doi:10.19306/j.cnki.2095-8110.2021.01.004

基于强化学习的多无人机避碰计算制导方法

赵　毓,郭继峰,郑红星,白成超

(哈尔滨工业大学航天学院,哈尔滨 150001)

摘　要:针对大量固定翼无人机在有限空域内的协同避碰问题,提出了一种基于多智能体深度强化学习的计算制导方法。首先,将避碰制导过程抽象为序列决策问题,通过马尔可夫博弈理论对其进行数学描述。然后提出了一种基于深度神经网络技术的自主避碰制导决策方法,该网络使用改进的 Actor-Critic 模型进行训练,设计了实现该方法的机器学习架构,并给出了相关神经网络结构和机间协调机制。最后建立了一个实体数量可变的飞行场景模拟器,在其中进行"集中训练"和"分布执行"。为了验证算法的性能,在高航路密度场景中进行了仿真实验。仿真结果表明,提出的在线计算制导方法能够有效地降低多无人机在飞行过程中的碰撞概率,且对高航路密度场景具有很好的适应性。

关键词:多智能体;强化学习;计算制导;固定翼;避碰

中图分类号:V249;TN967.5　**文献标志码**:A　**开放科学(资源服务)标识码(OSID)**:

文章编号:2095-8110(2021)01-0031-10

A Reinforcement Learning Based Computational Guidance Approach for UAVs Collision Avoidance

ZHAO Yu, GUO Ji-feng, ZHENG Hong-xing, BAI Cheng-chao

(School of Astronautics, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Aiming at the problem of cooperative collision avoidance for a large number of fixed wing UAVs in limited airspace, a computational guidance method based on multi-agent deep reinforcement learning is proposed. Firstly, the process of collision avoidance and guidance is formulated as a sequential decision problem, which is mathematically described by Markov game theory. Then, a decision-making method of autonomous collision avoidance guidance based on multilayer neural network technology is proposed. The network is trained by the improved Actor-Critic model. Furthermore, the machine learning architecture is designed to implement the method. The relevant neural network structure and coordination mechanisms among UAVs are given. Finally, a flight simulator with variable number of entities is established, in which centralized training and distributed execution are performed. In order to verify the performance of the algorithm, several simulation experiments are carried out in the scene of high traffic density. The simulation results show that the proposed onboard computational guidance method can effectively reduce the collision probability of multiple UAVs in flight process have a good adaptability to the scene of high route density.

Key words: Multi-agent; Reinforcement learning; Computational guidance; Fixed wing; Collision avoidance

收稿日期:2020-09-14;**修订日期**:2020-10-08

基金项目:国家自然科学基金(61973101);航空科学基金(20180577005)

作者简介:赵毓(1992-),男,博士研究生,主要研究方向为多智能体深度强化学习技术。E-mail: hitzhaoyu@hit.edu.cn

通信作者:郭继峰(1977-),男,博士,教授,主要从事多智能体强化学习方面的研究。E-mail: guojifeng@hit.edu.cn

0 引言

伴随着无人机行业的高速发展，多无人机在有限空域内协同执行任务成为可能[1]。无论在协同侦查搜索等作战任务中，还是在快递配送或飞行表演等日常场景中，多无人机间协同飞行避碰问题一直是相关制导技术研究的重点方向[2]。固定翼无人机因无法悬停及速度控制范围有限等技术特点，在航路密度较高的环境中，如果发生碰撞、损毁等安全事故，容易导致财产损失甚至人员受伤。因此，局部空域内大量固定翼无人机飞行碰撞冲突已成为相关领域亟待解决的突出问题。

无人机在自主执行任务期间主要依靠自身制导系统进行轨迹规划与目标跟踪。随着任务动态性的提高和执行任务期间无人机数量需求的增加，传统制导方法的自主性已难以满足相关性能要求。Lu 等最早提出了计算制导控制的概念，以描述具备更高自主性的新兴制导算法[3]。早期的计算制导方法研究主要集中在航天领域，模型预测控制可以被称作计算制导的前身[4]。Jiang 等针对行星动力下降过程设计了一种计算制导方法，并使用协同优化算法对其进行求解[5]。近期 Yang 等将计算制导方法应用在空中交通管制问题上，使用蒙特卡罗树搜索方法解决了临近空域内的飞行冲突问题[6]。本文研究的任务场景与其类似，但研究对象为具有更高动态特性的固定翼无人机，对制导系统的可靠性有着更为严格的要求。

早期国内外学者对多无人机避碰的航路规划或制导方法多是基于地面控制站实现。美国 McLain 等[7]最早基于最优控制的思想解决多无人机协调问题。国内的周炜等[8]基于层次分解法对多无人机避障问题展开研究，并实现了次优航路规划。杨秀霞等[9]分别基于时间约束和比例导引两方面对避障问题给出解决方案，通过数值解算得到了避碰时间估计方程和比例导引系数范围。

在机器人编队控制方向关于多智能体避碰问题的研究成果较多，近期无人机领域的相关学者也针对此类问题开展了研究。温家鑫等[10]使用改进的人工势场法进行无人机三维路径规划，有效地解决了传统虚拟力方法易陷于局部最优的问题。Horn 等[11]提出了一种基于虚拟结构的避碰算法，在势场法的基础上引入虚拟点理论，实现了对多智能体运动的控制。李相民等[12]将航迹规划抽象为滚动在线优化问题，基于模型预测控制法实现了 4 架无人机的避碰飞行。Everett 等[13]使用强化学习方法对不确定环境下无人车轨迹规划问题进行了研究，其训练成型的神经网络具有很好的避碰性能。本文也使用了强化学习方法训练制导决策神经网络，与前人不同之处在于使用了训练效率更高的 Actor-Critic 模型，而且设计的神经网络结构更简单，可以适应多无人机场景的高动态性要求。

