delta\theta_t| < |\Delta\theta_{t-1}|) + \omega_3(|q_t| < |q_{t-1}|) \\ r_3 = r_V^+ + r_\theta^+ \\ r_4 = -\omega_{\delta_e}|\delta_{e_{t-1}}| - \omega_{\delta_T}|\delta_{T_{t-1}}| \\ r_t = r_1 + r_2 + r_3 + r_4 \end{cases} \tag{22}$$

式中: ω 表示对应无人机变量在达成某种判定后的奖励系数。

即时奖励r_t中将所有的跟踪误差 (速度跟踪误差$|\Delta V_t|$、俯仰角跟踪误差$|\Delta\theta_t|$) 和俯仰角速度 ($|q_t|$) 设为惩罚项, 这些项乘以一定的比例系数构建奖励函数r_1, 当无人机状态接近控制目标时, 惩罚项的值逐渐减小直至接近零。

为提高算法的学习效率, 避免探索的过程中难以获得正奖励, 导致学习缓慢甚至无法进行学习的稀疏奖励问题, 同时满足飞行控制器的控制要求, 引入正值激励因子, 当 DDPG 输出的控制策略在控制无人机过程中达到激励因子的判断条件时, 会得到一定的正值奖励, 正值奖励激励智能体更快学习到哪些行为是有益的, 并倾向于更快收敛到最优或接近最优策略。

考虑到控制的快速性, 引入误差变化量及俯仰角速度在飞行过程中的变化情况构建奖励函数r_2, 即当前时刻的跟踪误差和俯仰角速度小于上一时刻时, 给出正值奖励, 值分别为ω_1、ω_2、ω_3, 否则, 奖励为 0, 当跟踪误差和俯仰角速度快速变小时, r_2给出的奖励变大, 从而促使 DDPG 智能体输出的控制信号能使无人机快速接近目标状态。接着, 引

入奖励函数r_3，其中，r_V^+为速度正值激励因子，当速度跟踪误差ΔV_t的绝对值在某个范围内时，给予正值奖励，否则，奖励为 0。类似地，r_θ^+为俯仰角正值激励因子，当俯仰角跟踪误差$\Delta\theta_t$的绝对值在某个范围内时，给予正值奖励，否则，奖励为 0。这些奖励激励因子在每个时间步均需要进行判断，当控制精度越高时，正值奖励越大，从而影响智能体改进控制策略。

最后，引入控制信号构建奖励函数r_4，以求在飞行过程中尽可能降低能量消耗和减小升降舵舵面的偏转幅度。

本文采用全连接前馈多隐层结构神经网络构建 DDPG 神经网络结构：在 DDPG 算法中，目标网络与对应的神经网络结构一致，Actor 网络有 5 层，其输入层包含 10 个神经元，中间有 3 个全连接的隐藏层，隐藏层每层都有 64 个神经元，前 4 层激活函数为线性整流 (ReLU) 函数，输出层有 2 个神经元，激活函数为双曲正切 (tanh) 函数，对该函数进行修改，使其更适配无人机对象的控制信号数值变化范围。Critic 网络有 5 层，其输入层包含 12 个神经元，输出层有 1 个神经元，中间有 3 个全连接的隐藏层，每个隐藏层包含 64 个神经元，激活函数都为 ReLU 函数。

3 DDPG 控制算法仿真与鲁棒性测试

3.1 智能体训练

基于 MATLAB R2021b/simulink 环境搭建了固定翼无人机纵向非线性模型，整个固定翼无人机智能飞行控制器的训练过程在一台搭载 i7-10700 CPU 和 NVIDIA Quadro P1000 GPU 的戴尔服务器上进行。

在训练中，给定的速度指令为$V_c = 10$ m/s,俯仰角指令为$\theta_c = 2°$，每个周期内选定一个不稳定状态作为无人机的初始状态，即在速度为 0.1 m/s、迎角为 0.01 rad、航迹角为 0.01 rad、俯仰角速度为 0 rad/s 附近选取。当 DDPG 算法学习的控制策略能使无人机从不稳定状态达到目标状态并能保持稳定，则认为策略是优秀的，其训练参数如表 1 所示。

