航 空 学 报
Acta Aeronautica et Astronautica Sinica

Apr. 25 2021 Vol. 42 No. 4
ISSN 1000-6893 CN 11-1929/V

http://hkxb.buaa.edu.cn hkxb@buaa.edu.cn

引用格式：相晓嘉，闫超，王菖，等. 基于深度强化学习的固定翼无人机编队协调控制方法[J]. 航空学报，2021，42(4)：524009. XIANG X J，YAN C，WANG C，et al. Coordination control method for fixed-wing UAV formation through deep reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica，2021，42(4)：524009 (in Chinese). doi：10.7527/S1000-6893.2020.24009

基于深度强化学习的固定翼无人机编队协调控制方法

相晓嘉，闫超*，王菖，尹栋

国防科技大学 智能科学学院，长沙 410073

摘 要：由于运动学的复杂性和环境的动态性，控制一组无人机遂行任务目前仍面临较大挑战。首先，以固定翼无人机为研究对象，考虑复杂动态环境的随机性和不确定性，提出了基于无模型深度强化学习的无人机编队协调控制方法。然后，为平衡探索和利用，将 ε-greedy 策略与模仿策略相结合，提出了 ε-imitation 动作选择策略；结合双重 Q 学习和竞争架构对 DQN(Deep Q-Network)算法进行改进，提出了 ID3QN(Imitative Dueling Double Deep Q-Network)算法以提高算法的学习效率。最后，构建高保真半实物仿真系统进行硬件在环仿真飞行实验，验证了所提算法的适应性和实用性。

关键词：固定翼无人机；无人机编队；协调控制；深度强化学习；神经网络

中图分类号：V249.1；V279 **文献标识码**：A **文章编号**：1000-6893(2021)04-524009-14

近年来，随着传感器技术、无线通信技术以及智能控制技术的不断发展与进步，无人机(Unmanned Aerial Vehicle，UAV)在军事和民用领域得到了广泛的应用，并取得了显著的成功[1]。但受限于平台功能少、有效载荷轻、感知范围小等固有缺陷，单架无人机在复杂环境下执行多样化任务仍面临较大困难[2]；而多架无人机组成协同编队能够有效弥补单机性能的不足，大幅提高系统的整体性能，在执行复杂作战任务时有着诸多优势[3]，如区域覆盖范围广、侦查和搜救成功率高等，作战效能远远高于各自为战的无人机。在可以预见的未来，随着战场环境和作战任务的日趋复杂，无人机编队将是执行作战任务的主要载体[4]。因此，无人机编队协调控制技术业已成为无人机系统技术领域的一个研究热点。

国内外学者针对该问题进行了广泛的研究。现有的解决方法，如模型预测控制[5]、一致性理论[6]等通常需要平台和扰动的精确模型进行控制率设计。但是，这一模型通常具有复杂、时变、非线性的特点，加之传感器误差、环境扰动等随机因素的影响，往往难以精确建模[7-8]。这严重限制了传统分析方法的适用范围。作为一种代替方法，应用无模型强化学习方法解决上述矛盾得到了越来越多的关注。

强化学习[9-10](Reinforcement Learning，RL)是机器学习领域的一个重要分支，主要用于解决序贯决策问题。强化学习任务通常可用马尔科夫决策过程(Markov Decision Process，MDP)来描述，其目标是在与环境的交互过程中，根据环境状态、动作和奖励学习一个最佳策略，使智能体(Agent)选择的动作能够从环境中获取最大的累积奖励。强化学习可以不依赖于环境模型，适用于未知环境中的决策控制问题，在机器人领域已取得了大量较为成功的应用，如路径规划[11-12]、

收稿日期：2020-03-24；退修日期：2020-05-18；录用日期：2020-06-30；网络出版时间：2020-07-07 11:03
网络出版地址：http://hkxb.buaa.edu.cn/CN/html/20210429.html
基金项目：国家自然科学基金(61906203)；西北工业大学无人机特种技术重点实验室基金(614230110080817)
* 通信作者. E-mail：yanchao17@nudt.edu.cn

导航避障[13-14]等。

目前，已有研究人员将强化学习融入其编队协调控制问题的解决方案中，并在仿真环境下对方案的可行性和有效性进行了初步的验证。强化学习在协调控制中的应用研究最早由 Tomimasu 等[15]开展，在该仿真研究中，Agent 采用 Q 学习算法和势场力方法学习聚集策略。不久之后，Morihiro 等[16]基于 Q 学习算法提出了一种多智能体自组织群集行为控制框架。仿真试验表明，Agent 在完成群集任务的同时，也表现出了反捕食行为以躲避捕食者。近年来，La 等[17-18]相继发布多项有关集群协调控制的研究成果，该团队提出了一种将强化学习和群集控制相结合的混合系统，并通过仿真和实验验证了系统的可扩展性和有效性。该系统由低层集群控制器和高层 RL 模块组成，这一结合方式使系统能在保持网络拓扑和连通性的同时躲避捕食者。混合系统中的 RL 模块采用 Q 学习算法，并通过共享 Q 表的方式实现分布式合作学习。试验结果表明，该方式可加速学习过程，并能获取更高的累积奖励。Wang 等[19]基于深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)算法，提出一种无人机编队协调控制算法，使无人机能够在大规模复杂环境中以完全分散的方式聚集并执行导航任务。

上述应用均采用质点 Agent 模型，所得控制方案仅适用于旋翼无人机。与旋翼无人机不同，由于固定翼无人机飞行动力学的非完整约束，固定翼无人机编队协调控制更加复杂，需要采用有别于旋翼机的控制策略与方法。此外，固定翼无人机更易受空速、侧风等环境扰动的影响，在动态不确定环境中学习到的策略会随着环境的变化而变化，导致强化学习算法难以收敛。到目前为止，将强化学习算法应用于固定翼无人机编队协调控制中的研究成果依然较少。

Hung 等[8,20]对该问题进行了初步的研究：2015 年，其在无模型强化学习的背景下，研究了小型固定翼无人机在非平稳环境下的聚集问题[20]；该研究采用变学习率 Dyna-$Q(\lambda)$ 算法学习 Leader-Follower 拓扑下的协调控制策略；仿真结果表明，所提变学习率方法具有更快的收敛速度；此外，所提方法还通过学习环境模型、并用规划的方式产生大量的模拟经验提高采样效率、加快学习过程。2017 年，Hung 和 Givigi 又在此基础上进一步提出了面向随机环境的无人机群集 Q 学习方法[8]；该研究以小型固定翼无人机为研究对象，基于无模型 RL 提出了固定翼无人机协调控制框架；在该框架中，Agent 采用变学习速率 $Q(\lambda)$ 算法在 Leader-Follower 拓扑中学习群集策略，并对抗环境的随机扰动；非平稳环境中的仿真试验验证了算法的可行性。

上述基于强化学习的固定翼无人机编队协调控制方法仍有一些问题尚未得到妥善解决：为解决维度灾难问题，Hung 等[8,20]将状态空间离散化以缩减状态空间的维度。这种处理方式虽然降低了问题的求解难度，但却未必十分合理。此外，Hung 等[8,20]仅在数值仿真环境对算法进行了初步的验证，所提算法的实用性和泛化性仍需进一步验证。

本文在 Hung 等[8,20]的研究基础上，聚焦动态不确定环境下固定翼无人机编队协调控制问题，基于深度强化学习算法构建端到端协调控制框架，实现多架无人僚机自主跟随长机组成编队协同飞行。首先，将 ε-greedy 策略与模仿策略相结合，提出 ε-imitation 动作选择策略以更好地平衡探索和利用；然后，结合双重 Q 学习和竞争架构对深度 Q 网络(Deep Q-Network，DQN)算法进行改进，提出 ID3QN(Imitative Dueling Double Deep Q-Network)协调控制算法以提高学习效率；最后，构建高保真半实物仿真系统验证算法的有效性和可迁移性。