本文通过构建自主决策神经网络的方法，解决了同一飞行高度共空域多固定翼无人机的实时飞行避碰制导问题。基于马尔可夫博弈(Markov game)理论，对多智能体序列制导决策问题进行数学建模，并确定优化目标。为了实现决策网络功能，根据 Actor-Critic 模型建立了集中训练和分布执行的多智能体强化学习训练系统。针对该自学习系统，分别设计了用于计算制导的执行网络(Actor)结构和用于评价联合动作的值函数网络(Critic)结构，并给出了基于 logit 协调机制的相关训练流程。在建立的多航路飞行场景模拟器中应用本文算法进行仿真，结果表明，算法在高航路密度场景中可以实现多无人机协同避碰飞行。

本文算法相对于传统基于数学模型或数值解析理论方法的优势，在于该方法对无人机数量可变环境具有更好的适应性，避免了多无人机控制领域自适应动态规划算法无法精确建模的问题。该研究内容为未来实物系统研究提供了方案参考和理论依据，具有一定的工程应用价值。

1 有限空域多无人机避碰建模

1.1 避碰问题描述

本文提出了一种在多无人机环境中分布式执行的自主制导决策神经网络算法，该方法适用于飞机总数可变的场景。经过训练后的制导神经网络以环境状态为输入，可以在线为无人机提供制导指令，这些指令被用于引导无人机飞向各自目的地的同时能够避免发生机间碰撞。在本研究中，仅考虑了飞机做水平运动的情况，即限定空域内所有飞机均在同一高度层飞行。这样假设主要有两个原因：1)为了增加无控条件下无人机的碰撞概率，进而有效验证算法性能；2)不失一般性地模拟了真实飞行情况，由于载荷和动力系统的限制，固定翼无人机协同执行任务时在同高度飞行情景较为常见[14]。为了降低复杂度，研究中假设所有无人机均通过无

延迟的可靠通信进行状态信息交换。

为更贴近现实并提高算法的适用性，本文基于中心地理论将训练场景设定为正六边形空域[15]，如图 1 所示。后续研究可以根据密铺原理对飞行空域进行扩展。在六边形每个顶点处，以随机时间间隔产生初始航向指向非相邻点的无人机。各无人机的初始速度在一定范围内随机选取，在飞行过程中速度存在扰动误差。

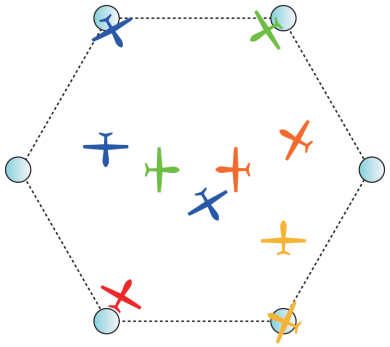


图 1　训练场景示意图

Fig. 1　Sketch map of training case

在实际应用中的固定翼无人机有多种构型[16]。由于本文研究算法为通用方法，不妨假设场景中无人机是碰撞半径为 50m 的空间球形包络，于是当空域内任意两机间距离小于 100m 时，即认定发生碰撞；同样，当飞机与空域边界距离小于 50m 时，认定为出界。无论发生碰撞还是出界，相关无人机都将从仿真场景中移除。

基于以上设定，本文提出算法的目标有 2 个：1）引导所有飞机到达各自目的地；2）避免飞行过程中发生碰撞或出界。

1.2　运动学模型和各项约束

许多学者已经对固定翼无人机的动力学问题展开了深入研究[17-18]。为了降低问题的复杂程度，本文仅考虑无人机在水平面内的运动控制问题，于是得到如下无人机二维运动学简化模型

$$\begin{aligned}\dot{x} &= v\cos\varphi\\ \dot{y} &= v\sin\varphi\\ \dot{\varphi} &= a_c\end{aligned} \tag{1}$$

其中，(x,y) 为无人机实时位置坐标；v 为无人机巡航速度；φ 为飞行航向角；a_c 为本文控制量，对应于航向角转动角速度。

在本文中，为了验证高速场景下的算法性能，将无人机的初速设定为 60m/s。巡航速度在每一仿真步长中会附加均值为 5m/s 的随机噪声，但实时飞行速度被限制在一定范围内，其值不大于 80m/s，不小于 40m/s。由于固定翼飞机相比旋翼飞机速度高很多，其控制系统必然存在扰动误差，速度噪声的引入是为了更加真实地模拟实际工程情况。

在生成每架无人机的初始阶段，随机选择非相邻节点为目标点，于是每架飞机从出生点指向目标点方向的航向角即为初始航向角。设定所有无人机在无控飞行状态下均保持初始航向角不变。在每一仿真步长中，文中制导算法为无人机选取某一确定的航向角转动角速度，该速度不大于 5(°)/s。本文制导决策神经网络会根据当前全局状态实时为无人机提供制导指令，为控制系统选择精确的转向角速度。无人机通过执行相应动作，实现对空域内其他飞机的规避并飞向目标点。

本文使用 off-policy 形式的强化学习方法对制导神经网络进行训练，训练过程每一回合(Episode)中产生 200 架无人机。对于每一个出生点生成的前后 2 架飞机，其时间间隔将会在 60～180s 内随机选取。关于更详细的参数信息将会在第 4 节中给出。