表 1 DDPG 训练参数

Table 1 DDPG training parameters

Critic 学习率	Actor 学习率	优化器	批数量	经验回放池大小	惯性更新率	折扣系数	单次训练周期/个
0.001	0.000 5	Adam	128	1 000 000	0.001	0.98	1 000

注：一个单次训练周期为10 s。

经过多次训练，并对奖励函数参数进行调整，最终奖励函数中的参数设置如下：r_1中的参数设置为$\omega_V = 1, \omega_\theta = 1, \omega_q = 0.5$。俯仰角速度的惩罚因子小于其余两项的原因是避免俯仰角速度优先收敛到 0 时使无人机提前达到非控制目标的某个稳态点。r_2中的参数设置为$\omega_1 = 1.5, \omega_2 = 2, \omega_3 = 3$。分别在 1 m/s、0.1 m/s、0.01 m/s 时分别给出 3、10、25 的正值奖励，否则，奖励为 0。类似地，当俯仰角跟踪误差$\Delta\theta_t$的绝对值在 1°、0.1°、0.01°时，分别给出 5、10、35 的正值奖励，否则，奖励为 0。ω_{δ_e}和ω_{δ_T}的值均取−0.005。

为了验证加入的正值激励因子可以有效提高算法的学习效率，将奖励函数中的激励因子去除，其余设置与保留激励因子的算法相同，训练相同的回合数，进行控制效果对比，奖励变化如图 3 所示。可以看到，训练一定回合数后，奖励曲线基本上收敛。取训练了 600 回合的 2 个智能体分别构建飞行控制器进行仿真测试，控制结果如图 4 所示。从俯仰角跟踪和速度跟踪任务的表现来看，加入了激励因子的智能体已经可以有效跟踪俯仰角和速度，未加入激励因子的智能体并不能有效跟踪

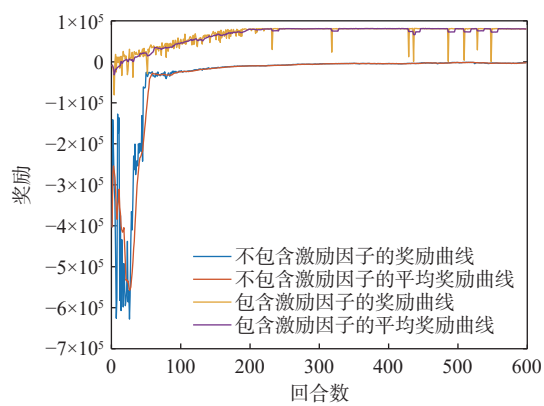


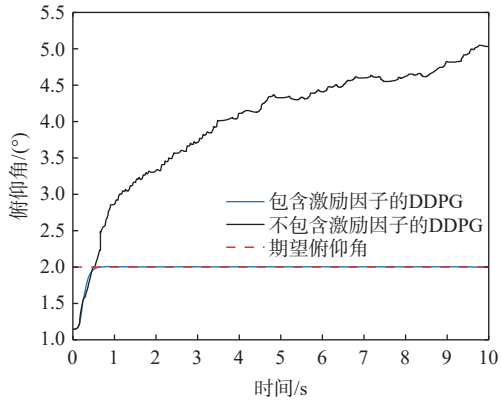
图 3 奖励变化对比曲线

Fig. 3 Comparison of reward variation curves

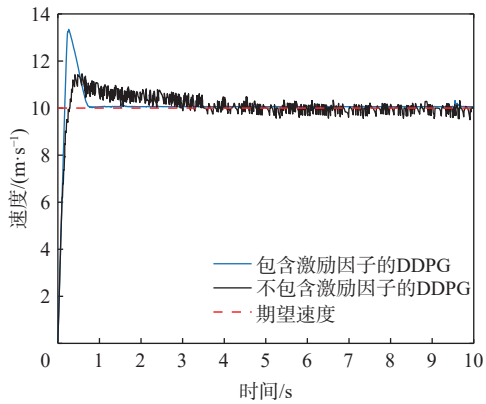
目标俯仰角，且速度跟踪振荡较大。可以看出，激励因子的加入可以有效提高算法的学习效率。

