电 子 科 技 大 学

学术学位研究生学位论文开题报告表

攻读学位级别： □博士 ☑硕士

学科专业： 控制科学与工程

学 院： 航空航天学院

学 号： 202221100238

姓 名： 唐 媛

论文题目： 深度强化学习在边缘无人

集群控制中的应用研究

指导教师： 刘 民 岷

填表日期： 2023 年 11 月 15 日

电子科技大学研究生院

1. 学位论文研究内容

|  |  |  |
| --- | --- | --- |
| 课题类型 | | □基础研究 ☑应用基础研究 □应用研究 |
| 课题来源 | | □纵向 ☑横向 □自拟 |
| 学  位  论  文  研  究  内  容 | 学位论文的研究目标、研究内容及拟解决的关键性问题（可续页）  **一、研究目标**  边缘是相对于中心而言的，边缘作战作为一种作战形式可以说由来已久，任务环境边缘，即该类体系执行的是高、远、边、深和快环境下的严酷任务。网络信息条件下的边缘作战是利用高速网络、电子通信和人工智能等高新技术为网络化和去中心化的边缘赋能，通过广泛的信息分享和团队协作使边缘具备自主发现任务、自主寻找资源、自主决定行动、自主调整改变和自主评估效果的敏捷反应能力。  无人机(Unmanned Aerial Vehicle)作为一种新型的具有重量轻、体积小、机动性高、成本低等特点轻型航空设备。无人机用于观察和战术规划。这项技术现在可用于应急响应领域，以协助机组人员。然而，随着现代战争战场环境和任务要求的日益复杂，许多复杂作战任务已经无法交由单架无人机来完成。  无人机集群（Unmanned Aerial Vehicle Swarm，UAVS）已经成为无人机研究的一个重要方向，通过有效的协作，无人机集群系统可以涌现出比人工系统更卓越的自主性、协调性和智能性[4]，成为未来空中作战的一种主要形式。相比单架无人机的作战应用，无人机集群通过协同合作的方式，能够以更高的效率完成任务,同时具有更好的适应能力、容错能力。无人机集群控制是集群高效作战的重要保证，无人机集群控制中的路径规划和集群避障是确保整个集群安全、高效执行任务的关键要素。路径规划使得无人机能够智能选择最优路径，不仅考虑了目标位置，还综合考虑了飞行距离、能耗等因素，以最小化任务执行成本，即保证了集群在有限资源下能够更有效地执行任务，提高了任务的执行效率。而集群避障是为了确保无人机在执行任务时能够安全避免障碍物。面对复杂的环境，无人机集群能够及时发现障碍物，通过协同规避机制，使得整个集群能够更加安全地运行。  深度强化学习（Deep Reinforcement Learning，DRL）是深度学习（Deep Learning， DL）和强化学习（Reinforcement Learning， RL）的结合，即融合了深度学习的结构和强化学习的思想，具备强大的泛化能力。通过在训练中学习到的模型和策略，能够使智能体适应新的、未知的环境。深度强化学习在集群控制技术领域是一种十分重要的方法，它使得集群可以自主地对复杂环境进行探索，并从探索到的历史经验中学习更好地集群行为和避障行为。  目前，在信息化、边缘化作战场景中，人工智能的发展给无人机集群控制领域的研究提供了新的思路，如何将智能算法应用于无人机集群控制，保证集群系统具有自主性、灵活性以及自适应性已成为国内外科研的热点方向。针对未知复杂的边缘作战环境，早期的强化学习方法不能适用于这种连续的动作空间，在时间上具有一定的迟滞性，又需要处理大量的历史探索经验，所以采用深度强化学习方法去解决这些问题，使之既拥有了强化学习中的决策能力，又具备基于深度学习中神经网络的数据处理能力。  因此，本文的研究目标是：在未知复杂的边缘作战环境下，将深度强化学习应用在无人机集群控制中，能够使无人集群在复杂的无人边缘作战环境中实现集群路径规划和动态避障，从而提高无人集群应对未知复杂边缘作战环境的智能化水平。  **二、研究内容**  在本开题报告中，单无人机的自主控制为路径规划和自主避障，无人集群控制为集群路径规划和动态避障。所以，单无人机控制与无人集群控制无论从任务特点还是研究方法来说都具备相似性。  **1、以单智能体深度强化学习为基础，研究无人机在边缘环境下的自主控制问题**  单个无人机的路径规划和避障方法通过一定程度的扩展能够运用到无人机编队协同避障上来。无人机集群协同避障属于多智能体领域，并且研究方法又涉及到深度强化学习领域，整个多无人机系统较为复杂，而单个无人机的路径规划和避障任务较为简单，因此有必要首先对单无人机的路径规划和避障进行研究。在单无人机避障导航中，主要考虑两个问题：一是需要将无人机导航至目标区域，二是无人机在航行过程中遇到障碍物时能实时避障。  **2、以多智能体深度强化学习为基础，研究在边缘环境下无人集群控制问题**  在多无人机领域，无人机之间常以编队的形式协作执行任务，与单无人机相比，在灵活性、鲁棒性和容错性和上具有明显的优势，是未来无人机执行复杂任务的重要模式。在上述的单无人机路径规划和避障的基础上，拟采用多智能体深度强化学习算法控制多无人机系统形成编队，并在航向目标的过程中保持一定的编队队形，完成在边缘环境下的协同避障任务。  **三、拟解决的关键性问题**   1. **基于单无人机与无人机编队问题的马尔科夫决策（博弈）建模**   针对具体问题建立无人机运动模型，建立无人机视域模型，确定无人机的动作空间、状态空间和奖励函数(马尔可夫决策建模)，同时考虑无人机的通信能力，确定无人机编队的联合动作空间、状态空间和奖励函数(马尔可夫博弈建模)，采用深度神经网络对其进行特征提取。   1. **基于单智能体深度强化学习的无人机路径规划和自主避障**   将任务初步设定为无人机由初始区域航向目标区域的巡航任务，并且要求无人机在航行过程中避开障碍物。基于单智能体深度强化学习算法对无人机进行训练，使其能完成未知环境下的避障导航任务。   1. **基于多智能体深度强化学习的无人集群路径规划和动态避障**   在单无人机路径规划和避障的基础上，考虑无人机集群从无序状态以一定的编队队形航向目标区域，并且在航行区域中存在障碍物。基于多智能体深度强化学习算法对无人机编队进行训练，使其能在避障过程中完成队形变换以成功避开障碍物并到达目标地。 | |