1.3　多智能体马尔可夫博弈模型

当空域内只有 1 架无人机与环境发生交互，其制导过程可以视为序列决策问题，使用传统的马尔可夫决策过程(Markov Decision Process，MDP)可以对其进行建模并求解。然而，本文研究空域内存在多个无人机对象，环境的整体状态受到所有飞机联合动作影响，对单架无人机来讲环境失去稳定性，MDP 方法不再适用于此场景。本文使用马尔可夫博弈理论对空域内可变数量飞行器的制导决策问题进行建模，该理论是 MDP 在多智能体条件下的一种自然扩展[19]。

多智能体的马尔可夫博弈问题可以用一个元组 $\{n,\boldsymbol{S},\boldsymbol{A}_1,\cdots,\boldsymbol{A}_n,\gamma,R,T\}$ 来表示，其中 n 代表环境中智能体的总数；$\boldsymbol{S}$ 代表整个系统可能状态的有限集合，也称为状态空间；$\boldsymbol{A}_i, i\in[1,n]$ 代表第 i 个智能体的可选动作集合，因本文智能体有相同可选动作集，此处可称为动作空间；γ 是奖励折扣系数；R 是联合奖励值，由环境受联合动作 $\boldsymbol{a}$ 影响后产生；T 是状态转移函数。某一时刻系统状态改变受所有智能体联合动作 $\boldsymbol{a}=(a_1,\cdots,a_n), a_i\in\boldsymbol{A}_i$ 影响，其中第 i 个智能体的动作 a_i 通过自身策略 π_i 选

择产生。

多智能体系统在联合策略 $\pi=(\pi_1,\cdots,\pi_n)$ 的指导下，有累计折扣奖励定义如下

$$J^{\pi}(s_t)=E^{\pi}\left\{\sum_{t=0}^{T_m}\gamma^t R(t+1)\right\},s_t\in S \quad (2)$$

其中，T_m 为总时间；t 为当前仿真时刻；s_t 为当前时刻环境的状态。多智能体马尔可夫博弈的终极目标是找到最优的联合策略 π^*，使得整个系统的累计期望回报值最大。

多智能体马尔可夫博弈的不稳定性表现在，某一智能体最优策略 π_i^* 会随着其余智能体策略的改变而变化。为了解决不稳定性问题，很多学者对纳什均衡和长期稳定行为进行了大量研究[20]。然而，当系统只有一个均衡点时，在有限时间内通过仿真或计算方式求得纳什均衡十分困难。本文设计的强化学习方法是在训练回合有限的情况下，寻找次优联合策略，后文为了方便表述也用 π^* 表示。

为了降低环境的不稳定性，本研究对场景内所有智能体使用 Logit 策略进行动作协调[21]。具体来说，在 Logit 策略中仅存在一个高级别智能体，其余智能体均保持原始动作不变。当环境中高级别智能体做出决策和选择动作后，向其余智能体发送这一信息，然后自动变为低级智能体。顺序选择下一智能体为高级别智能体，循环迭代进行决策，直到所有智能体完成动作选择。通过这种方式可以在智能体进行策略更新时固化环境影响，进而降低系统的不稳定性。

2 基于多智能体强化学习的计算制导方法

2.1 多智能体 Actor-Critic 算法原理

本文使用强化学习领域最流行的 Actor-Critic 方法解决多智能体避碰制导决策神经网络的训练问题[22]。在集中训练过程中，使用 2 个异步更新的 Actor 神经网络来拟合策略(Policy)，同样使用 2 个 Critic 网络逼近评价值(Q-Value)函数。环境中的所有无人机都共享一套神经网络结构，但各自网络参数存在一定噪声，通过这种方式鼓励协作的同时扩大系统的探索能力。在分布执行过程中，每架无人机各自使用一个独立的 Actor 网络生成制导指令。

本文提出的多智能体强化学习算法具有如下特点：1)所有智能体共享同一套 Actor-Critic 网络结构和参数，但个体网络参数存在噪声；2)理论上，由于使用了 Logit 策略，本文算法中不同智能体的动作选择是异步更新的；3)文中算法通过引入长短期记忆(Long Short Term Memory，LSTM)网络，能够处理智能体数目可变场景下的制导决策问题。下面将对多智能体计算制导方法进行详细分析。

本文所述强化学习过程中共有 4 个神经网络，分别是决策 Actor(ActorD)、估计 Actor(ActorE)、决策 Critic(CriticD)和估计 Critic(CriticE)。其中 ActorD 用来拟合制导策略，可以用参数 θ 描述，它是唯一在训练过程和执行过程都被使用的神经网络结构；CriticD 用来逼近评价函数，可以用参数 ω 描述。以第 i 个智能体为例，其在自学习系统中的训练过程如图 2 所示。