1 背景介绍

1.1 强化学习

在强化学习中，智能体以试错的方式不断地与环境进行交互，旨在学习一个最佳策略，使得其从环境中获取的累积奖励达到最大[21]。强化学习问题可用 MDP 框架形式化描述。通常情况下，MDP 可用一个四元组 $(S,A,P(s,s',a),R(s,s',a))$ 定义，其中 S 表示状态空间；A 表示动作空间；$P(s,s',a)$ 表示状态转移概率函数(模型)，该模型定义了智能体执行动作 $a\in A$ 后，环境状态 $s\in S$ 转移到新状态 $s'\in S$ 的概率；$R(s,s',a)$ 表示

回报函数，其含义为智能体执行动作 $a\in A$ 后，环境状态 $s\in S$ 转移到新状态 $s'\in S$ 所带来的奖励。

在智能体与环境交互中的每一时间步 t，智能体观测环境状态为 s_t，进而根据策略 $\pi(a_t|s_t)$ 从动作空间 A 中选择动作 a_t。执行动作 a_t 后，环境状态以 $P(s_{t+1}|s_t,a_t)$ 的概率转移到新状态 s_{t+1}，并将回报值 r_t 反馈给智能体。智能体的目标在于学习一个最优策略 $\pi^*:S\rightarrow A$，即状态空间到动作空间的映射，以最大化期望折扣回报 R_t：

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{1}$$

式中：T 为终止时刻；γ 为折扣因子，用于平衡未来回报对累积回报的影响，$0\leqslant\gamma\leqslant 1$；$r_t$ 表示 t 时刻的立即回报。

1.2 Q 学习与深度 Q 网络

Q 学习(Q-learning)算法是强化学习领域最为经典且最为重要的算法之一，是由 Watkins 和 Dayan[22]提出的一种无模型(model-free)异策略(off-policy)的强化学习算法。该算法定义了 Q 值函数(Q-value)，并使用如式(2)和式(3)所示的更新规则迭代优化 Q 值函数：

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\delta_t \tag{2}$$

$$\delta_t=r_{t+1}+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t) \tag{3}$$

式中：δ_t 为 TD(Temporal-Difference)误差；s_t 为当前状态；a_t 为当前动作；s_{t+1} 为执行 a_t 后的环境状态；r_{t+1} 为立即回报值；α 为学习率，$0<\alpha<1$。

Q 值函数一旦确定，即可根据 Q 值函数确定最优策略：智能体以贪婪策略选择动作，即在每一时间步选择最大 Q 值定义的动作。Q 学习算法实现简单、应用广泛，但依然面临“维度灾难”的问题。该算法通常以表格的形式存储 Q 值，并不适用于高维或连续状态空间中的强化学习问题。

为解决“维度灾难”问题，利用深度神经网络(Deep Neural Network，DNN)作为函数逼近器估计 Q 值成为一种替代方案。Mnih 等[23]将卷积神经网络(Convolutional Neural Network，CNN)和经验回放技术引入 Q 学习算法，提出 DQN 算法，在 Atari 游戏中达到了人类玩家的水平。较之于 Q 学习算法，DQN 除了使用 CNN 作为函数逼近器并引入经验回放技术提高训练效率外，还设置单独的目标网络来产生目标 Q 值，以提高算法的稳定性[23]：

$$y_t^{\mathrm{DQN}}=r_{t+1}+\gamma\max_{a'}Q(s_{t+1},a';\theta^-) \tag{4}$$

式中：y_t^{DQN} 为 DQN 的目标 Q 值；θ^- 为 DQN 目标网络的参数。

DQN 通过最小化损失函数

$$L(\theta)=\mathbb{E}[(y_t^{\mathrm{DQN}}-Q(s_t,a_t|\theta))^2] \tag{5}$$

即主网络输出的估计 Q 值与目标网络输出的目标 Q 值之差来实时更新主网络参数 θ。与主网络实时更新参数不同，目标网络参数每隔若干时间步更新一次。具体而言，每隔 N 时间步，将主网络参数复制给目标网络，从而完成目标网络参数 θ^- 的更新。

1.3 双重 Q 学习与竞争架构

DQN 使用单独的目标网络产生 Q 值。尽管该技巧降低了预测 Q 值(主网络输出)与目标 Q 值(目标网络输出)之间的相关性，在一定程度上缓解神经网络近似值函数时出现的不稳定问题，但 Q 值“过估计”[24]的问题仍然没有得到解决。为更好地分析这一问题，将式(4)展开，有

$$y_t^{\mathrm{DQN}}=r_{t+1}+\gamma Q(s_{t+1},\arg\max_{a'}Q(s_{t+1},a';\theta^-);\theta^-) \tag{6}$$

显然，DQN 的 max 操作使用相同的值函数(同一套参数 θ^-)进行动作选择和动作评估。这极易导致过高地估计 Q 值。为解决这一问题，Van Hasselt 等[25]提出了双重 DQN 算法(Double DQN，DDQN)。该算法使用两个不同的值函数(两套参数)解耦动作选择与策略评估。DDQN 的目标 Q 值可表示为

$$y_t^{\mathrm{DDQN}}=r_{t+1}+\gamma Q(s_{t+1},\arg\max_{a'}Q(s_{t+1},a';\theta);\theta^-) \tag{7}$$

式中：θ 为主网络参数，用于选择最优动作；θ^- 为目标网络的参数，用于评估该动作的价值。

除目标 Q 值的形式不同外，DDQN 均与 DQN 保持一致。Atari 游戏中的实验结果表明，DDQN 能够更精确地估计 Q 值，获得更稳定有效的策略[25]。

DQN 与 DDQN 均是直接利用 DNN 近似状态-动作值函数，一旦给定当前状态，DNN 将评估所有状态-动作对的 Q 值。然而，对于某些状态

而言，并没有必要估计每个动作选择的价值。受此启发，Wang 等[26]提出使用竞争架构(Dueling Architecture)进一步提升 DQN 的性能。该架构构建了两个支路的全连接网络，分别用于逼近状态值函数和动作优势函数(Advantage Function)，最后通过特殊的“聚合”操作将二者组合起来，从而得到每个有效动作的 Q 值。假设 V 代表一条支路近似得到的状态值函数，$\mathcal{A}$ 表示另一支路近似得到的动作优势函数，则上述聚合操作可表示为

$$Q(s,a)=V(s)+\left(\mathcal{A}(s,a)-\frac{1}{|A|}\sum_{a'}\mathcal{A}(s,a')\right) \tag{8}$$

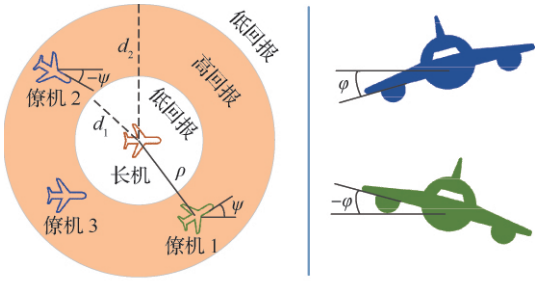
式中：$V(s)$为状态值函数；$|A|$为动作空间 A 的维度。

竞争架构可以简便地融入 DQN 或 DDQN 算法中。实验结果表明，基于竞争架构的 DQN 算法能够获得更好的结果[26]。

2 问题描述

在想定的协调控制场景中，无人机编队采用 Leader-Follower 拓扑，即一架长机带领若干架僚机组成编队遂行任务。长机的控制策略由飞行员根据具体任务类型(跟踪、侦察等)和战场态势确定。长机通过通信链路将自身位置与姿态信息广播给僚机，僚机需要根据机载传感器感知到的自身状态信息和接收到的长机状态信息，实时选择最佳的控制指令(如滚转角)。假设僚机在不同固定高度层飞行，故不必考虑飞机之间的避碰问题[8,20]，因此不同僚机可使用相同的控制策略。每一架僚机均配备有自驾仪，每隔 1 s[8,20]，控制策略根据当前系统状态输出新的控制指令，并发送给自驾仪，自驾仪使用 PID 控制器完成控制指令的底层闭环控制。