由于 DRL 并不能保证智能体性能可以持续改善，即可能出现训练次数增加、智能体学习的控制策略控制效果反而变差的情况，因此，本文选取在一定训练次数后，训练效果较好的智能体来构建飞行控制器，并进行仿真测试。

图 5 为使用 DDPG 智能体和 PID 方法构建的飞行控制器控制效果对比。可以看到，DDPG 控制



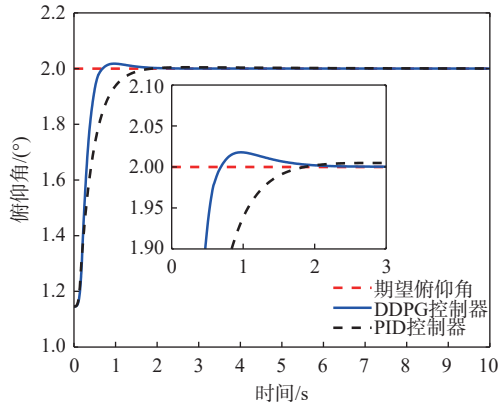
(a) 俯仰角跟踪



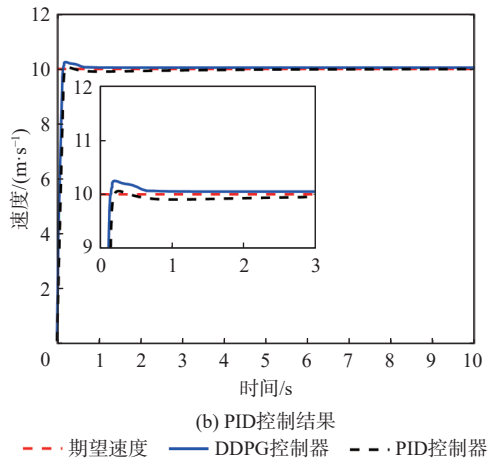
(b) 速度跟踪

图 4　DDPG 控制结果对比

Fig. 4　Comparison of DDPG control results



(a) DDPG控制结果



(b) PID控制结果

图 5　DDPG 与 PID 控制结果对比

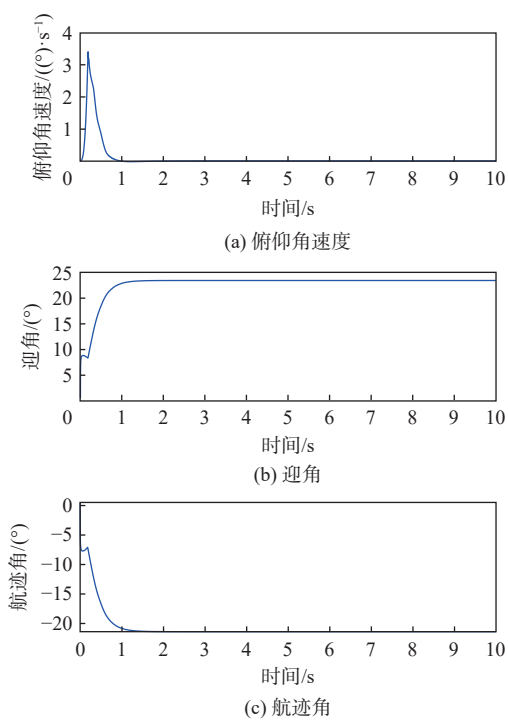
Fig. 5　Comparison of control results between DDPG and PID