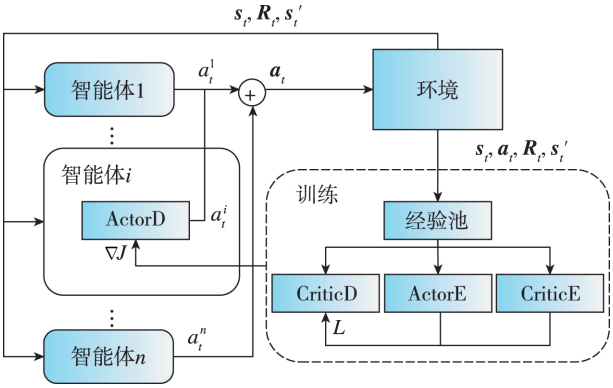


图 2 多智能体训练过程示意图

Fig. 2 Multi-agent training framework diagram

在训练期间，ActorD 实时生成决策，即通过获取当前环境中全局状态信息为智能体选择动作。在每一个时间步长 t 中，所有智能体都通过自身策略选择各自动作，虽然这些策略有相同的 ActorD 网络结构，但每次应用时都为参数附加了误差，并且每个个体的输入和输出是差异化的。环境中所有被选择的动作最终合成为联合动作 a_t。环境状态 s_t 在执行联合动作 a_t 后，通过状态转移函数更新为新的状态 s_t'，并返回即时奖励 R_t。每一仿真步长最终产生一个用集合 (s_t,a_t,R_t,s_t') 表示的案例，所有案例将会被存储在经验池 D 中。当经验池中案例达到一定数目时，随机从中抽取 M 组案例对所有神经网络进行训练。类似于深度 Q 学习(Deep Q-Learning)的思路，用 off-policy 这样的方式来提高动态环境中训练曲线的收敛速度。

案例训练过程中，CriticD 首先根据联合状态 s_t 和联合动作 a_t 生成评价值 Q_t；然后 ActorE 根据新

的联合状态 s_t' 生成估计联合动作 a_t'；CriticE 将会基于 s_t' 和 a_t' 生成 Q_t'。于是可以得到 CriticD 网络的损失方程为

$$L(\omega)=E[(Q_t(\boldsymbol{s}_t,\boldsymbol{a}_t)-\boldsymbol{R}_t-\gamma Q_t'(\boldsymbol{s}_t',\boldsymbol{a}_t'))^2] \tag{3}$$

基于求得的损失值 $L(\omega)$ 更新 CriticD 网络参数。

用式(4)可以求取 ActorD 网络的梯度

$$\nabla_\theta J(\pi)=E[\nabla_a Q(\boldsymbol{s}_t,\boldsymbol{a}_t)\cdot\nabla_\theta\pi(\boldsymbol{s}_t)] \tag{4}$$

其中，$J(\pi)=E[R]$ 由式(2)给出。使用 Adam 优化器根据上述梯度值更新 ActorD 的神经网络参数。

综上，本文强化学习训练的目的是通过不断调整各网络参数，使 ActorD 能够根据环境内全局状态为智能体选择最理想的动作。

2.2　多智能体强化学习算法设计

本文将多无人机的计算制导问题求解过程抽象为强化学习过程，将每架飞机看作一个智能体。制导算法的最终目标是形成一个决策神经网络，为智能体选择合适的动作，在避免碰撞的同时引导其到达目的地。由于空域中飞机随机产生，所以本文研究的是一个动态场景，正因如此，需要每个智能体能够具备对突发情况的临时决策能力。虽然给出了飞行速度约束，但在研究中只对无人机的飞行航向角进行控制，即选择航向角的改变角速率作为动作集合。前文已经提到可用航向角速度的最大值为 5 (°)/s，所以需要优化的动作集合是有界且连续的。以下将给出本文所用强化学习方法中状态空间、动作空间、回报函数和其他一些参数的相关定义。

(1)状态空间

每架无人机的状态信息包括实时位置 (x,y)、速度 v、飞行航向角 φ 和目标点 (g_x,g_y)。在 t 时刻，空域环境中所有智能体的状态信息联合构成了完整的全局状态信息 s_t，定义如下

$$\boldsymbol{s}_t=\{\boldsymbol{s}_t^1,\boldsymbol{s}_t^2,\cdots,\boldsymbol{s}_t^n\},\quad \boldsymbol{s}_t\in\boldsymbol{S} \tag{5}$$

其中，s_t 是一个 $n\times 6$ 的矩阵，n 是环境中智能体的总数。本文认为在仿真过程中所有智能体都能及时获得完整的全局状态信息。

(2)动作空间

每架无人机在仿真步长内可以选择一个固定的角速度值做航向角转向动作，于是智能体的动作空间为 $\boldsymbol{A}_i\in[-5,5]$ (°)/s，正值代表右转，负值代表左转。然而，通过大量仿真实验发现，在训练中使用连续的动作空间将会耗费大量计算时间，使得系统难以找到最优策略。为了提高训练速度，在本研究中将动作空间离散为 $\boldsymbol{A}_i=\{-5,-3,0,3,5\}$，其中 0 表示不发生转向。

(3)回报函数

在前文中已经给出了发生机间碰撞的距离为 100m，为了进一步提高算法的训练速度，在此还设定了机间警告距离为 500m，如果 2 架飞机间距离小于此值时，他们各自的回报值中会计入一个惩罚。基于以上设定，可以得到每个智能体的回报函数定义

$$r_i(\boldsymbol{s}_t)=\begin{cases}-2 & \text{碰撞}\\ -1 & \text{出界/警告}\\ -\dfrac{d_g}{d_{g\max}T_{\max}^i} & \text{其余情况}\\ 2 & \text{到达目标}\end{cases} \tag{6}$$

其中，d_g 是无人机与目标点之间的距离；$d_{g\max}$ 设定为限定空域内最大距离(对角线)；$T_{\max}^i$ 为智能体 i 的最大仿真步长，是仿真系统的超参数。于是环境的联合回报定义如下

$$R(\boldsymbol{s}_t)=\sum_{i=1}^{n}r_i(\boldsymbol{s}_t) \tag{7}$$

基于这样的回报函数设定，可以通过训练提高环境累积回报值以提升算法性能。

(4)其他相关参数

由于转移方程和回报函数已经确定，本文中强化学习训练的目标即是为多智能体问题找到优化的策略。为了解决这一随机博弈问题，需要找到可以使所有无人机未来累计折扣回报总和尽量大的策略 π^*。因为环境中所有智能体都有相同的回报函数，并在环境返回时累加所有智能体的回报值，而且有相同的决策神经网络结构，所以属于完全合作情景。如 2.1 节所述，将动作-值函数定义为 $Q(\boldsymbol{s}_t,a_1,\cdots,a_n)$，它是由当前全局状态和所有无人机联合共同决定的。