目标是让僚机在无任何先验知识的情况下，学习一种自主跟随长机编队飞行的控制策略。该策略能够根据获取的自身及长机的状态信息，确定当前给定状态的最佳滚转角设定值(自驾仪据此设定值完成闭环控制)，维持僚机与长机之间合理的位置关系(即僚机在以长机为中心的圆环内，如图 1 所示)，以实现 Leader-Follower 拓扑下的无人机编队协调控制。



(a) 俯视图　　(b) 正视图

图 1 长机与僚机期望位置关系

Fig. 1 Positional relationship between leader and followers

2.1 无人机运动学模型

试错学习是无模型强化学习重要的特征之一。由于无人机的特殊性，在真实环境中进行试错是不现实的，且在高保真的仿真环境下进行学习亦需要花费大量的时间。为提高学习的效率，思路为根据真实飞机运动学的经验特性，考虑环境扰动建立无人机运动学数值模型，并以此为基础应用深度强化学习方法学习无人机编队的协调控制策略，进而将该策略应用(迁移)到真实世界。

在真实世界，无人机运动学通常由六自由度模型描述。考虑到无人机保持定高飞行，该模型可简化至四自由度。为了弥补简化带来的损失，同时考虑环境扰动的影响，故而在滚转、空速等各个子状态引入随机性[27]，所得随机无人机运动学模型为

$$\xi=\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}x\\y\\\psi\\\varphi\end{bmatrix}=\begin{bmatrix}v\cos\psi+\eta_x\\v\sin\psi+\eta_y\\-(\alpha_g/v)\tan\varphi+\eta_\psi\\f(\varphi,\varphi_d)\end{bmatrix} \tag{9}$$

式中：$\xi:=[x,y,\psi,\varphi]$为无人机状态，其中(x,y)为 x-y 平面位置，ψ 为航向角，φ 为滚转角；α_g 为重力加速度；v 为无人机速度，服从均值为 $\bar{v}$ 方差为 σ_v 的正态分布 $\mathbb{N}(\bar{v},\sigma_v^2)$；$\eta_x$、$\eta_y$ 和 η_ψ 分别为平面位置和航向角的扰动项，分别服从均值为 $\bar{\eta}_x$、$\bar{\eta}_y$、$\bar{\eta}_\psi$，方差为 σ_x、σ_y、σ_ψ 的正态分布 $\mathbb{N}(\bar{\eta}_x,\sigma_x^2)$、$\mathbb{N}(\bar{\eta}_y,\sigma_y^2)$ 和 $\mathbb{N}(\bar{\eta}_\psi,\sigma_\psi^2)$，用于模拟无人机位置和航向因环境因素而产生的扰动；$f(\varphi,\varphi_d)$ 为期望滚转角 φ_d(输入)与实际滚转角 φ(响应)之间的

关系。

使用二阶系统响应模拟无人机滚转通道的动态响应[8,20]，并引入随机项使得该响应更具真实性。假定滚转通道二阶系统的无阻尼自然频率 ω_n 和阻尼系数 ζ 分别服从均值为 $\bar{\omega}_n$、$\bar{\zeta}$，方差为 σ_ω、σ_ζ 的正态分布 $\mathbb{N}(\bar{\omega}_n,\sigma_\omega^2)$ 和 $\mathbb{N}(\bar{\zeta},\sigma_\zeta^2)$，所需参数可根据自驾仪滚转指令的实际响应情况确定。

由于随机性的影响，无人机在同一初始状态下执行相同动作会产生不同的终止状态。如图 2 所示，初始时刻，无人机位于原点（$x=0$，$y=0$），并朝向 $+x$ 方向（$\psi=0$），执行同一控制指令后，无人机可能位于完全不同的位置。这说明所建运动学模型中引入的随机项能够模拟真实世界的随机性。

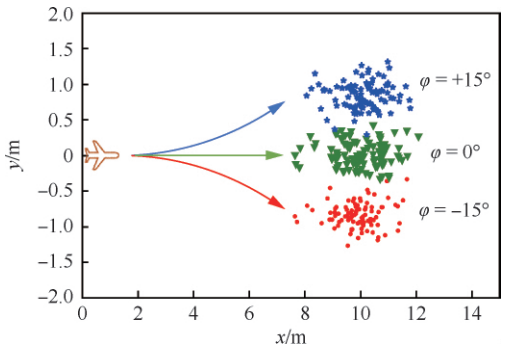


图 2 随机性对无人机状态影响

Fig. 2 Collection of possible resulting UAV states due to stochasticity

2.2 协调控制 MDP 模型

在无模型强化学习的背景下，将无人机编队协调控制问题建模为马尔可夫决策过程。依次对该模型的 3 个要素，即状态表示、动作空间和回报函数进行定义。

2.2.1 状态表示

由式(9)可知，无人机的状态可以通过四维数组 $\xi:=[x,y,\psi,\varphi]$ 表示。在 Leader-Follower 拓扑下的编队协调控制问题中，长机与僚机之间的相对关系（如距离、航向差等）对于控制策略的制定有着至关重要的影响。假定 $\xi_l:=[x_l,y_l,\psi_l,\varphi_l]$ 代表长机状态，$\xi_f:=[x_f,y_f,\psi_f,\varphi_f]$ 代表僚机状态，若定义系统联合状态为 $\boldsymbol{s}:=[s_1,s_2,s_3,s_4,s_5,s_6]$，则

$$\begin{cases}\begin{bmatrix}s_1\\s_2\end{bmatrix}=\begin{bmatrix}\cos\psi_l & \sin\psi_l\\-\sin\psi_l & \cos\psi_l\end{bmatrix}\begin{bmatrix}x_f-x_l\\y_f-y_l\end{bmatrix}\\ s_3=\psi_f-\psi_l\\ s_4=\varphi_f\\ s_5=\varphi_l\\ s_6=\varphi_d^l\end{cases}\tag{10}$$

式中：(s_1,s_2) 为僚机相对于长机的平面位置；s_3 为僚机与长机航向角之差；φ_d^l 为长机的期望滚转角。在实际应用中，长机的滚转指令是飞行员根据任务和战场态势输入确定的。为了让模型获得对各种可能输入的适应性，在训练时采用随机函数生成长机的滚转指令，以增加系统的不确定性。

需要指出的是，不同于文献[8,20]，本文没有对状态空间进行离散化以简化问题，而是直接在连续状态空间中求解无人机编队协调控制问题。

2.2.2 动作空间

如前所述，无人机的操控通过改变滚转角设定值实现。控制策略每隔 1 s 更新一次滚转指令，间隔时间内由自驾仪完成底层闭环控制。考虑到无人机的最大加速度，并避免滚转角的剧烈变化影响无人机的安全飞行，定义滚转动作空间 $a\in A$ 为

$$A:=\{-a_{\max}^r,0,+a_{\max}^r\}\tag{11}$$

式中：$a_{\max}^r$ 为最大候选滚转动作。若僚机当前滚转角为 φ，则下一时刻期望滚转角 φ_d 为

$$\varphi_d=\begin{cases}r_{bd} & \varphi+a>r_{bd}\\ -r_{bd} & \varphi+a<-r_{bd}\\ \varphi+a & \text{otherwise}\end{cases}\tag{12}$$