1. 学位论文研究依据

|  |
| --- |
| 学位论文的选题依据和研究意义，国内外研究现状和发展态势，主要参考文献，以及已有的工作积累和研究成果。（2000字）   1. **选题依据和研究意义**   依托于教研室项目：“XXX 边缘指控战--多无人机作战”。  边缘作战概念中，边缘是相对于中心而言的，边缘作战作为一种作战形式可以说由来已久，任务环境边缘， 即该类体系执行的是高、远、边、深和快环境下的严酷任务。网络信息条件下的边缘作战是利用高速网络、电子通信和人工智能等高新技术为网络化和去中心化的边缘赋能，通过广泛的信息分享和团队协作使边缘具备自主发现任务、自主寻找资源、自主决定行动、自主调整改变和自主评估效果的敏捷反应能力[1]。  无人机（Unmanned Aerial Vehicle，无人机）是一种由动力驱动、机上无人驾驶、可重复使用的航空器[2]，具备自主飞行和独立执行任务能力的新型作战平台，无人机最早在上个世纪二十年代出现，早期主要用于军事领域，凭借自身优势，在信息化战中被广泛应用，不仅能够执行军事侦察、监视、空中预警，实施干扰，通信中继，搜索、目标指向，毁伤评估等非攻击性任务，而且还能够执行对地攻击和目标轰炸等作战任务[3,4]。相较于其他航空器而言， 无人机具有体积小、造价低、环境要求低、生存能力强等优点[5]，非常适合代替人类执行一些危险任务,无人机的使用有效降低了飞行员的强度、减少了不必要的伤亡，例如在两伊战争、海湾战争等局部战争中，无人机初步崭露头角；阿富汗战争中“捕食者”挂装“海尔法” 反坦克导弹的成功应用，宣告了无人战斗机时代的到来[6]；纳卡冲突中阿塞拜疆无人机袭击亚美尼亚士兵的画面在互联网上曝光，无人机在战场上的杀伤力再次让全世界惊叹。随着无人机技术的快速发展，越来越多的无人机将应用在未来战场。  随着战场环境、空战战术和无人机应用模式复杂程度的提高，体积、载荷、能源等条件的限制，单架无人机的应用模式已难以充分应对现代战争的需求[7,8]。为了达到整体高效、高性价比、高安全性的需求，美国空军科学顾问委员会指出，单架无人机作战模式应逐步向多架无人机作战模式转变，无人机应当以机群的方式协同工作，而不是单独行动[9]，由多架相同或不同型号的 无人机 组成多无人机系统，协同作业，共同完成任务，多无人机协同模式极大弥补了单无人机模式下的战力不足、资源短缺和策略单一等缺陷。无人机群由多种型号的有人或无人机组成的混合机群，甚至无人机与其他有人或无人作战机器如无人战斗车辆等组成的混合异构群组，必然是未来战争无人作战机器的主要作战形式。  强化学习(Reinforcement Learning，RL)所具有的试错和延迟回报的特点使其拥有优秀的决策能力，在人工智能领域得到了广泛的应用。传统的强化学习只能应对状态空间和动作空间离散的简单情况，然而在比较复杂的现实任务中状态空间和动作空间连续，传统强化学习对其往往“束手无策”。深度学习(Deep Learning，DL)具备强大的表征能力和泛化性能，可以应对高维的数据输入，能利用神经网络对数据进行有效的特征提取，在学术界和工业界得到了普及。深度强化学习(Deep Reinforcement Learning，DRL)同时具备深度学习和强化学习的优点，使得智能体能够在复杂的高维状态空间和动作空间中进行端到端的感知决策，与此同时，随着多智能体领域的发展，深度强化学习也在多智能体环境中得到了扩展[10]。  无人机编队在执行任务过程中常保持一定的编队队形，并且在任务驱动模式下(任务改变、编队成员数量增减、遇到障碍物)按要求变换队形[10]。其中最为复杂的编队控制及协同避障问题是多智能体领域的重点研究方向，传统的编队控制及协同避障方法往往需要很多理想的假设条件，并且大都将协同规划层与控制层分离，这在实际应用中具有一定的局限性。另外，在面对复杂的作战任务和难以预测的作战环境时，无人机编队处于高度动态的对抗环境，此时需要无人机编队对当前态势进行认知和评价并迅速采取合适的协同策略，这对系统的实时性与智能性要求较高，无人机编队智能化水平的高低决定着能否在未来空战中占据绝对的战场优势，因此，对无人机集群的研究具有极其重要的意义。  人工智能的发展给多智能体领域的研究提供了新的思路，其中深度强化学习算法可通过规划层直接输出控制指令，进而使得无人机编队能在未知的动态环境中快速地做出最优的决策，增强无人机集群对环境的适应能力，提高无人机集群的智能化水平。