连通保持约束下的无人机集群动态覆盖控制

学生姓名：xxx　　学　　号：xxx
学　　院：xxx　　指导老师：xxx
指导单位：电子科技大学 xxx 学院

摘要 基于无人机集群的动态覆盖控制可以快速部署到高风险或高污染的环境中，具有较大的应用前景。在恶劣环境中，通信网络的连通保持问题是集群执行任务的关键。因此，以点覆盖作为主要研究对象，本文使用多智能体强化学习方法解决在连通保持约束下的无人机集群的覆盖轨迹规划问题，主要工作如下：

1) 建立了基于覆盖能量和覆盖功率的无人机集群动态点覆盖任务模型；
2) 基于上述模型，使用 MADDPG 方法解决了无连通约束下的轨迹规划；
3) 基于 CTDE 框架提出了动作矫正器的连通保持算法；

关键词 动态覆盖控制，连通保持，多智能体深度强化学习

Abstract Coverage control based on UAV swarms can be quickly deployed to high-risk or high-pollution environments, and has great application prospects. In harsh environments, the problem of connectivity preservation of the communication network is the key to the task of the cluster. Therefore, taking point coverage as the main research object, this thesis uses a multi-agent reinforcement learning method to solve the coverage trajectory planning problem of UAV swarms under the constraint of connectivity preservation. The main work is as follows:

1) This thesis establishes a dynamic point coverage model for UAV swarms based on coverage energy and coverage power;
2) Based on the above model, this thesis uses the MADDPG method to solve the trajectory planning without the constraint of connectivity preservation;
3) Based on the CTDE framework, this thesis proposes a connectivity preservation algorithm based on motion correctors;

Keywords dynamic coverage control, Connectivity preservation, Multi-agent Deep Reinforcement Learning (MADRL)

第一章 绪论

无人机集群是指由一组具有数据采集和信息处理功能的无人机组成的多智能体系统。集群可以通过通信来协调彼此行动，以协作的方式实现任务，表现出单架无人机难以达到的优势[1]。基于无人机集群的动态覆盖控制是近年来的研究热点，其基本任务是使用无线传感器网络对一定的空间进行探测覆盖和信息采集[2]，基本问题是通过优化无人机集群的运动轨迹来提高覆盖性能。

基于无人机集群的动态覆盖方法在提高系统的机动性的同时，也带来了连通性保持上的挑战。集群的连通性取决于集群中的个体之间能否建立直接或者间接的信息通道来协调行动，是恶劣环境中执行任务的关键[3]。动态覆盖会使集群在空间中分散以实现对于任务空间的全面覆盖，连通性保持会限制集群扩展来保持通信连通，二者在动力学行为上相反且矛盾的表现使得连通保持约束下的集群动态覆盖控制设计更为复杂。近年来，多智能体强化学习（MARL）展现了求解具有复杂约束的优化问题的强大性能[4]。对比蚁群、粒子群等群体优化算法[5]，MARL 具有信息处理能力更强，收敛更快等优点。综上所述，本文主要旨在借助 MARL，研究连通保持约束下的无人机集群动态覆盖控制的相关问题。

第二章 研究基础及概述

2.1 覆盖控制

设 K 维任务空间 $Q \in ¡^{K}$，$q \in Q$ 为其上一点，密度函数使用 $\phi(q,t)$ 表示，$X^t = \{x_1^t, \ldots, x_N^t\}$ 表示集群中所有传感器的轨迹，传感器的探测能力使用 $f(x^t, q)$ 表示，$g(\cdot)$ 表示代价函数，则目标函数有以下形式：

$$\min_{X(t)} H[X(t)] = \int_{\tau}\int_{Q} g\left(\sum_{x^t \in X^t} f(x^{\tau}, q)\phi(q,\tau)\right) dq\, d\tau \tag{2-1}$$

式(2-1)定义了动态覆盖控制的基本任务，根据密度函数的不同，覆盖控制可以分类为区域覆盖、栅栏覆盖和点覆盖等[6]。本文选取与实际需求更匹配的点覆盖作为主要研究对象。基于式(2-1)所述，本文在 3.1 中将继续对任务进行建模。

2.2 强化学习

强化学习是一种过程学习方法，其方式可以概括为：智能体以试错的方式与未

知环境交互并获得反馈，根据反馈对控制策略进行改进。这种顺序决策可以使用马尔可夫决策过程（MDP）描述[7]。对状态$s_t \in \mathrm{S}$，智能体可以选择做出不同的动作$a_t \in \mathrm{A}$，此时环境会根据状态转移函数P 生成下一个状态，$s_{t+1} \sim \mathrm{P}(s' \mid s_t, a_t)$，并给出奖励$R_{t+1}$。使用折扣率加权，折扣长期回报为$G_t = \sum_{t=0}^{\infty} \gamma^t R_t$。智能体的目标是在未知概率P 的前提下，学习到一个策略$\pi : \mathrm{S} \to \mathrm{A}$，来最大化期望回报。