式中：a 为选定的滚转动作；$[-r_{bd},r_{bd}]$ 为无人机滚转角的范围。

2.2.3 回报函数

在强化学习中，设计合理的回报函数至关重要。参考文献[28]设计的成本函数，定义回报函数为

$$\begin{cases} r = -\text{Cost} \\ \text{Cost} = \max\left\{d, \dfrac{d_1 \mid s_3 \mid}{\pi(1+\omega d)}\right\} \\ d = \max\{d_1 - \rho, 0, \rho - d_2\} \\ \rho = \sqrt{s_1^2 + s_1^2} \end{cases} \tag{13}$$

式中：r 为立即回报值；d_1 和 d_2 分别为圆环的内半径和外半径(以长机为中心，见图 1)；d 为僚机到圆环的距离；ω 为调整因子，用以调整 d 的权重；ρ 为长机与僚机之间的距离。

图 3 为长机与僚机相对位置关系对回报函数的影响。可知，当僚机位于以长机为中心的圆环内时，回报函数值最高；在圆环外部，当僚机靠近或远离长机时，回报函数值降低。这与图 1 所描述的场景想定是一致的。

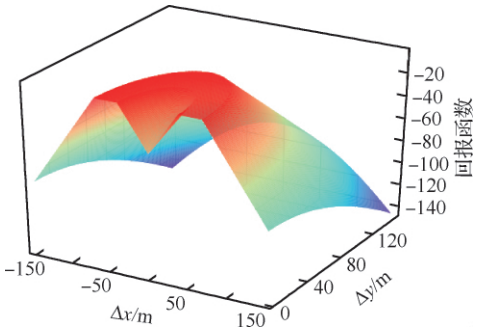


图 3 长僚机相对位置与回报函数的关系

Fig. 3 Relation between position of follower relative to leader and reward function

3 ID3QN 协调控制算法

DQN 算法结合了深度学习和强化学习的优势，能够较好地处理高维连续状态空间下的 RL 问题。因此，该算法在机器人领域得到了广泛的应用[29-30]。结合双重 Q 学习和竞争网络，在 DQN 算法的基础之上进行，提出 ID3QN 算法，并应用该算法解决连续状态空间中无人机编队的协调控制问题。

3.1 动作选择策略

为提高训练阶段 D3QN 的学习效率，将 ε-greedy策略与模仿策略相结合，提出 ε-imitation 动作选择策略平衡探索与利用。所谓模仿策略，是指僚机模仿长机行为(滚转指令)、参照长机的状态信息选择自身的滚转指令。ε-imitation 动作选择策略的主要实现步骤见算法 1。

算法1 ε-imitation 动作选择策略

输入： 待选动作Q值Q；长机航向角ψ_{l}；僚机滚转角φ_{f}；僚机航向角ψ_{f}；阈值ψ_{δ}

输出： 动作a

1: 生成随机数 $p, p \in (0,1)$
2: **if** $p > \varepsilon$ (ε为探索率)
3: $a \leftarrow \arg\max_{a'} Q(Q(s, a'))$
4: **else**
5: **if** $\psi_{\mathrm{f}} - \psi_{\mathrm{l}} > \psi_{\delta}$ **then**
6: $a \leftarrow +a^{\mathrm{r}}_{\max}$
7: **else if** $\psi_{\mathrm{f}} - \psi_{\mathrm{l}} < -\psi_{\delta}$ **then**
8: $a \leftarrow -a^{\mathrm{r}}_{\max}$
9: **else**
10: **if** $\varphi^{\mathrm{l}}_{\mathrm{d}} - \varphi_{\mathrm{f}} > a^{\mathrm{r}}_{\max}/2$ **then**
11: $a \leftarrow +a^{\mathrm{r}}_{\max}$
12: **else if** $\varphi^{\mathrm{l}}_{\mathrm{d}} - \varphi_{\mathrm{f}} < -a^{\mathrm{r}}_{\max}/2$ **then**
13: $a \leftarrow -a^{\mathrm{r}}_{\max}$
14: **else**
15: $a \leftarrow 0$
16: **end if**
17: **end if**
18: **end if**

该策略的主要思路为：当僚机以 $1-\varepsilon$ 的概率从动作集合选择动作时，僚机参考长机的状态信息和期望滚转角选择动作，而不是盲目地随机选择。具体而言：① 当两机的飞行方向相差较大时，僚机主要参考长机的航向角确定自身滚转动作；例如，当僚机的航向角远大于(超过阈值 ψ_{δ})长机的航向角时(Line 5)，滚转动作应选择 $+a^{\mathrm{r}}_{\max}$ (Line 6)，以缩小两机之间的航向差。② 当两机飞行方向大致相同时(Line 9)，僚机主要参考长机的滚转角确定自身滚转动作；例如，当长机期望的滚转角与僚机的滚转角相差不大(在 $\pm a^{\mathrm{r}}_{\max}/2$ 之间)时(Line 14)，僚机应保持当前滚转角，滚转动作应选择 0°(Line 15)。通过这种方式，僚机即可尽可能地与长机期望滚转角保持一致。

该策略降低了初始阶段僚机的盲目性，减少了无效探索的次数，增加了经验池中正样本的数量，有助于训练效率的提升。

3.2 D3QN 网络结构

为准确地估计 Q 函数，构建如图 4 所示的 D3QN 网络模型。该网络以系统联合状态为输入，输出为所有有效动作的 Q 值。上述 D3QN 由

两个子网络组成：多层感知机和竞争网络。多层感知机包含 3 层全连接层（Fully-Connected，FC），隐含节点数分别为 64、256 和 128，均使用 ReLU 激活函数[31]。竞争网络包含两个支路：状态值函数支路和优势函数支路。状态值函数支路和优势函数支路均包含两层全连接层，两支路第 1 层全连接层的隐含节点数均为 64，亦使用 ReLU 激活函数[31]。状态值函数支路第 2 层全连接层的网络节点数为 1，输出值为当前状态的值函数；而优势函数支路第 2 层全连接层的网络节点数为 3，输出值表示动作空间中 3 个待选动作的优势函数。D3QN（Dueling Double Deep Q-Network）输出层的输出为当前状态下各个待选动作的 Q 值，其值可通过“聚合”两支路的输出值得出。“聚合”操作的计算公式由式（8）定义。

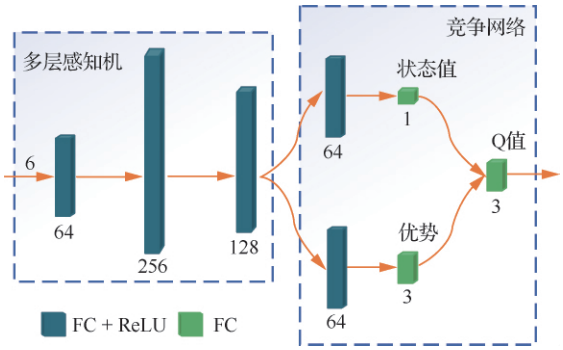


图 4 D3QN 网络结构

Fig. 4 Network structure for D3QN

3.3 算法实现

采用 ID3QN 算法实现固定翼无人机编队协调控制，训练过程如图 5 所示。僚机被映射为 RL 中的智能体，智能体在与环境的不断交互中学习控制策略，更新网络参数。僚机获取长机的状态信息及自身的状态信息，组成联合系统状态 s 输入到 D3QN 网络中，ε-imitation 动作选择策略根据 D3QN 的输出选取僚机的滚转动作 a；分别将长机（长机的滚转动作随机产生以增加系统的随机性）和僚机的滚转指令输入随机无人机运动学模型，得到长机和僚机下一时刻的状态；回报函数值 r 和下一时刻系统状态 s' 亦可随之得出。交互过程中所产生的元组数据（s，a，r，s'）均被保持到经验池中。在每一时间步，从经验池中进行随机采样，批次更新 D3QN 的网络参数。当每回合的时间步达到一定步数，结束该回合，重新开始下一回合的学习。基于 ID3QN 的协调控制算法的主要实现步骤见算法 2。