因此本文针对未来无人机集群协同控制的作战需求，基于深度强化学习算法，研究无人机集群的协同控制问题，以期提高无人机集群的环境适应性和智能性，为未来无人机集群在监视侦察、协同打击和合围歼灭等方面提供新的战术和战法。   1. **国内外研究现状和发展态势**   近年来，各军事强国着眼在复杂战场环境下执行作战任务的新需求，以全面应付多维度、多元化、多层面现实安全威胁为导向，以有效制衡潜在作战威胁对象反介入/区域拒止 能力为牵引，以进一步强化态势认知、自主协同和体系集成为重点，积极推进无人机集群系统颠覆性技术研究[11]。尤其是以美国为首的世界军事强国正通过顶层设计、理论研究、项目规划、关键技术攻关以及演示验证等手段来促进无人机集群系统的快速发展[12]。美国国防部及相关部门，重点依托以机器学习、强化学习、自主作战等为代表的人工智能技术，着力推进各类无人机集群系统的研究、实验及演示验证工作。目前，相关国家的无人机集群项目的研发已取得阶段性成果，并引起外界的高度关注。  无人机集群路径规划问题获得了国内外的众多学者的重视和研究。国内外科研人员根据路径规划的多样化需求提出了多种算法。这些算法在各自所针对的领域有着独特的优势，固然在面临不同环境时也有着对应的不足之处。路径规划首要的准备工作是对规划空间进行描述，才能得到可经由计算机处理的数学模型，这里可以使用的方法包括基于路线图的表示方法，比如通视图法、Voronoi 图法概率地图法等，或者使用基于格栅图的表示方法。关于多无人机或机群协同航路规划研究现状，J.D.Wolef 等人[13]着重研究采用编队控制器控制以保持编队队形的多无人机编队控制方法。刘敏, 孔繁峨等人[14]提出了一种基于 A\*算法的可消解冲突的多无人机协同航路规划算法，可以将节点扩展过程加入传统 A\*算法的约束条件中，从而达到消解冲突的目的，为了达成多条航迹之间的时间协同，再将无人机飞行速度通过扩展不同的步长进行反映。YongBo Chen 等人[15]的方案则是首先得到集群航路，再利用人工势能法让僚机的路径靠近得到长机与僚机之间的运动模型，最终达到保持队形的效果。Jiangsong Lu 等人[16]在解决多无人机时间协同航路规划的问题上使用了人工免疫算法和蚁群算法，通过将以上两者的结合，在为单架无人机完成路径规划后，再通过约束足时间协同的蚁群算法选择出其中的最优路径。  无人机集群编队是个较为前沿的科研领域，在扩展原先单架无人机的机动半径、提升作战效率方面有着深远的发展前景，是近年来国内外研究人员的研究热点和重点发展方向。人们对无人机的各项功能要求随着国内外无人机研究和应用的不断深入也日渐提高。单台无人机无法胜任高难度的复杂任务，因此无人机集群编队系统应运而生。多年来，人们通过对技术的理解和从分析生物群体的社会现象中得到一定的启发，逐渐发觉无人机编队可以用来弥补单个无人机的不足。Seungkeun Kim 等人[17]提出了基于行为方法的三维最优控制器的无人机编队控制器，该方法假定各个无人机能够获得相互的位置、速度等信息，建立了各无人机运动学模型。因为该方法基于无人机基础行为，所以能够对不同的输入作出迅速的响应。 Krishnaprasad 等人[18]提出了一种被称为“李氏群设置”的基于编队内无人机载荷以及路径规划方案的简化版控制系统，其核心原理是通过Frenet-Serret 方程来对无人机在二维平面内的飞行轨迹进行分析。针对复杂多变的任务需求，北京航空航天大学的朱杰斌等人[19]提出了基于无人机控制策略的无人机编队飞行的相关控制策略。南京航空航天大学的何真[20]等人提出了一种基于分布式的无人机编队队形控制系统。该系统在对无人机编队飞行数学模型进行梳理和分析后，将原本复杂的多模块总控制系统模型分解成相对简单的两个部分，分别是解耦和耦合模型，再对以上两个模型分别设计各自的控制器系统，最终对无人机编队控制器进行了简化。  在过去的十年中，人工智能（AI）领域迅猛发展，并取得了令人瞩目的成果。人工智能在图像识别，语音识别和高度复杂的游戏方面取得了重大成就。深度学习属于机器学习的范畴，而机器学习是人工智能（AI）的子集。深度强化学习算法在20世纪80年代末开始发展，但是直至 2009 年才有专家学者将 Q-Learning算法与无人机应用结合起来。目前，已经有研究人员将深度强化学习与无人机系统结合，解决了许多无人机系统应用上的问题。郭万春等[21]使用了一种深度强化学习的无人机反追击自主机动决策方法，该方法训练的智能无人机胜率超过传统方法训练的智能无人机；深度强化学习算法在无人机的其他方面也有很多应用，如基于深度强化学习的多无人机协同、空战机动、空战决策、路径规划与避障、目标跟踪等。