的无人机在 0.67 s 达到目标俯仰角, 超调在 0.01°以内, 稳态误差在 0.000 44°以内; 而 PID 控制器在 1.86 s 达到目标俯仰角, 稳态误差为 0.000 5°。对于无人机速度跟踪任务, DDPG 控制器在 0.15 s 达到目标速度, 超调为 0.25 m/s, 稳态误差在 0.050 3 m/s 以内; PID 控制器在 0.2 s 达到目标速度, 超调为 0.06 m/s, 稳态误差为 0.003 m/s。从上述仿真结果可以看出, DDPG 控制器的跟踪速度更快, 在俯仰角跟踪任务中的控制精度比 PID 控制器更高。

从图 6 可以看出, DDPG 控制器在完成跟踪任务的同时, 其他状态量也达到了稳定状态。图 7 为 DDPG 控制器输出的控制信号。可以看出, 控制信号最终都趋于收敛, 且在收敛前的波动较小。

为了进一步测试 DRL 飞行控制器的性能, 进行控制目标变化仿真, 改变控制目标, 在 4 s 后将期望俯仰角从 2°跳变为 3°, 期望速度不变, 与 PID 控制器的对比仿真结果如图 8 所示。

在期望俯仰角仍为 2°时, 仿真结果与图 5 相同, DDPG 控制器相比于 PID 控制器具备更快的跟踪速度, 在第 4 s 期望俯仰角跳变至 3°后, DDPG 控制器在期望俯仰角跳变后仍然能进行有效跟踪, 无超调, 且稳态误差仅为 0.000 15°, 而 PID 控制器在期望俯仰角跳变后, 出现了 0.093 5°的超调, 且调节时间较长, 未能有效跟踪期望俯仰角。可以看出, DDPG 控制器在俯仰角跟踪任务的表现优于 PID