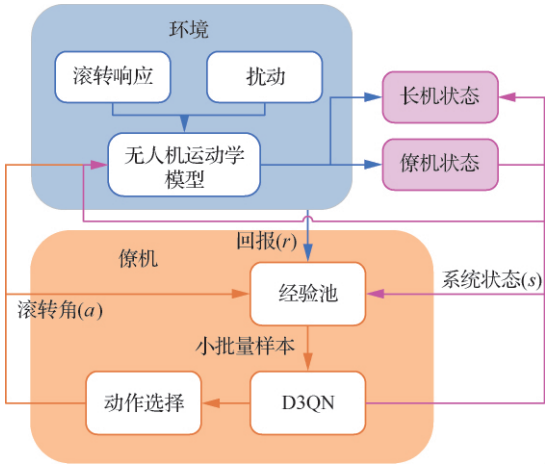


图 5 ID3QN 协调控制算法训练框图

Fig. 5 Block diagram of ID3QN coordination control algorithm

算法 2 ID3QN 算法

输入：单回合最大时间步 N_s；最大训练回合数 N_{max}

初始化经验池 D（最大容量为 N）；
随机初始化 D3QN 主网络参数 θ；
初始化目标网络参数 $\theta^- \leftarrow \theta$
repeat（for 每一回合）
  随机初始化系统状态 $s \leftarrow (\xi_l, \xi_f, \varphi_d^l)$；$t=1$
  while $t \leqslant N_s$ do
    根据 ε-imitation 动作选择策略（算法 1）
    选取僚机滚转动作 a
    由式（12）计算僚机滚转角设定值 φ_d^f
    将控制指令 φ_d^f 应用到无人机运动学模型（式（9））中，生成僚机下一时刻状态 ξ_f'
    观测下一时刻长机的状态 ξ_l' 和滚转角设定值 $\varphi_d'^l$
    由式（10）构建系统状态 $s' \leftarrow (\xi_l', \xi_f', \varphi_d'^l)$
    根据式（13）计算立即回报 r
    将状态转移数据元组（s，a，r，s'）保存到经验池 D 中
    若经验池溢出，即 $\|D\| > N$，则删除 D 中最早的经验数据
    从经验池 D 中随机抽取 N_b 个样本 $(s^j, a^j, r^j, s^{j+1})(j = 1, 2, \cdots, N_b)$
    计算每一元组数据的目标 Q 值：$y^j = r^j + \gamma Q(s^{j+1}, \arg\max_{a'} Q(s^{j+1}, a'; \theta); \theta^-)$
    根据损失函数更新主网络参数 θ：$L = \frac{1}{N_b}\sum_j \| y^j - Q(s^j, a^j; \theta) \|^2$
    朝向主网络更新目标网络参数 θ^-：$\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$（$\tau$ 为软更新率）
    $s \leftarrow s'$；$[\xi_l, \xi_f, \varphi_d^l] \leftarrow [\xi_l', \xi_f', \varphi_d'^l]$；
    $t \leftarrow t+1$
  end while
until 最大训练回合数

4 仿真验证及性能分析

4.1 参数设置

在 Python 环境中基于 TensorFlow 框架构建 D3QN 网络。D3QN 的网络参数均使用 Adam 优化算法进行更新，batch size(N_b)设为 32。共进行 50 000 回合的训练，每回合的仿真时间为 30 s，即最大训练回合数 $N_{max}=50\,000$，每回合的最大时间步 $N_s=30$。需要指出的是，在正式训练前进行 200 回合的预训练，用于收集经验数据以进行批次训练。在训练过程中，探索率 ε 在 10 000 回合内从初始值 1.0 线性衰减到最小值 0.1；D3QN 主网络参数的学习率 α 与目标网络的更新率 τ 从初始值(0.010，0.0010)指数衰减到最小值(0.001，0.0001)，衰减频率为 1 000 回合，衰减率为 0.9，即每隔 1 000 回合衰减为原来的0.9 倍。训练过程中所需参数的经验值详见表 1。

表 1 ID3QN 参数设置

Table 1 Parameter settings for ID3QN

参数	取值	参数	取值
d_1	40	d_2	65
ω	0.05	α_g	9.8
$\bar{v}$	10	σ_v	0.8
$\bar{\omega}_n$	6.3	σ_ω	0.1
$\bar{\zeta}$	0.5561	σ_ζ	0.01
$\bar{\eta}_x$，$\bar{\eta}_y$，$\bar{\eta}_\psi$	0	σ_x，σ_y，σ_ψ	1
a^r_{max}/(°)	15	r_{bd}/(°)	30
ψ_{δ}/(°)	20	N_s	30
α	$10^{-3}\rightarrow10^{-4}$	τ	$10^{-2}\rightarrow10^{-3}$
ε	1→0.1	γ	0.95
N_b	32	N	100 000

4.2 数值仿真实验

4.2.1 训练结果分析

为对策略进行有效的评价分析，使用单位回合内(如 N_e 回合)每一时间步的平均回报 G_{Ave} 作为度量标准来评价策略的优劣，其定义为

$$G_{Ave}=\frac{1}{N_e N_s}\sum_{n=1}^{N_e}\sum_{t=1}^{N_s} r_t^n \tag{14}$$

式中：r 为立即回报，由式(13)确定。

为验证提出的 ID3QN 协调控制算法的可行性和有效性，分别使用 DDQN、D3QN 和 ID3QN 算法进行对比实验。其中，D3QN 使用 ε-greedy 动作选择策略，其他流程与 ID3QN 完全相同；DDQN 与 D3QN 算法流程完全相同，二者唯一的区别在于网络结构的不同：D3QN 多层感知机分为两个支路分别估计状态值函数和优势函数，而后通过式(8)定义的“聚合”操作产生 Q 值，而 DDQN 仅构造单个支路的全连接层直接近似 Q 函数。为保证对比实验的公平性，上述 3 种算法均使用相同的深度网络结构(见图 4，DDQN 没有进行拆分操作，仅有 1 个支路)和参数设置(见表 1)。在整个训练过程中，每隔 100 回合(即 $N_e=100$)记录一次平均回报 G_{Ave} 的值，上述 3 种算法的学习曲线如图 6 所示。

由图 6 可知，在训练初期，3 种算法的回报曲线均快速上升；在大约 10 000 回合的训练后，3 种算法获取的平均回报逐渐趋于稳定。DDQN 与 D3QN 的回报曲线几乎重合，这意味着两种算法具有大体相当的性能；而在训练初期，D3QN 的回报曲线增长速度略高于 DDQN，这表明竞争网络可以更有效地学习 Q 函数。与以上两种算法相比，ID3QN 算法无论是在初始阶段还是在收敛阶段都能够获取最高的平均回报；这意味着在 ε-imitation 动作选择策略的引导下，ID3QN 算法能够更快更有效地学习最佳策略。

4.2.2 测试结果分析

完成 4.2.1 节的训练过程后，对训练后的协调控制策略进行测试分析。测试实验中，两架僚机与一架长机组成编队。每隔 1 s，长机随机选择滚转动作，而僚机根据训练后 D3QN 网络的输出选择最大 Q 值所对应的滚转动作。实验中，最大时间步(N_s)设置为 120，即仿真时间为 2 min。编队的飞行轨迹见图 7，飞行过程中立即回报值 r、僚机与长机之间的距离 ρ 和航向差 $\Delta\psi$ 的变化曲线情况见图 8。