李东华等[22]提出了一种基于多智能体 Q-Learning 算法的飞行路径规划算法，该算法采用两个能力不同、能够互补的智能体分别对飞行区域进行局部和全局路径规划，同时对状态和动作空间进行了划分和抽象，减少了状态的数量，有效地解决了因无人机飞行区域的扩大和状态的细分而产生的维数灾难问题。传统的强化学习算法受限于策略表征能力，只能处理一些简单的、维数不高的决策问题，深度强化学习的出现打破了这一限制，使其在无人机系统应用上更加广泛。  最后，通过国内外研究现状综合分析，无论对单无人机还是无人机集群来说，现阶段的避障方法存在容易陷入局部最优、假设条件较难满足等问题，导致应用场景受限，使得无人机难以适应复杂的动态环境。至于在无人机集群作战领域，目前对其仅处于理论研究层面，在实际应用方面尚未成熟，且现阶段的研究方法并未体现无人机集群控制编队的高智能化水平。例如在目标围捕场景中，一般假设目标是静止的或者按照预定的轨迹运动而不具有较强的逃逸能力，目标并不具有“智能性”，这导致编队的智能化水平较低，难以适应瞬息万变的战场环境。  随着AlphaGO[23]和AlphaGO Zero[24]在围棋领域的成功，人工智能得到了广泛关注，特别是深度强化学习已经在游戏[25]、机器人[26]、自动驾驶[27]、自然语言处理[28]等领域获得了丰硕的研究成果。深度强化学习具备强大的自学习与自进化能力，能在高维空间实现端到端的感知决策[29,30]，完美适用于复杂的边缘作战环境下的无人机集群感知决策场景，因此本文基于深度强化学习，研究无人机集群控制在未知动态边缘环境下的相关问题，提高无人机集群作战的智能化水平，让无人机集群控制实现真正的“智能”。   1. **主要参考文献** 2. 张维明.指挥控制的新范式：边缘指挥控制[J].指挥信息系统与技术,2021.12(12)1-7. 3. 曹菊红,高晓光.多架无人机协同作战智能指挥控制系统[J]. 火力与指挥控制, 2003, 28(005): 22-24,30. 4. Tang J, Piera M A, Ling Y, et al. Extended Traffic Alert Information to Improve TCAS Performance by means of Causal Models[J], 2015, 2015(PT.15): 1-11. 5. 杜永浩,邢立宁,蔡昭权.无人飞行器集群智能调度技术综述 [J].自动化学报,2020,46(02):222-241. 6. 樊邦奎. 国外无人机大全[M]. 航空工业出版社 2001. 7. 秦枫. 无人机指挥控制系统设计与实现[D]. 国防科学技术大学, 2017. 8. Chen YB, Yu JQ, Su XL, et al. Path planning for multi-UAV formation[J]. Journal of Intelligent & Robotic Systems, 2015, 77(1): 229-246. 9. Arslan G, Wolfe J D, Shamma J, et al. Optimal planning for autonomous air vehicle battle management[C]//Proceedings of the 41st IEEE Conference on Decision and Control, 2002. IEEE, 2002, 4: 3782-3787. 10. 胡纯德, 祝延军, 高随祥. 基于人工免疫算法和蚁群算法求解旅行商问题[J]. 计算机工程与应用, 2004(34):63-66. 11. 关旭宁, 魏瑞轩, 郭庆等. 多无人机时序到达协同控制方法[J]. 电光与控制, 2014, 21(1): 18-22. 12. 朱收涛, 曹林平, 翁兴伟. 采用改进粒子群算法的无人机协同航迹规划[J]. 电光与控制, 2012, 19(12)：76-80. 13. 喻蓉, 刘敏, 孔繁峨等.基于冲突消解多无人机协同航路规划算法仿真[J]. 电光与控制, 2011(06):45-48+92. 