在仿真过程中，模拟器会生成固定数量的无人机。如果某一架无人机发生了碰撞、出界或者到达终点，它将会被从场景中移除。为了防止训练早期发生无人机无限循环飞行的情况，设定了每一回合中的最大仿真步数 $T_{\max}$，一旦有飞机总飞行时间达到 $T_{\max}$，则同样被从场景中移除。当所有飞机从场景中移除后，一个仿真回合结束。

3 自学习系统设计

3.1 网络结构设计

本文研究的避碰场景中，环境内每一时刻的无人机总数是变化的，因此智能体需要处理输入数据数目可变的情况。许多使用反馈式神经网络结构的强化学习方法中需要固定的输入维度，这限制了他们无法应用在智能体数量可变的场景中。为了解决这一问题，引入了 LSTM 网络结构，将长度变化的状态输入数据编码为固定长度的向量，进而提供给决策神经网络进行应用。本文设计的神经网络结构示意图如图 3 所示，是以第 i 个智能体的 ActorD 网络为例。其中 s_t^i 是智能体 i 在 t 时刻的状态信息；$(s_t^1,\cdots,s_t^n)$ 是除 i 以外环境中所有智能体的状态信息，他们共同构成了完整的全局状态。LSTM 网络的输出是一个经过编码的固定长度的隐藏状态 h_n，将被当作输入数据 (s_t^e) 传入后端的神经网络中。

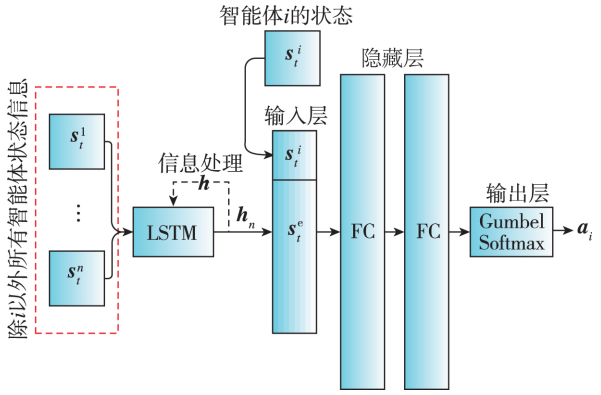


图 3 Actor 的神经网络结构示意图

Fig. 3 Illustration of the Actor neural network architecture

虽然 LSTM 网络结构经常被用于处理序列数据，但是只用到其对序列输入相关信息的存储能力，没有考虑时序相关性。在每一个仿真步长中，除自身外其余所有的智能体状态都被输入到 LSTM 网络中，以距离由远到近的顺序输入，这样可以保证距离最近的智能体在最终获得的隐藏状态中具有最大的影响力。每输入一个智能体的状态，LSTM 网络都会生成一个隐藏状态，并传入下一步计算中。将最后的隐藏状态作为一个固定长度的、编码过的全局状态提供给决策神经网路用于选择动作。

本文中 Actor 和 Critic 网络具有相似的前端结构，他们都有一个 LSTM 网络处理输入，有 2 个全连接隐层处理信息。但是，Critic 网络的输入中不仅包含了全局状态信息，还加入了所有智能体的当前动作信息，进一步增强了智能体之间的协同。如前文所述，Actor 神经网络的作用是拟合为智能体选择最优动作的策略函数，由于仿真中动作空间是离散的，所以使用 Gumbel-Softmax 估计器作为 Actor 的输出层[23]。Critic 网络的输出层只有一个神经元，用于输出评价值 Q。

3.2 算法流程

根据前文中相关描述和参数信息，表 1 给出了本文使用的强化学习方法算法流程。

表 1 强化学习计算制导算法流程

Tab. 1 The reinforcement learning based computational guidance algorithm process

算法：基于多智能体强化学习的计算制导方法	
1	分别用参数 $\theta,\omega,\theta'=\theta,\omega'=\omega$ 初始化 ActorD、CriticD、ActorE 和 CriticE 的神经网络.
2	初始化经验池 D 和计步器 $C_d=0$.
3	**for** episode = 1 to Max-Episode **do**
4	初始化联合动作 $a=(0,\cdots,0)\in\mathbf{R}^n$.
5	重置环境状态 S.
6	设定回合内最大仿真步数 $T_{\max}$.
7	**for** $t=1$ to $T_{\max}$ **do**
8	场景中随机生成固定翼无人机.
9	使用 ActorD 网络通过 Logit 策略生成各无人机动作 a_i. 将所有动作组成联合动作 $a_t=(a_t^1,\cdots,a_t^n)$.
10	执行联合动作 a_t，环境从 s_t 转移为 s_t' 并返回 R_t.
11	存储 (s_t,a_t,R_t,s_t') 到经验池，C_d ++.
12	用新状态 s_t' 替换原状态 s_t.
13	**if** (C_d > 最小样本数) and (t mod 采样周期 = 0) **do**
14	从经验池 D 中采样 M 组数据.
15	CriticD 计算 Q_t，ActorE 计算 a_t'，CriticE 计算 Q_t'.
16	通过最小化损失 $L(\omega)$ 更新 CriticD.
17	使用策略梯度更新 ActorD 的参数.
18	**if** (t mod 估计网络更新步长 = 0) **do**
19	用下式更新 ActorE 和 CriticE 的网络参数 $\theta'=\tau\theta+(1-\tau)\theta'$ $\omega'=\tau\omega+(1-\tau)\omega'$
20	**if** 无人机数量 > max_num **do**
21	结束循环
22	**end for** (t)
23	**end for**(episode)