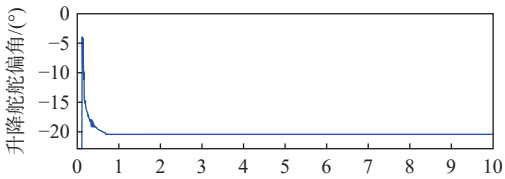
(a) 俯仰角速度

(b) 迎角

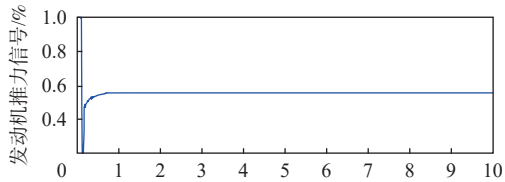
(c) 航迹角

图 6　DDPG 控制器控制结果

Fig. 6　Control results of DDPG controller



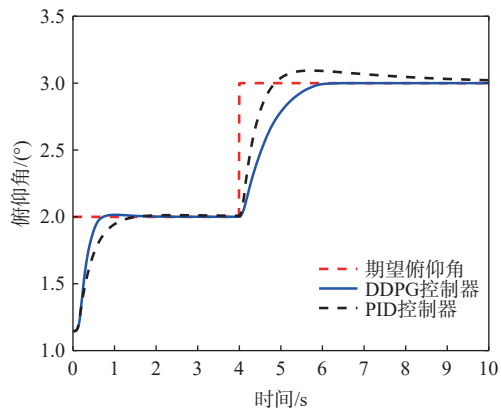
(a) 升降舵舵偏角



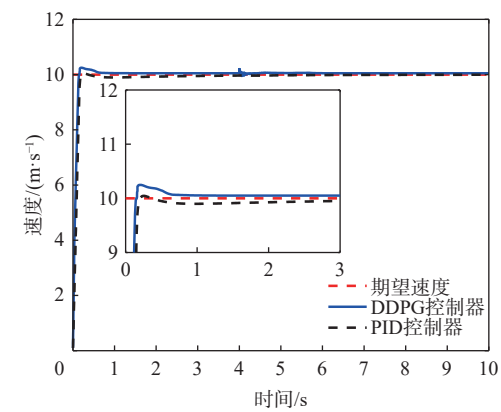
(b) 发动机推力信号

图 7 DDPG 输出的控制信号

Fig. 7 Control signal output of DDPG



(a) DDPG控制结果



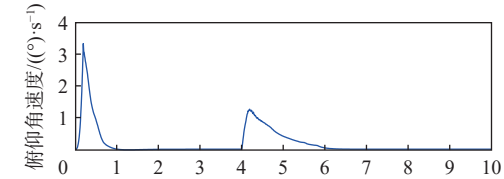
(b) PID控制结果

图 8 变期望俯仰角后 DDPG 与 PID 控制结果对比

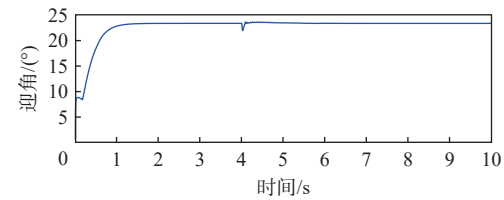
Fig. 8 Comparison of DDPG and PID control results after changing desired pitch angle

控制器。在速度跟踪任务中，DDPG 控制器相比于 PID 控制器，跟踪速度更快，在期望俯仰角跳变后虽然出现轻微扰动，但能快速稳定到期望速度。

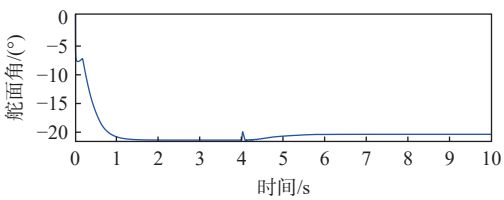
图 9 为 DDPG 控制器控制下无人机的其他状态量变化曲线。可以看到，其他状态量在仿真初始阶段稍微波动后达到稳态，在目标状态出现跳变后，其他状态也能在调整后重新收敛到稳定状态。



(a) 俯仰角速度



(b) 迎角



(c) 舵面角

图 9 变期望俯仰角的 DDPG 控制结果

Fig. 9 DDPG control results for changing desired pitch angle

3.2 参数不确定性测试

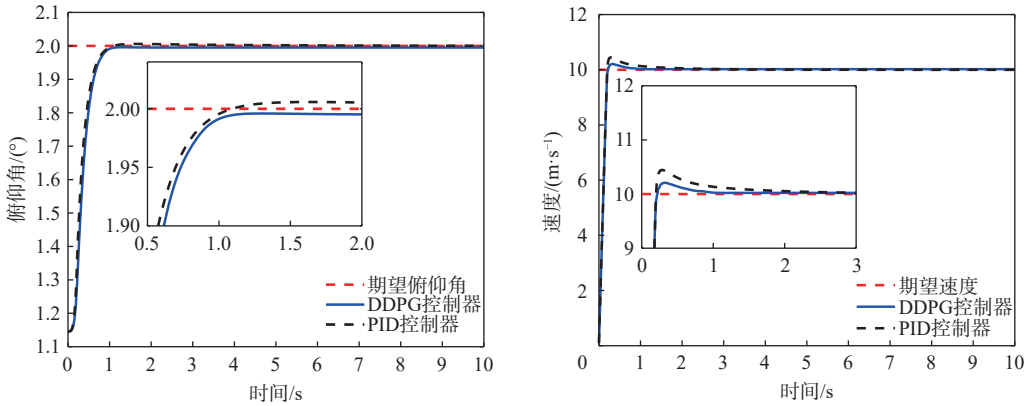
由于无人机模型中的参数不确定性影响无人机飞行任务的执行，在训练过程中未考虑模型不确定性的情况下，对质量、转动惯量、大气密度及气动参数分别引入一定比例的摄动，摄动比例如表 2 所示。测试了 DDPG 控制器的鲁棒性，仿真结果如图 10 所示，在无人机模型参数大比例摄动的情况下，DDPG 控制器在跟踪速度和俯仰角的超调都比 PID 控制器小，表明在当前情况下，DDPG 控制器性能优于 PID 控制器。