14. Kim S, Choi J, Kim Y. Fault detection and diagnosis of aircraft actuators using fuzzy-tuning IMM filter[J]. IEEE Transactions on Aerospace and Electronic Systems, 2008, 44(3): 940-952. 15. Bloch A M, Krishnaprasad P S, Marsden J E, et al. Nonholonomic mechanical systems with symmetry[J]. Archive for Rational Mechanics and Analysis, 1996, 136(1): 21-99. 16. 朱杰斌,秦世引. 无人机编队飞行的分布式控制策略与控制器设计[J]. 智能系统学报(5):392-399. 17. 何真, 陆宇平. 无人机编队队形保持控制器的分散设计方法[J]. 航空学报, 2008(B05):55-60. 18. 范卫刚[1], 袁冬莉[1]. 基于模糊自适应 PID 的无人机纵向姿态控制研究[J]. 计算机测量与控制, 2011, 19(3):586-588. 19. 丁家如, 杜昌平, 赵耀, 等. 基于改进人工势场法的无人机路径规划算法[J]. 计算机应用, 2016, 36(1): 287-290. 20. Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]//Advances in neural information processing systems. 2012: 1223-1231. 21. Mairal J, Bach F, Ponce J, et al. Online dictionary learning for sparse coding[C]//Proceedings of the 26th annual international conference on machine learning. 2009: 689-696. 22. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788. 23. Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey[J]. Computers and electronics in agriculture, 2018, 147: 70-90. 24. Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587):484-489 25. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676):354-359. 26. Bergdahl J, Gordillo C, Tollmar K, et al. Augmenting Automated Game Testing with Deep Reinforcement Learning[J]. 2021. 27. Silva I, Perico D H, Homem T, et al. Deep Reinforcement Learning for a Humanoid Robot Soccer Player[J]. Journal of Intelligent & Robotic Systems, 2021, 102(3):69. 28. Klose P, Mester R. Simulated Autonomous Driving on Realistic Road Networks using Deep Reinforcement Learning[J]. 2017. 29. Shah A, Vuong T. Natural Language Person Search Using Deep Reinforcement Learning[J]. 2018. 30. 唐振韬,邵坤,赵冬斌,朱圆恒.深度强化学习进展:从AlphaGo到AlphaGo Zero[J].控制理论与应用,2017,34(12):1529-1546. 31. 李天旭.基于深度强化学习的多智能体协同算法研究[D].中国矿业大学,2020. 32. **已有的工作积累和研究成果**   1、选题意义和研究路线明确。  2、已阅读大量文献，了解了深度强化学和无人机集群控制相关的知识，并对研究过程有了一定程度的理解。 |