在仿真期间，发生碰撞的次数将会被记录下来，并在每回合仿真最后输出。当模拟器生成了最

大数量为 max_num 的无人机后不再生成新的无人机。需要指出的是，为了提高训练速度，本文使用 AdaGrad 优化器对 LSTM 网络进行更新操作。

4　仿真分析

4.1　仿真条件设定

为了更好地训练和验证算法的有效性，基于 OpenAI 的 Gym 环境建立了一个大量飞机自由飞行的 2D 模拟器，模拟器中空域范围为 34km×34km。模拟器的输入为无人机的联合动作 a_t，输出为新的全局状态 s'_t 和当前即时回报 R_t。算法根据全局状态中参数对的个数判断当前环境中的无人机数目。每回合中最大仿真步数为 $T_{\max}=500\times$ max_num。

本文设置 LSTM 网络的隐层有 32 个节点，其输入数据需要正则化，选择 Softsign 函数作为其激活函数。各网络结构中的全连接层均为 128 个节点，使用 ReLU 函数作为激活函数。

操作系统环境为 Windows10 x64，使用软件工具包版本为 Python3.7 和 TensorFlow 2.1.0。硬件信息为 Intel i5-9600K、DDR4 16GB 和 240GB SSD。

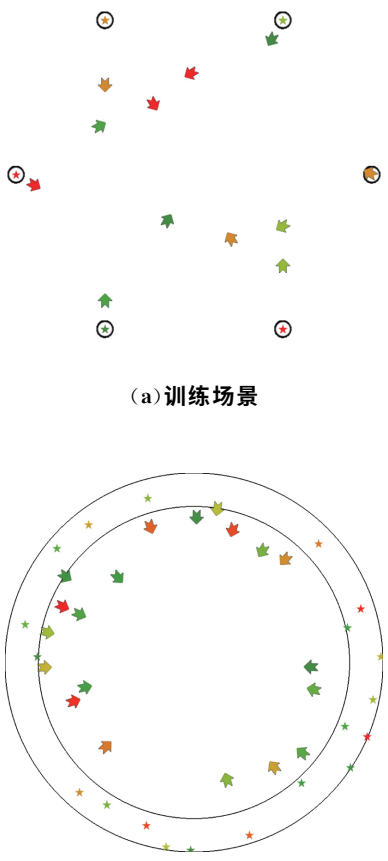
4.2　仿真结果及分析

为了研究多智能体计算制导神经网络的性能，分别设计了训练案例和压力测试案例。训练案例用来训练并验证文中算法的有效性，而压力测试案例则用于检验训练好的决策神经网络对高航路密度场景的适应性。下面分别给出了算法在各案例中的仿真结果。

(1)强化学习训练和验证

在此场景中，有 6 个出生点均匀分布在正六边形的各顶点处，相邻顶点间距为 16km。在每个出生点以 60～180s 的随机时间间隔生成新的无人机，并以非临近顶点作为其终点设定初始航路。图 4(a)是该场景在仿真过程中的截图。

将训练过程的最大回合数设定为 10 万次，在每个回合中生成 200 架无人机，单回合最大仿真步数 $T_{\max}=10^5$。强化学习系统有如下相关参数：网络更新率 $\tau=0.01$，折扣因子 $\gamma=0.98$，最小样本数为 1000 组，$M=10^3$，各优化器使用默认参数，神经网络的相关参数每 1000 回合存储 1 次。



(a)训练场景　　(b)压力测试场景

图 4　仿真过程场景截图

Fig. 4　Screenshots in simulation running scenarios

环境累计回报值随训练回合数变化的曲线如图 5 所示。从图 5 中可以看出，决策神经网络在经历了 2 万回合训练后才逐渐学习到有一定效果的策略，且整体收敛过程较慢。分析其原因，应是 LSTM 网络在训练前期难以收敛导致。由于后端决策网络输入了大量依赖 LSTM 网络输出的隐藏状态信息，LSTM 网络的性能严重限制了算法的整

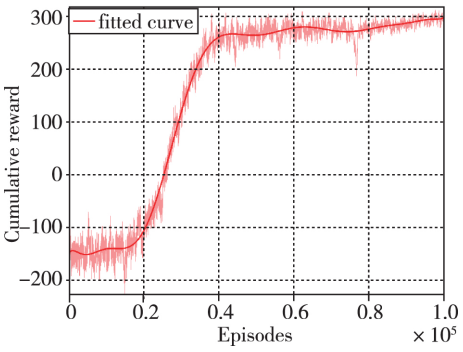


图 5　累积回报训练曲线

Fig. 5　Cumulative reward with the training episodes

体效果。曲线最终逐渐收敛于回报值 300 附近，说明本文算法能够有效避免航路冲突。

制导决策神经网络训练完成后，将训练好的 ActorD 应用在每架无人机上，进行 1000 回合独立仿真实验并统计发生碰撞次数信息。对飞机无制导自由飞行场景也进行了 1000 回合独立仿真实验，并将取得的结果作为对比。表 2 展示了本文算法的性能，说明提出的算法可以有效降低限定空域内多无人机飞行过程中的碰撞概率。由于初始航向即指向目标点，自由飞行场景中未发生碰撞的飞机全部到达目标点。在使用本文算法的场景中，所有飞机均到达目标点。统计了随机 100 回合中每个仿真步长内空域中无人机密度信息，并绘制图 6 所示直方图。综合表 2 和图 6 可以看出，大多数时间空域内有 20～26 架飞机在飞行，最大数量可达 30 架次，本文算法能够在此航路密度下保证不发生机间碰撞，说明其具有较好的性能。