表 2 控制器对比仿真使用的参数不确定性

Table 2 Parameters uncertainties used for controllers comparison simulation

参数	不确定性/%
m	25
I_y	25
ρ	25
C_{L_α}	25
C_{L_0}	10
$C_{L_{\delta_e}}$	10
C_{m_0}	−20
C_{m_α}	−20
$C_{m_{\delta_e}}$	−20

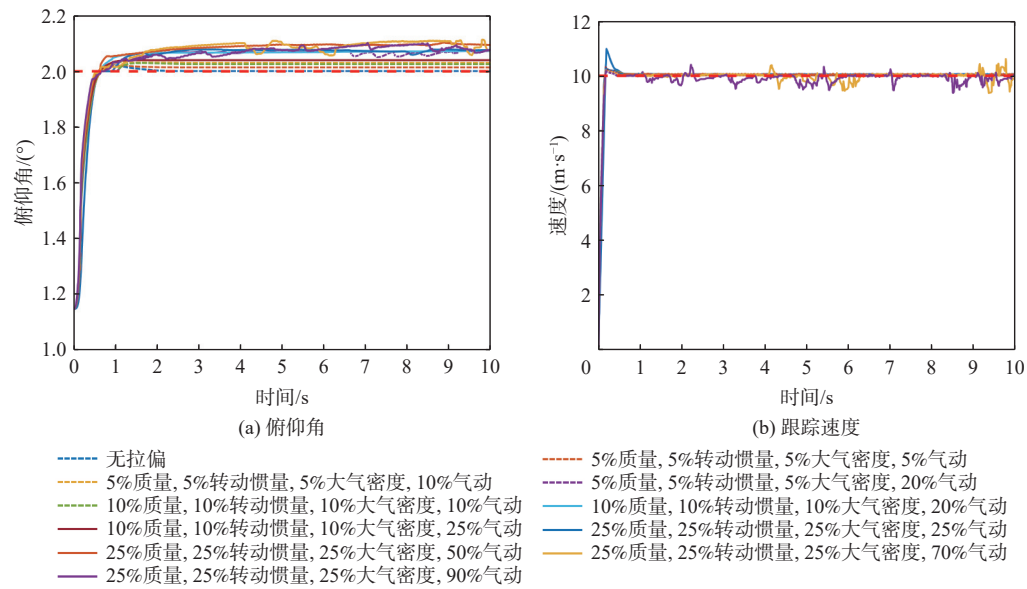
如图 11 所示，在不同比例的参数摄动情况下，DDPG 控制器仍然能够完成飞行任务，并且仍然保



(a) 俯仰角　　(b) 跟踪速度

图 10　DDPG 和 PID 在模型参数摄动下的控制结果对比

Fig. 10　Comparison of control results between DDPG and PID controllers under perturbed model parameters



(a) 俯仰角　　(b) 跟踪速度

图 11　DDPG 在模型参数摄动的控制结果

Fig. 11　Control results of DDPG with model parameter perturbation

持一定的控制精度。这表明训练出的控制器在面对一定范围内的模型参数变化时, 具备一定的鲁棒性。继续拉大摄动比例以测试该控制器的极限性能, 当质量、转动惯量和大气密度拉偏 25%、气动参数拉偏 50% 及以上时, 控制性能开始显著下降, 虽然仍能保持在跟踪目标附近, 但已经开始出现较大的稳态误差和振荡。

4　结　论

本文使用 DDPG 算法训练智能体, 进行固定翼无人机纵向控制器设计, 实现了从无人机状态到控制面的端到端控制。智能体训练和仿真结果表明:

1) 相对于将当前时刻的无人机状态作为智能体的观测状态, 将多时刻的跟踪误差信息作为输入能有效减小稳态误差。

2) 针对控制目标及控制要求, 根据跟踪误差、跟踪速度及无人机控制信号对奖励函数进行细致地设置可以在提高智能体学习的效率的同时, 使智能体学习到更优秀的控制策略。

3) 作为无模型的 DRL 算法, DDPG 智能体学习到的控制策略在完成训练的飞行任务时, 不仅具备比 PID 控制器更快的跟踪速度, 还能在控制目标变化和引入模型参数不确定性的情况下完成飞行任务, 且控制性能优于 PID 控制器, 体现了 DDPG 控制器具有一定的泛化能力和鲁棒性。

参考文献 (References)

[1] GHAMARI M, RANGEL P, MEHRUBEOGLU M, et al. Unmanned aerial vehicle communications for civil applications: a review[J]. IEEE Access, 2022, 10: 102492-102531.