1. 学位论文研究计划及预期目标

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1.拟采取的主要理论、研究方法、技术路线和实施方案（可续页）  **一、拟采用的主要理论**  **1、深度强化学习理论**  深度强化学习是深度学习（Deep Learning，DL） 和 强 化 学 习（Reinforcement Learning， RL）的结合，如图 3.1 所示，具体来说是结合了深度学习的结构和强化学习的思想，侧重点在于强化学习。深度强化学习解决的仍然是决策问题，只是借助神经网络强大的表征能力去拟合 Q 表或者直接拟合策略以解决状态-动作空间过大或联系状态-动作空间问题。    图 3. 1 深度学习与强化学习的结合  深度学习的概念来自人工神经网络 （Artificial Neural Network，ANN）。使用深度学习对数据进行提炼，不需要选择特征、压缩维度、转换格式等数据处理方式，拥有比传统机器学习方法更强的特征表示能力［4］。深度学习模型通常是由多层非线性神经网络运算层组成。每层的输入为该层前一层的输出，通过该方式深度学习模型可以自动地从大量训练数据中学习到高维的抽象特征表示，实现数据的分布式表示。  强化学习问题一般利用马尔可夫决策过程(Markov Decision Process, MDP)对其进行理论建模。MDP由一个五元组组成，为状态空间，包括智能体自身状态和环境状态，为动作空间，为智能体选择动作，系统由状态转移到的概率分布函数，为状态转移过程中的奖励函数，为奖励折扣因子，用于权衡长期奖励与立即奖励。    图 3. 2强化学习标准框架  强化学习的标准框架如图3.2所示，其基本原理是让智能体与环境持续地交互反馈，智能体不断进化从而获得最优或较优的策略。智能体与环境的整个交互过程如下：在任意时刻，智能体根据当前时刻状态和策略选择动作，到达下一状态，同时得到相应的奖励，智能体重复此行为直到回合结束。  按选择策略的方法分类，强化学习可分为基于策略(Policy-Based)和基于值 (Valua-Based)的强化学习。基于策略的强化学习算法直接通过确定性策略或策略的概率分布选择动作，如策略梯度(Policy Gradient，PG)等，基于值的强化学习算法主要是通过值函数来选取策略，如Q-Learning、SARSA(State-Action-Reward-State-Action)、DQN(Deep Q Network)等。结合两者的优点，基于策略-评价(Actor-critic，AC)的强化学习算法(也称为基于AC框架的深度强化学习算法)在很多领域取得了一系列不错的成果，算法采用深度神经网络对状态进行特征提取，actor网络(策略网络)根据当前状态输出动作，critic网络(评价网络)则对该状态下输出的动作价值进行估计，随后不断地更新策略使得智能体具备优秀的感知决策能力。  在多智能体强化学习领域，MDP问题被泛化为马尔可夫博弈问题。马尔可夫博弈仍然可以表示为一个五元组，为状态空间，包括所有智能体的状态和环境状态，为所有智能体的联合动作空间，表示每个智能体的动作空间，为智能体的数量，为多智能体选择联合动作，系统由状态转移到的概率分布函数，为状态转移过程中的奖励函数，为智能体在选择联合动作后获得的奖励，为奖励折扣因子。    图 3. 3多智能体强化学习框架  多智能体强化学习的标准框架如图3.3所示，其基本原理是让多智能体与环境持续地交互反馈，多智能体不断更新彼此的策略从而获得最优或较优的策略。多智能体与环境的整个交互过程如下：在任意时刻，多智能体根据当前时刻状态和策略选择动作，到达下一状态，同时得到相应的奖励，多智能体重复此联合行为直到回合结束。  **2、单智能体深度强化学习相关算法**  PPO (Proximal Policy Optimization) [66]、DDPG (Deep Deterministic Policy Gradient) [67]和SAC (Soft Actor-Critic) 算法[68-69]等是目前比较热门的单智能体深度强化学习算法。  (1) PPO算法  传统PG算法对于更新步长十分敏感，步长过大会导致学习的策略存在较大扰动，不会收敛，步长太小则会降低学习速度。PPO算法是TD与PG的结合，可用来解决连续动作控制问题，通过引入新旧策略的比值，限制策略的更新幅度，从而解决了PG算法训练不稳定的问题。传统PG算法对于更新步长十分敏感，步长过大会导致学习的策略存在较大扰动，不会收敛，步长太小则会降低学习速度。PPO算法是TD与PG的结合，可用来解决连续动作控制问题，通过引入新旧策略的比值，限制策略的更新幅度，从而解决了PG算法训练不稳定的问题。  在传统PG算法中，actor网络更新的目标函数为：   |  |  |  | | --- | --- | --- | |  |  | (3.1) |   对其求微分可得到梯度估计量：   |  |  |  | | --- | --- | --- | |  |  | (3.2) |   网络参数更新：   |  |  |  | | --- | --- | --- | |  |  | (3.3) |   其中为时刻的优势函数，一般通过蒙特卡洛采样获得，为当前策略，为学习率，代表网络参数的更新步长。