表 2 测试案例碰撞统计表

Tab. 2 Collision statistics of test cases

仿真条件	平均碰撞概率/%	飞机碰撞中位数/次	目标点到达率/%
计算制导	0	0	100
自由飞行	62.77	132	37.23

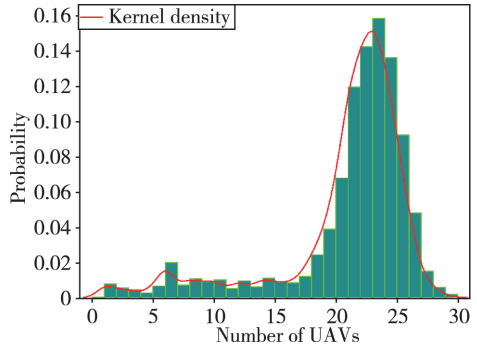


图 6 单步长内无人机数量统计

Fig. 6 Statistical histogram of UAVs in a single step

为了测试算法的运算速度，统计了随机 100 回合中某一步长内无人机数量与计算时间的关系数据，取均值后绘制成图 7 所示曲线。由图 7 可以看出，随着空域内无人机数量的增加，算法总的计算时间也快速增大。当空域内无人机数量不超过 22 架时，算法运算时间不超过 1s，在实际应用中可以接受。后续可以通过压缩神经网络或优化决策系统结构来进一步提升运算速度。

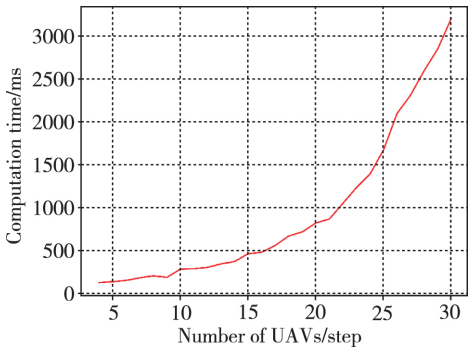


图 7 单步长内计算时间随无人机数量变化曲线

Fig. 7 Computation time with the number of UAVs in a single step

通过观察模拟器中无人机飞行路径发现，随着训练回合数的增加，无人机规避动作数量大量减少，飞向目标点的曲线愈发平滑。但这一结果并未得到充分统计和论证，未来将对此点进行深入研究。

(2)压力测试

为进一步评估算法性能，本文设计了一个限定空域内多无人机相遇的压力测试场景，此场景有助于研究算法在高航路密度条件下解决碰撞冲突的能力。在压力测试场景的每一回合中，产生从 10 架到 50 架不等数量的无人机，随机分布在一个内径为 28km、外径为 34km 的环形空域上。限定任意 2 架飞机的初始位置最小距离不小于 1km。每架飞机的终点被设定为空域中的中心对称位置，因此所有飞机的初始航向都必然指向空域中心，以确保他们都存在碰撞可能。图 4(b)展示了在场景中有 20 架飞机的仿真过程截图。

图 8 展示了随着飞机数量的增加，本文计算制导算法在测试场景中的性能，其中每个数据点表示 100 个独立回合中统计结果的均值。从图 8 中可以看出，如果空域内飞机不采取制导措施，60% 以上的飞机将在测试中发生碰撞，这一概率随着飞机数目的增长而逐渐增加。然而在应用本文算法的测试场景中，无人机的最大碰撞概率不超过 10%，在总数 28 架以下的场景中可以保证无碰撞发生。由此表明，即使在高航路密度情况下，本文提出的计算制导方法也具有良好的适应性。

5 结论

针对有限空域内多固定翼无人机避碰飞行问题，提出了一种实时分布式计算制导方法。算法分析与实验结果表明：

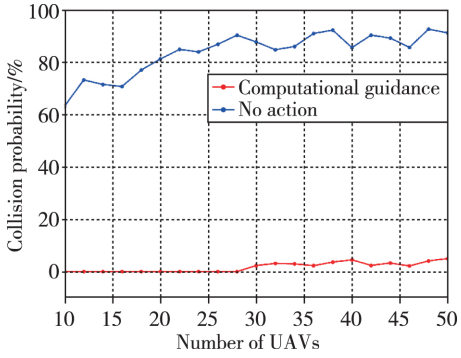


图 8　碰撞概率随飞机总数变化图

Fig. 8　Collision probability as the number of UAVs increases

1)提出了一种解决限定空域内多飞行器碰撞冲突问题的方法,为未来三维空间制导算法和实物仿真提供了新的研究思路;

2)基于全局状态信息的神经网络决策方法可以解决多智能体计算制导问题,能够有效降低多无人机在飞行过程中的碰撞概率;

3)基于强化学习的计算制导算法在高航路密度场景下具有良好的性能,能够有效避免碰撞发生,对提高限定空域内无人机容量有一定帮助。

文中算法在仿真过程中仅对空间进行抽象,并未考虑受硬件影响的精确时间因素;受限于硬件性能致使训练时间过长,未能对连续动作空间问题进行深入分析。未来需要对以上两点展开进一步研究。

参考文献

[1] 许晓伟,赖际舟,吕品,等. 多无人机协同导航技术研究现状及进展[J]. 导航定位与授时,2017,4(4):1-9.
Xu Xiaowei, Lai Jizhou, Lyu Pin, et al. A literature review on the research status and progress of cooperative navigation technology for multiple UAVs[J]. Navigation Positioning and Timing, 2017, 4(4): 1-9 (in Chinese).