[2] 符文星, 郭行, 闫杰. 智能无人飞行器技术发展趋势综述[J]. 无人系统技术, 2019, 2(4): 31-37.
FU W X, GUO H, YAN J. Overview on the technology development trend of intelligent unmanned aerial vehicle[J]. Unmanned Systems Technology, 2019, 2(4): 31-37(in Chinese).

[3] SHAN Y Q, WANG S, KONVISAROVA A, et al. Attitude control of flying wing UAV based on advanced ADRC[J]. IOP Conference Series: Materials Science and Engineering, 2019, 677(5): 137-142.

[4] YU Z Q, ZHANG Y M, JIANG B. PID-type fault-tolerant prescribed performance control of fixed-wing UAV[J]. Journal of Systems Engineering and Electronics, 2021, 32(5): 1053-1061.

[5] ZHAO X C, YUAN M N, CHENG P Y, et al. Robust H_∞/S-plane controller of longitudinal control for UAVs[J]. IEEE Access, 2019, 7: 91367-91374.

[6] ZHENG F Y, ZHEN Z Y, GONG H J. Observer-based backstepping longitudinal control for carrier-based UAV with actuator faults[J]. Journal of Systems Engineering and Electronics, 2017, 28(2): 322-377.

[7] SEOKWON L, JIHOON L, SOMANG L, et al. Sliding mode guidance and control for UAV carrier landing[J]. IEEE Transactions on Aerospace and Electronic Systems, 2019, 55(2): 951-966.

[8] ZHANG J L, ZHANG P, YAN J G. Distributed adaptive finite-time compensation control for UAV swarm with uncertain disturbances [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, 68(2): 829-841.

[9] YU K Y, JIN K, DENG X Y. Review of deep reinforcement learning[C]//Proceedings of the 2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2022: 41-48.

[10] 甄岩, 袁健全, 池庆玺, 等. 深度强化学习方法在飞行器控制中的应用研究[J]. 战术导弹技术, 2020(4): 112-118.
ZHEN Y, YUAN J Q, CHI Q X, et al. Research on application of deep reinforcement learning method in aircraft control[J]. Tactical Missile Technology, 2020(4): 112-118(in Chinese).

[11] 程林, 蒋方华, 李俊峰. 深度学习在飞行器动力学与控制中的应用研究综述[J]. 力学与实践, 2020, 42(3): 267-276.
CHENG L, JIANG F H, LI J F. A review on the applications of deep learning in aircraft dynamics and control[J]. Mechanics in Engineering, 2020, 42(3): 267-276(in Chinese).

[12] PI C H, HU K C, CHENG S, et al. Low-level autonomous control and tracking of quadrotor using reinforcement learning[J]. Control Engineering Practice, 2020, 95: 104222.

[13] 孙丹, 高东, 郑建华, 等. 引入积分补偿的四旋翼确定性策略梯度控制器[J]. 计算机工程与设计, 2023, 44(1): 255-261.
SUN D, GAO D, ZHENG J H, et al. Deterministic policy gradient controller with integral compensator for quadrotor[J]. Computer Engineering and Design, 2023, 44(1): 255-261(in Chinese).

[14] YOO J, JANG D, KIM H J, et al. Hybrid reinforcement learning control for a micro quadrotor flight[J]. IEEE Control Systems Letters, 2021, 5(2): 505-510.

[15] HAN H R, CHENG J, XI Z L, et al. Cascade flight control of quadrotors based on deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 11134-11141.