可以看出，PG算法的actor网络是对式(3.1)的目标函数进行梯度上升更新从而找到最好的优势值。  (2) DDPG算法  DDPG算法是DPG(Deterministic Policy Gradient)[70]与DQN的结合，DDPG摒弃了用概率分布表示策略的方法，转而用一个确定性的函数表示。  DDPG算法中critic网络更新的目标函数同样为TD-error：   |  |  |  | | --- | --- | --- | |  |  | (3.4) |   其中D表示优先经验池，为目标状态动作值函数。  DDPG 算法中actor网络更新的目标函数为：   |  |  |  | | --- | --- | --- | |  |  | (3.5) |   其中最优策略为最大化:。  (3) SAC算法  SAC算法以熵最大化与累积奖励和最大化为目标训练智能体。SAC集成了PPO和DDPG的优点，在优化目标的基础上加入了熵项，增强了探索能力，SAC的策略使用的是与PPO相同的随机策略，基于能量模型进行策略的选择，而且仍然利用了DDPG中采用的经验回放池进行历史数据的存储，采样效率高且算法稳定，能够在多模态奖励函数下寻找到更好的策略。其中“多模态”是指智能体不局限于学习单一任务的单一解决方法，而是学习多样任务的多样解决方法，如编队协同避障任务中存在避障、导航、编队集结和编队保持等任务，SAC算法可以协调好各个任务奖励函数的权重从而让智能体学习合适的策略来完成整个任务。  基于最大熵深度强化学习SAC算法的目标是学习一个策略使得累积奖励值与动作熵的和最大，即：   |  |  |  | | --- | --- | --- | |  |  | (3.6) |   其中，表示当前策略的随机化程度，为熵正则化系数，用于权衡奖励与熵之间的权重。  SAC算法critic网络更新的目标函数为：   |  |  |  | | --- | --- | --- | |  |  | (3.7) |   其中为目标状态动作值函数。  SAC算法actor网络更新的目标函数为：   |  |  |  | | --- | --- | --- | |  |  | (3.8) |   其中为归一化项，actor网络采用梯度下降的方法对更新使得输出策略与基于能量模型的策略散度最小。  **2、多智能体深度强化学习相关算法**  多智能体深度强化学习算法为MADDPG(Muti-Agent Deep Deterministic Policy Gradient)和MASAC(Muti-Agent Soft Actor-Critic)算法等。  （1）MADDPG算法  MADDPG算法是对DDPG算法的扩展，适用于多智能体环境。DDPG是一种连续动作空间的深度强化学习算法，MADDPG通过在每个智能体的训练中共享经验池和目标网络，以实现多智能体协同学习。  由个智能体组成的MAS，为智能体的actor网络参数，为智能体的联合策略，那么智能体的actor网络更新的目标函数，针对随机策略求梯度：   |  |  |  | | --- | --- | --- | |  |  | (3.9) |   其中，表示第个智能体的观测，为观测向量(状态),为第个智能体的联合状态动作值函数，每个智能体可以独立学习自己的，从而完成规定的协同任务。  式(3.9)为随机策略梯度算法，MADDPG算法采用确定性策略，因此将其拓展到确定性策略梯度：   |  |  |  | | --- | --- | --- | |  |  | (3.103) |   其中经验回放池D中元素组成为。  MADDPG算法的critic网络更新的目标函数仍然为TD-error，与DDPG算法相似，只不过是将动作值函数变为联合状态动作值函数：   |  |  |  | | --- | --- | --- | |  |  | (3.14) |   其中为目标状态动作值函数，更新的目标是最小化。  （2）MASAC算法  MASAC算法是基于Soft Actor-Critic（SAC）算法的多智能体版本。SAC是一种用于连续动作空间的深度强化学习算法，而MASAC通过引入额外的关于其他智能体策略的信息，以促进智能体之间的协同学习。  同样，MASAC算法与MADDPG算法相似，MASAC算法采用随机策略梯度更新，因此actor网络更新策略梯度的形式与式(3.9)类似，得到在MASAC算法中智能体更新actor网络需要最小化的目标函数：   |  |  |  |  | | --- | --- | --- | --- | |  |  |  | (3.15) |   MASAC算法critic网络更新的目标函数为：   |  |  |  | | --- | --- | --- | |  |  | (3.11) |   其中为目标状态动作值函数，更新的目标是最小化。  **二、研究方法**  本研究利用相关研究成果，采用理论建模，算法推导，仿真实验的研究方法，开展针对边缘环境下，深度强化学习在无人机集群控制中的应用研究。  首先，深度强化学习理论研究。通过分析所研究的无人机的路径规划和自主避障、无人机集群的协同避障问题的特点，确定解决其问题所应用到的单智能体和多智能体深度强化学习算法。  其次，基于单智能深度强化学习算法的无人机避障导航研究。对单无人机避障导航问题建立马尔可夫决策模型，采用单智能体深度强化学习算法研究无人机在边缘未知动态环境下的避障导航问题，最终训练出合适的网络模型，并与其他传统深度强化学习算法作对比来体现其性能的优劣。  最后，基于多智能深度强化学习算法的无人集群路径规划和动态避障研究。在单无人机路径规划和自主避障研究的基础上，采用多智能体深度强化学习算法研究合作场景下的无人机集群编队协同避障问题，对其建立马尔可夫博弈模型，拟将整个编队协同避障任务分为四个子任务(编队集结、编队保持、靠近与到达目标、避障避碰)分别设计奖励函数，最后通过仿真验证算法的有效性。  **三、技术路线**    图 3. 4 技术路线框图  **四、实施方案**  深度强化学习在边缘无人集群控制的实施方案将涉及多个关键步骤。  首先，选择和设计适用于单无人机自主控制的深度强化学习算法。其次，在此基础上，再选择和设计出适用于集群控制的深度强化学习算法。在算法选择后，通过构建深度神经网络对每个无人机的策略进行训练，使其能够实时感知环境、规划路径，并协同避开障碍物。将利用深度神经网络对集群中每个无人机的策略进行训练，使其能够在协同环境中作出智能决策。最后，将建立合适的仿真环境，以模拟边缘无人机集群控制场景，提供大量的训练数据。 |
| 2.研究计划可行性，研究条件落实情况，可能存在的问题及解决办法（可续页）  **一、研究计划可行性**  对深度强化学习在边缘无人集群控制中的研究，可行性在于其在智能体间协同决策和复杂环境中学习的卓越能力。通过利用深度学习算法，可以使无人机集群在未知、动态的边缘环境下更智能地协同工作，实现高效的任务执行和避障行为。可借助先进的感知、路径规划和集群协同控制方法，结合深度强化学习的优势，提高集群整体性能。在现有深度学习技术支持下，该研究计划在解决无人机集群面临的复杂控制问题上具备较高的可行性，并有望为未来智能边缘计算环境下的无人机集群控制提供创新的解决方案。  **二、研究条件落实情况**  硬件条件：处理器为R7 5800H，内存为16GB，显卡为NVIDIA GeForce RTX 3060。软件条件：利用OpenAI MPE、Unity、pybullet和MATLAB等软件对集群控制进行研究。  **三、可能存在的问题及解决办法**  在研究过程中，针对深度强化学习算法收敛较慢的问题，拟结合一些传统算法去改善算法收敛速度，提升算法整体效率。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2023.11-2024.4 | 国内外研究现状梳理，进行文献调研，完成开题报告撰写，对现有理论、算法和模型进行理论研究和复现 |
| 2024.4-2024.8 | 针对具体需求，对问题进行建模，优化单智能体深度强化学习算法、多智能体深度强化学习算法，并设计相关仿真验证实验，进行理论验证并分析相关结果，进一步改进相关理论研究内容 |
| 2024.8-2024.12 | 针对深度强化学习在无人机集群控制中的路径规划和动态避障问题的应用研究，设计相关仿真验证实验，进行理论验证并分析相关结果，进一步改进相关理论研究内容 |
| 2024.12-2025.3 | 根据前期成果，完善整理硕士学位论文，准备毕业答辩 |
| 预  期  创  新  点  及  成  果  形  式 | **1、创新点：**  (1)针对深度强化学习算法收敛较慢的问题，拟结合一些传统算法去改善算法收敛速度，提升算法整体效率  (2)由单智能体深度强化学习算法，引申到多智能体深度强化学习算法结合传统算法，实现无人机集群在复杂边缘动态环境下的协同避障。  **2、成果形式**  算法、论文以及申请相关专利一项。 | |

1. 开题报告审查意见

|  |  |  |  |
| --- | --- | --- | --- |
| 1.导师对学位论文选题和论文计划可行性意见，是否同意开题： | | | |
| 导师（组）签字： 年 月 日 | | | |
| **2.开题报告考评组意见** | | | |
| 开题日期 |  | 开题地点 |  |
| 考评专家 |  | | |
| 考评成绩 | 合格 票 基本合格 票 不合格 票 | | |
| 结 论 | □通过 □原则通过 □不通过  **通过：**表决票均为合格  **原则通过：**表决票中有1票为基本合格或不合格，其余为合格和基本合格  **不通过：**表决票中有2票及以上为不合格 | | |
| 考评组对学位论文的选题、研究计划及方案实施的可行性的意见和建议： | | | |
| 考评组签名：  年 月 日 | | | |
| **3.学院意见：** | | | |
| 负责人签名： 年 月 日 | | | |