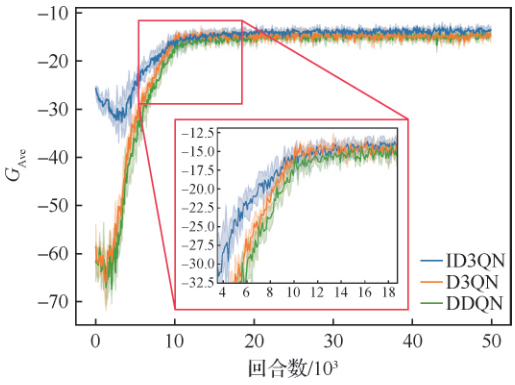


图 6 3 种算法的学习曲线

Fig. 6 Learning curves of three algorithms

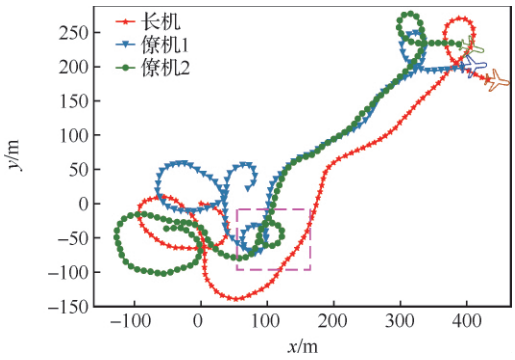


图 7 数值仿真中 ID3QN 策略的测试结果

Fig. 7 Testing results of ID3QN policy in numerical simulation

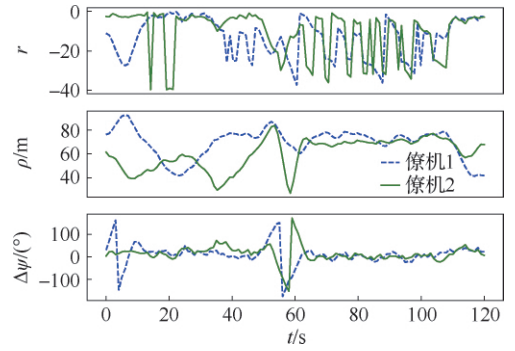


图 8 数值仿真中 ID3QN 策略的性能曲线

Fig. 8 Performance curves of ID3QN policy in numerical simulation

图 7 直观地展示了 ID3QN 协调控制策略的效果。无论是在前期和后期的转弯阶段，还是在中期的平直飞行阶段，两架僚机均能较好地跟随长机飞行。值得注意的是，在 55 s 左右，两架僚机均位于长机前方且距长机较远。在之后的十多秒内，两架僚机通过大滚转角机动实现了绕圈飞行（见图 7 紫色方框内）。这是因为滚转角是僚机唯一的控制量，两架僚机只能通过盘旋来缩小与长机之间的距离。在之后的飞行中，僚机可以维持与长机之间的距离在 70 m 上下，航向差大致在±25°的范围之内。

除以上的定性评价外，继续进行定量测试以进一步分析所得协调控制器的有效性。在定量测试实验中，4 架僚机分别使用 3.1 节提出的模仿策略和 4.2.1 节训练得到的 DDQN、D3QN 和 ID3QN 3 种控制策略跟随长机协同飞行。实验共进行 100 回合，每回合的仿真时间（N_s）设为 120 s。在每回合的实验中，长机的初始状态和滚转指令随机产生。为保证测试实验的公平性，4 架僚机的初始状态随机产生并保持一致。4 种策略的测试结果见表 2。

由表 2 可知，3 种基于 DQN 的深度强化学习算法（即 DDQN、D3QN 和 ID3QN）获得了远高于模仿策略的平均回报；同时，3 种算法所得平均回报的方差远低于模仿策略。这意味着基于 DQN 的无人机编队协调控制策略具有良好的可行性和稳定性。大体来看，3 种策略所获取的平均回报相差不大，ID3QN 所得平均回报略高。与 D3QN 和 DDQN 相比，ID3QN 策略的方差最低，这意味 ID3QN 具有更好的鲁棒性。上述结果表明，提出的 ID3QN 算法的性能优于 D3QN 和 DDQN 算法。

表 2 测试阶段 4 种策略性能对比

Table 2 Comparison results of four policies in testing stage

算法	G_{Ave}	方差
ID3QN	−11.04	22.35
D3QN	−11.61	26.71
DDQN	−11.86	32.27
Imitation	−39.39	596.42

4.3 硬件在环实验

为展示所提 ID3QN 协调控制算法的泛化能力和应用价值，基于 X-Plane 10 飞行仿真器建立高保真半实物仿真系统进行硬件在环实验，验证

所得策略的实用性。

4.3.1 半实物仿真系统

如图 9 所示，搭建的高保真半实物仿真系统由地面控制站、飞行仿真器、自动驾驶仪和机载处理器组成：

1）使用课题组开发的多机控制站 SuperStation 作为地面控制站，完成对多架无人机的控制，如模式切换、航线规划等。

2）使用商业飞行模拟软件 X-Plane 10 作为飞行仿真器，X-Plane 10 能够模拟风速变化、天气变化等环境扰动。

3）选择 PIXHAWK 作为自动驾驶仪的硬件平台。

4）使用英伟达 Jetson TX2 作为机载处理器。

长机和僚机共享一个地面控制站，即地面站可以同时监控长机和僚机。二者的机载处理器通过 RJ45 网线连接，模拟机间无线通信链路。

协调控制软件架构如图 10 所示，选用 PX4 开源飞控作为 PIXHAWK 自动驾驶仪的软件栈。ID3QN 协调控制策略运行在 TX2 上，TX2 上安装有 Ubuntu 14.04 操作系统和机器人操作系统（Robot Operating System，ROS）。TX2 与 PIXHAWK/PX4 通过 MavLink 协议连接。使用以下节点实现无人机编队的协调控制：

1）Communicator 节点：通过 UDP 协议接收长机状态信息。

2）Flocking Commander 节点：基于 ID3QN 算法完成上层协调控制。

3）Controller 节点：通过 PID 控制器完成底层闭环控制。

4）MAVROS 节点：通过 MavLink 协议同 PX4 建立连接获取自身状态信息。

4.3.2 实验结果分析

在半实物仿真实验中，一架僚机直接使用数值仿真环境中训练得到的 ID3QN 协调控制策略完成跟随长机飞行的任务。长机采用随机策略生成滚转指令，僚机根据训练后的 ID3QN 策略每隔 1 s 更新一次滚转指令，完成协调控制。二者的控制策略分别独立运行在各自的机载处理器上，二者的机载处理器通过网线连接，长机通过 UDP 协议将自身状态信息发送给僚机。半实物仿真实验流程如下：