[16] 梁吉, 王立松, 黄昱洲, 等. 基于深度强化学习的四旋翼无人机自主控制方法[J]. 计算机科学, 2023, 50(增刊 2): 13-19.
LIANG J, WANG L S, HUANG Y Z, et al. Autonomous control method of quadrotor UAV based on deep reinforcement learning[J]. Computer Science, 2023, 50(Sup 2): 13-19(in Chinese).

[17] 孙丹, 高东, 郑建华, 等. 示教知识辅助的无人机强化学习控制算法[J]. 北京航空航天大学学报, 2023, 49(6): 1424-1433.
SUN D, GAO D, ZHENG J H, et al. UAV reinforcement learning control algorithm with demonstrations[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(6): 1424-1433(in Chinese).

[18] 张经伦, 杨希祥, 邓小龙, 等. 基于深度强化学习的平流层浮空器高度控制[J]. 北京航空航天大学学报, 2023, 49(8): 2062-2070.
ZHANG J L, YANG X X, DENG X L, et al. Altitude control of stratospheric aerostat based on deep reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(8): 2062-2070(in Chinese).

[19] BØHN E, COATES E M, REINHARDT D, et al. Data-efficient deep reinforcement learning for attitude control of fixed-wing UAVs: field experiments[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(3): 3168-3180.

[20] BOHN E, COATES E M, MOE S, et al. Deep reinforcement learning attitude control of fixed-wing UAVs using proximal policy optimization[C]//Proceedings of the 2019 International Conference on Unmanned Aircraft Systems. Piscataway: IEEE Press, 2019: 523-533.

[21] 章胜, 杜昕, 肖娟, 等. 基于深度强化学习的固定翼飞行器六自由度飞行智能控制[J]. 指挥与控制学报, 2022, 8(2): 179-188.
ZHANG S, DU X, XIAO J, et al. Fixed-wing aircraft 6-DOF flight control based on deep reinforcement learning[J]. Journal of Command and Control, 2022, 8(2): 179-188(in Chinese).

[22] BEARD R W, MCLAIN T W. Small unmanned aircraft: theory and practice[M]. Princeton: Princeton University Press, 2012: 43-52.

Longitudinal control of fixed-wing UAV based on deep reinforcement learning

HE Haiyang，ZHAO Zhengen*，KONG Fei

(College of Automation Engineering，Nanjing University of Aeronautics and Astronautics，Nanjing 210016，China)

Abstract: As a typical nonlinear system, the dynamic characteristics of a fixed-wing unmanned aerial vehicle (UAV) become more and more complex. Traditional control methods are mainly designed based on model and experience, and lack adaptability to complex environments and tasks. Based on the deep deterministic policy gradient (DDPG) algorithm of multi-dimensional continuous state input and multi-dimensional continuous action output, a longitudinal flight controller of a fixed-wing UAV was designed. The speed, pitch angle tracking errors, and related quantities of multiple moments were taken as the input of the controller, and the output was the elevator deflection and throttle setting signals. To improve the learning efficiency of the algorithm and mitigate the impact of sparse rewards on learning, the reward function introduced positive reward incentives in addition to the dense penalty for tracking errors. These positive rewards were given when the tracking error fell within a certain range and when the agent quickly reached the tracking target. Ultimately, end-to-end control from the longitudinal state of the UAV to the control surface was achieved, and under various control targets and model parameter perturbations, simulations were performed to compare the proportional-integral-derivative (PID) controller with a deep reinforcement learning-based control system. According to the simulation results, the deep reinforcement learning (DRL)-based control system may accomplish control goals and show some degree of robustness and generalization, with control performance sometimes outperforming the PID controller.

Keywords: deep deterministic policy gradient; fixed-wing UAV; longitudinal control; model uncertainties; sparse reward

Received：2024-02-01；**Accepted**：2024-03-29；**Published Online**：2024-04-28 10：27

URL：link.cnki.net/urlid/11.2625.V.20240426.1301.001

Foundation items：National Natural Science Foundation of China(62233009,62003161)

* **Corresponding author**. E-mail：zhaozhengen@nuaa.edu.cn