[2] 韩亮,任章,董希旺,等. 多无人机协同控制方法及应用研究[J]. 导航定位与授时,2018,5(4):1-7.
Han Liang, Ren Zhang, Dong Xiwang, et al. Research on cooperative control method and application for multiple unmanned aerial vehicles[J]. Navigation Positioning and Timing, 2018, 5(4): 1-7(in Chinese).

[3] Lu P. Introducing computational guidance and control [J]. Journal of Guidance, Control, and Dynamics, 2017, 40(2): 193.

[4] Borelli F, Bemporad A, Morari M. Predictive control for linear and hybrid systems[M]. Cambridge University Press, 2017.

[5] Jiang X Q, Li S. Computational guidance for planetary powered descent using collaborative optimization [J]. Aerospace Science and Technology, 2018, 76 (5): 37-48.

[6] Yang X, Wei P. Scalable multiagent computational guidance with separation assurance for autonomous urban air mobility[J]. Journal of Guidance, Control, and Dynamics, 2020, 43(8): 1-14.

[7] McLain T, Chandler P, Pachter M. A decomposition strategy for optimal coordination of unmanned air vehicles[C]// Proceedings of 2000 American Control Conference. IEEE, 2000: 369-373.

[8] 周炜,魏瑞轩,董志兴. 基于层次分解策略无人机编队避障方法[J]. 系统工程与电子技术,2009,31(5):1152-1157.
Zhou Wei, Wei Ruixuan, Dong Zhixing. Formation method of obstacle avoidance for UAVs based on control architecture and decomposition strategy[J]. Systems Engineering and Electronics, 2009, 31(5): 1152-1157(in Chinese).

[9] 刘小伟,杨秀霞. 基于比例导引律的无人机避障研究[J]. 计算机仿真,2015,32(1):34-39+82.
Liu Xiaowei, Yang Xiuxia. Study on proportional navigation-based collision avoidance for UAV[J]. Computer Simulation, 2015, 32(1): 34-39+82(in Chinese).

[10] 温家鑫,赵国荣,赵超轮,等. 基于改进人工势场的无人机编队避障[J]. 飞行力学,2020,38(2):55-60.
Wen Jiaxin, Zhao Guorong, Zhao Chaolun, et al. Obstacle avoidance of UAV formation based on improved artificial potential field[J]. Flight Dynamics, 2020, 38(2): 55-60(in Chinese).

[11] Dang A D, Horn J. Collinear formation control of autonomous robots to move towards a target using artificial force fields[C]// Proceedings of IEEE International Conference on Technologies for Practical Robot Applications. IEEE, 2015: 1-6.

[12] 李相民,薄宁,代进进. 基于模型预测控制的多无人机避碰航迹规划研究[J]. 西北工业大学学报,2017,35(3):513-522.
Li Xiangmin, Bo Ning, Dai Jinjin. Study on collision avoidance path planning for multi-UAVs based on model predictive control[J]. Journal of Northwestern Polytechnical University, 2017, 35(3): 513-522(in

Chinese).

[13] Everett M, Chen Y, How J P. Motion planning among dynamic, decision-making agents with deep reinforcement learning [C]// Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2018: 3052-3059.

[14] Mueller E, Kopardekar P, Goodrich K. Enabling airspace integration for high-density on-demand mobility operations[C]// Proceedings of 17th AIAA Aviation Technology, Integration, and Operations Conference. AIAA, 2017: 1-24.

[15] Banaszak M, DzieRcielski M, Nijkamp P, et al. Geography in motion: hexagonal spatial systems in fuzzy gravitation [J]. Environment and Planning A: Economy and Space, 2018, 51(2): 1-10.

[16] Champasak P, Panagant N, Pholdee N, et al. Self-adaptive many-objective meta-heuristic based on decomposition for many-objective conceptual design of a fixed-wing unmanned aerial vehicle [J]. Aerospace Science and Technology, 2020, 100(2): 1-11.

[17] **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): 1-26.
Wang Xiangke, Liu Zhihong, Cong Yirui, et al. Miniature fixed-wing UAV swarms: review and outlook [J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(4): 1-26(in Chinese).

[18] He S, Shin H, Tsourdos A. Computational guidance using sparse gauss-hermite quadrature differential dynamic programming[C]// Proceedings of 21st IFAC Symposium on Automatic Control in Aerospace (ACA). 2019: 13-18.

[19] Littman M L. Markov games as a framework for multiagent reinforcement learning[C]// Proceedings of 11th International Conference on Machine Learning. 1994: 157-163.

[20] Papoudakis G, Christianos F, Rahman A, et al. Dealing with non-stationarity in multi-agent deep reinforcement learning [J]. arXiv preprint, arXiv: 1906.04737v1, 2019: 1-8.

[21] Stahl D O, Wilson P W. On players' models of other players: theory and experimental evidence[J]. Levines Working Paper Archive, 1995, 10(1): 218-254.

[22] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning[C]// Proceedings of 36th International Conference on Machine Learning. 2019: 1-14.

[23] Jang E, Gu S X, Poole B. Categorical reparameterization with gumbel-softmax[C]// Proceedings of 5th International Conference on Learning Representations. 2017: 1-13.

（编辑：李瑾）