1）在 MANUAL 模式下使用地面站控制长机与僚机起飞。

2）使用地面站控制飞机切入 MISSION 模式，两机按照预设航线飞行，并保持一定距离。

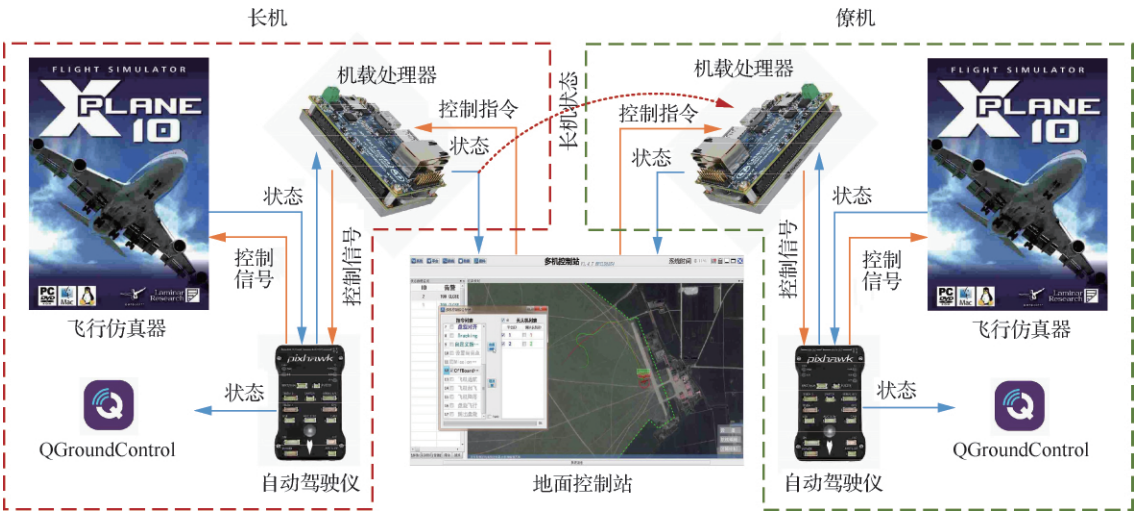


图 9 高保真半实物仿真系统

Fig. 9 High-fidelity semi-physical simulation system

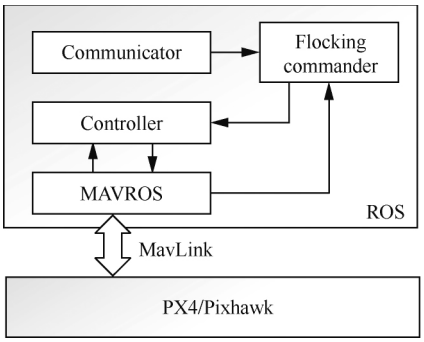


图 10 协调控制软件架构

Fig. 10 Software architecture for coordination control

3）使用地面站控制僚机切入 OFFBOARD 模式，僚机根据 ID3QN 策略完成跟随飞行任务。在每一时间步，Flocking Commander 节点根据 ID3QN 策略更新滚转指令，决策过程如下：① 从 Communicator 节点获取长机状态（即位置、姿态和速度）信息，而后将其与从 MAVROS 节点获取的自身状态相结合，构建系统状态；② 载入数值仿真环境中训练得到的 D3QN 网络模型参数；③ 以系统状态为输入，D3QN 网络输出滚转指令，进而生成滚转角设定值；④ 向 Controller 节点发布滚转角设定值，该节点据此通过 PID 控制器完成底层闭环控制。

4）一段时间后，使用地面站控制飞机切入 RETURN 模式，实验结束。

硬件在环仿真飞行实验共持续 120 s，长机的滚转角设定值在 $-10°\sim10°$ 之间随机产生，飞行速度设置为 10 m/s。实验中长僚机的飞行轨迹、航向角和滚转角的变化情况见图 11，飞行过程中的立即回报值 r、长僚机之间的距离 ρ 和航向差 $\Delta\psi$ 见图 12。在初始时刻，僚机与长机之间的距离高达 110 m，且僚机位于长机的前方。在随后 20 多秒的时间内，僚机通过盘旋飞行成功将两机之间的距离缩短到 75 m 之内。这是因为滚转角是僚机唯一的控制量，僚机只能通过盘旋缩小其与长机之间的距离。在之后的飞行中，无论长机平直飞行还是机动转弯，僚机均能及时做出反应，稳定地跟随长机飞行。需要指出的是，训练得到的控制策略在用于半实物仿真环境下的仿真飞行实验时并没有进行任何的参数调整。上述结果充分表明，所提ID3QN算法训练得到的协调控制

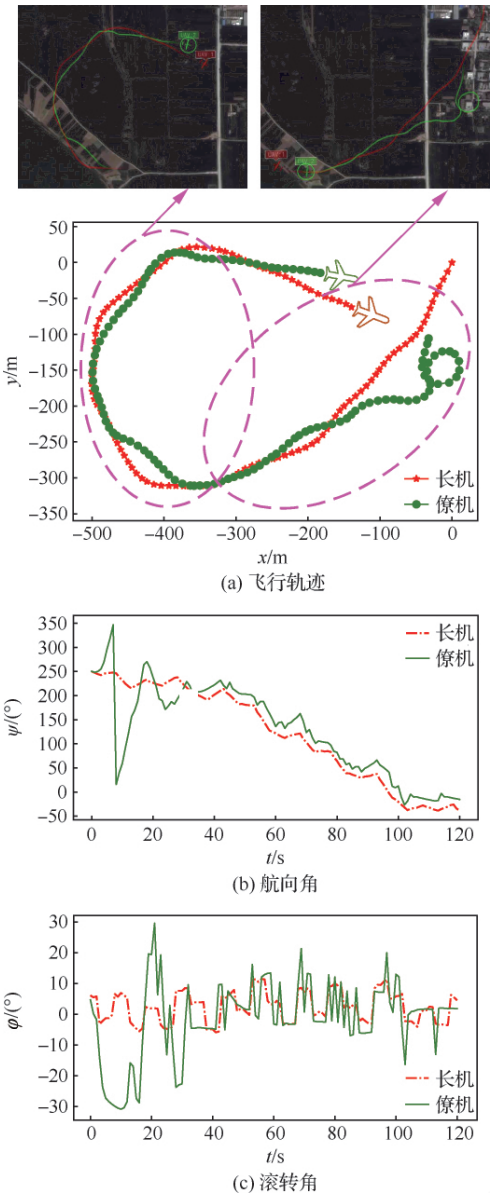


图 11 硬件在环实验结果

Fig. 11 Results of hardware-in-loop simulation

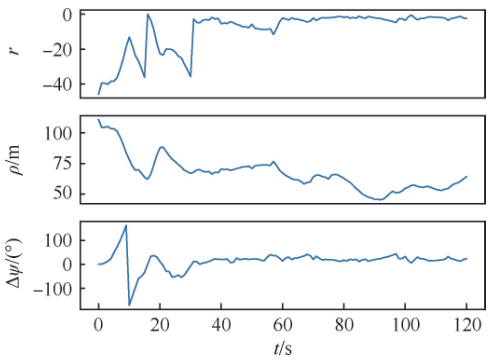


图 12 硬件在环实验中 ID3QN 策略的性能曲线

Fig. 12 Performance curves of ID3QN policy in hardware-in-loop simulation

策略可直接迁移到半实物仿真环境中，具有较强的适应性及良好的实用性。

5 结 论

聚焦动态不确定环境下的固定翼无人机编队协调控制问题，基于深度强化学习提出了无人机编队协调控制方法。首先在强化学习背景下对无人机协调控制问题进行了形式化描述，建立了协调控制 MDP 模型。进而将 ε-greedy 策略与模仿策略相结合，提出了 ε-imitation 动作选择策略，并将其引入 DQN 算法，提出了 ID3QN 算法以提高算法的学习效率。数组仿真环境下的训练结果和测试结果明：在 ε-imitation 动作选择策略的引导下，ID3QN 算法能够更快更有效地学习最佳策略。最后，构建高保真半实物仿真系统验证了算法的有效性和可迁移性。硬件在环飞行仿真实验显示，数值仿真环境下训练得到的控制策略无需任何参数调整即可直接迁移到半实物仿真系统中。这一结果表明，提出的 ID3QN 协调控制算法具有较强的适应性及良好的实用性。

参 考 文 献

[1] 宗群，王丹丹，邵士凯，等. 多无人机协同编队飞行控制研究现状及发展[J]. 哈尔滨工业大学学报，2017，49(3)：1-14.
ZONG Q，WANG D D，SHAO S K，et al. Research status and development of multi UAV coordinated formation flight control[J]. Journal of Harbin Institute of Technology，2017，49(3)：1-14 (in Chinese).

[2] 樊琼剑，杨忠，方挺，等. 多无人机协同编队飞行控制的研究现状[J]. 航空学报，2009，30(4)：683-691.
FAN Q J，YANG Z，FANG T，et al. Research status of coordinated formation flight control for multi-UAVs[J]. Acta Aeronautica et Astronautica Sinica，2009，30(4)：683-691 (in Chinese).

[3] 王祥科，刘志宏，丛一睿，等. 小型固定翼无人机集群综述和未来发展[J]. 航空学报，2020，41(4)：323732.
WANG X K，LIU Z H，CONG Y R，et al. Miniature fixed-wing UAV swarms：Survey and directions[J]. Acta Aeronautica et Astronautica Sinica，2020，41(4)：323732 (in Chinese).

[4] 贾永楠，田似营，李擎. 无人机集群研究进展综述[J]. 航空学报，2020，41(S1)：723738.
JIA Y N，TIAN S Y，LI Q. The development of unmanned aerial vehicle swarms[J]. Acta Aeronautica et Astronautica Sinica，2020，41(S1)：723738 (in Chinese).

[5] KURIKI Y，NAMERIKAWA T. Formation control with collision avoidance for a multi-UAV system using decentralized MPC and consensus-based control[J]. Journal of Control，Measurement，and System Integration，2015，8(4)：285-294.

[6] SAIF O，FANTONI I，ZAVALA-RIO A. Distributed integral control of multiple UAVs：Precise flocking and navigation[J]. IET Control Theory & Applications，2019，13(13)：2008-2017.

[7] PHAM H X，LA H M，FEIL-SEIFER D，et al. Cooperative and distributed reinforcement learning of drones for field coverage[JDB/OL]. arXiv preprint：1803.07250，2018.

[8] HUNG S M，GIVIGI S N. A Q-learning approach to flocking with UAVs in a stochastic environment[J]. IEEE Transactions on Cybernetics，2017，47(1)：186-197.

[9] SUTTON R S，BARTO A G. Reinforcement learning：An introduction[M]. Cambridge：MIT Press，1998.

[10] 高阳，陈世福，陆鑫. 强化学习研究综述[J]. 自动化学报，2004，30(1)：86-100.
GAO Y，CHEN S F，LU X. Research on reinforcement learning：A review[J]. Acta Automatic Sinica，2004，30(1)：86-100 (in Chinese).

[11] YAN C，XIANG X. A path planning algorithm for UAV based on improved Q-learning[C]// International Conference on Robotics and Automation Sciences (ICRAS). Piscataway：IEEE Press，2018：1-5.

[12] EVERETT M，CHEN Y F，HOW J P. Motion planning among dynamic，decision-making agents with deep reinforcement learning[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway：IEEE Press，2018：3052-3059.

[13] TAI L，PAOLO G，LIU M. Virtual-to-real deep reinforcement learning：Continuous control of mobile robots for mapless navigation[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway：IEEE Press，2017：31-36.

[14] LONG P，LIU W，PAN J. Deep-learned collision avoidance policy for distributed multiagent navigation[J]. IEEE Robotics and Automation Letters，2017，2(2)：656-663.

[15] TOMIMASU M，MORIHIRO K，NISHIMURA H，et al. A reinforcement learning scheme of adaptive flocking behavior[C]// International Symposium on Artificial Life and Robotics (AROB). Oita：ISAROB，2005：GS1-4.

[16] MORIHIRO K，ISOKAWA T，NISHIMURA H，et al. Characteristics of flocking behavior model by reinforcement learning scheme[C]// SICE-ICASE International Joint Conference. Piscataway：IEEE Press，2006：4551-4556.

[17] LA H M，SHENG W. Distributed sensor fusion for scalar

field mapping using mobile sensor networks[J]. IEEE Transactions on Cybernetics, 2013, 43(2): 766-778.

[18] LA H M, LIM R, SHENG W. Multirobot cooperative learning for predator avoidance[J]. IEEE Transactions on Control Systems Technology, 2015, 23(1): 52-63.

[19] WANG C, WANG J, ZHANG X, et al. A deep reinforcement learning approach to flocking and navigation of UAVs in large-scale complex environments[C]// IEEE Global Conference on Signal and Information Processing (GlobalSIP). Piscataway: IEEE Press, 2018: 1228-1232.

[20] HUNG S M, GIVIGI S N, NOURELDIN A. A dyna-$Q(\lambda)$ approach to flocking with fixed-wing UAVs in a stochastic environment[C]// IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE Press, 2015: 1918-1923.

[21] **左家亮, 杨任农, 张滢, 等. 基于启发式强化学习的空战机动智能决策[J]. 航空学报**, 2017, 38(10): 321168.
ZUO J L, YANG R N, ZHANG Y, et al. Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2017, 38(10): 321168 (in Chinese).

[22] WATKINS C, DAYAN P. Q-learning [J]. Machine Learning, 1992, 8(3): 279-292.

[23] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[24] VAN HASSELT H. Double Q-learning[C]// Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2010: 2613-2621.

[25] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]// AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2016: 2094-2100.

[26] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]// International Conference on Machine Learning (ICML). Brookline: JMLR, 2016: 1995-2003.

[27] WANG C, YAN C, XIANG X, et al. A continuous actor-critic reinforcement learning approach to flocking with fixed-wing UAVs[C]// Asian Conference on Machine Learning (ACML). Brookline: JMLR, 2019: 64-79.

[28] QUINTERO S A P, COLLINS G E, HESPANHA J P. Flocking with fixed-wing UAVs for distributed sensing: A stochastic optimal control approach[C]// American Control Conference. Piscataway: IEEE Press, 2013: 2025-2031.

[29] 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 & Robotic Systems, 2020, 98: 297-309.

[30] LV L, ZHANG S, DING D, et al. Path planning via an improved DQN-based learning policy[J]. IEEE Access, 2019, 7: 67319-67330.

[31] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// International Conference on Machine Learning (ICML). Brookline: JMLR, 2010: 807-814.

(责任编辑: 苏磊,王小辰)

Coordination control method for fixed-wing UAV formation through deep reinforcement learning

XIANG Xiaojia, YAN Chao*, WANG Chang, YIN Dong

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: Due to the complexity of kinematics and environmental dynamics, controlling a squad of fixed-wing Unmanned Aerial Vehicles (UAVs) remains a challenging problem. Considering the uncertainty of complex and dynamic environments, this paper solves the coordination control problem of UAV formation based on the model-free deep reinforcement learning algorithm. A new action selection strategy, ε-imitation strategy, is proposed by combining the ε-greedy strategy and the imitation strategy to balance the exploration and the exploitation. Based on this strategy, the double Q-learning technique, and the dueling architecture, the ID3QN (Imitative Dueling Double Deep Q-Network) algorithm is developed to boost learning efficiency. The results of the Hardware-In-Loop experiments conducted in a high-fidelity semi-physical simulation system demonstrate the adaptability and practicality of the proposed ID3QN coordinated control algorithm.

Keywords: fixed-wing UAVs; UAV formation; coordination control; deep reinforcement learning; neural networks

Received: 2020-03-24; Revised: 2020-05-18; Accepted: 2020-06-30; Published online: 2020-07-07 11:03
URL: http://hkxb.buaa.edu.cn/CN/html/20210429.html
Foundation items: National Natural Science Foundation of China (61906203); The Foundation of National Key Laboratory of Science and Technology on UAV, Northwestern Polytechnical University (614230110080817)
* Corresponding author. E-mail: yanchao17@nudt.edu.cn