“Actor-Critci”架构[9]使用两组参数分别表示策略函数$\pi(a \mid s, \theta)$与动作价值函数$Q_\pi(s, a, \psi)$。Critic 函数的训练过程与 SARSA 方法[8]类似，以 TD-error 作为损失函数，随着 TD-error 收敛到 0，Critic 函数对动作价值的估计也越来越准确。Actor 函数的训练则与 REINFORCE 方法略有不同，AC 架构舍弃了 MC 采样，而是使用 TD-target 估计动作价值函数，$G_t \leftarrow \sum_{a_t} R_t + \gamma Q_\pi(s_{t+1}, a_t, \psi)$，其中 Q 值由 Critic 函数给出估计。由于R_t的无偏性，最终 Actor 会与 Critic 同步收敛到π^*和Q^*。AC 框架整体上加快了收敛速度，已经成为目前强化学习最常使用的框架。深度确定性策略梯度算法[10]是 AC 框架在连续空间上的拓展，其示意图如图 2-1 所示。

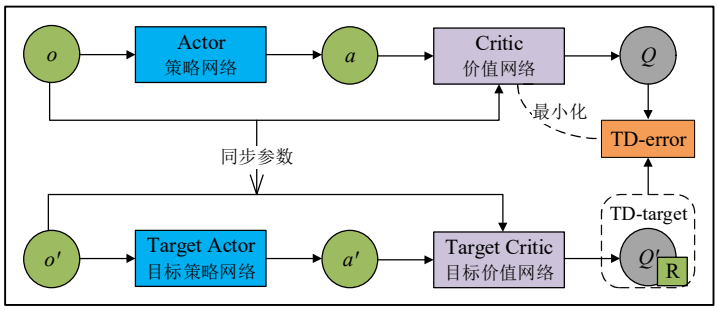


图 2-1 深度确定性策略梯度算法示意图

2.3 集群连通

考虑一个由N个无人机组成的集群，个体的通信半径为R_i，集群的通信网络使用图$\mathrm{G} = (\mathrm{V}, \mathrm{E}, A)$表示。其中，$\mathrm{V} = \{v_1, \ldots, v_N\}$为结点集，E 为边集，$A$为图G的邻接矩阵。 v_i与v_j之间的距离使用d_{ij}表示，若$d_{ij} < R_i + R_j$，则$\left(v_i, v_j\right) \in \mathrm{E}$，$a_{ij} = 1$。此时，表示$v_j$的信息可以被$v_i$接收。

若对$\forall v_i, v_j \in \mathrm{V}$，均存在一条路径使$v_j$对$v_i$可到达，则称G是强连通的。此时表明集群内的信息可以自由流通，任意结点都能接收集群中所有结点的信息。对非负邻接矩阵A，有

$$[A^m]_{ij} = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \mathrm{L} \sum_{k_{m-1}=1}^{n} a_{ik_1} a_{k_1 k_2} \ldots a_{k_{m-1}, j} \tag{2-2}$$

若$[A^m]_{ij} > 0$，则v_i与v_j之间存在$m-1$个结点使得$a_{ik_1} a_{k_1 k_2} \ldots a_{k_{m-1}, j} > 0$，即$v_j$到$v_i$存在

长度为m的路径。若$\left[\sum_{m=1}^{N-1}A^m\right]>0$，则存在路径使$v_j$对$v_i$可到达。若$\sum_{m=1}^{N-1}A^m$为正矩阵，则G是强连通的[11]。

第三章 基于多智能体强化学习的动态覆盖控制

3.1 无人机集群动态覆盖控制模型

首先，使用简化 Neyman-Pearson 模型[12]，即二维钟型函数来刻画传感器的信息获取率。传感器探测范围为一个圆，圆心与无人机位置重合。传感器有效探测半径为r，x表示传感器位置，p表示目标点的位置，$d=\|x-p\|_2$为传感器与目标点的距离。则任务空间上的信息获取率为：

$$g(x,p)=\begin{cases}\exp\left(-\left(\frac{x-p}{r}\right)^2\right) & ,d\le r\\ 0 & ,d>r\end{cases} \tag{3-1}$$

其次，使用二阶积分器对无人机动力学建模，设风阻系数为η，无人机质量为m，x_1表示无人机位置，x_2表示速度，依据动力学公式，有：

$$\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\end{pmatrix}=\begin{pmatrix}0 & 1\\ -\eta & 0\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\begin{pmatrix}0\\ 1/m\end{pmatrix}u \tag{3-2}$$

对覆盖任务，在任务空间$W\in\mathbb{R}^K$上离散地分布着M个待覆盖目标点(PoI)，PoI的位置使用$p_j\in W, j\in\{1,\dots,M\}$表示。执行覆盖任务的无人机集群由$N$架无人机$V=\{v_1,\dots,v_N\}$组成，其位置表示为$x_i^t\in W, i=\{1,\dots,N\}$。无人机$v_i$在 PoI 点$p_j$上提供的探测功率等于峰值探测功率$M_p$与传感器信息获取率之积。以时间为变量，对探测功率积分，得到无人机在一段时间内在任务空间上提供的探测能量。无人机v_i在$[0,t]$对 PoI 点p_j提供的探测能量为：

$$E_{i,j}^t=\int_0^t P_{i,j}(x_i^\tau,p_j)\,d\tau \tag{3-3}$$

将上式对i求和，集群提供的覆盖能量为：

$$E_{N,j}^t=\int_0^t\sum_{i=1}^{N}P_{i,j}(x_i^\tau,p_j)\,d\tau \tag{3-4}$$

每个 PoI 点p_j都有一定的覆盖能量需求E_j^*，当$E_{N,j}^t\ge E_j^*$时，认为对p_j的覆盖任务完成。当不等式对$\forall j=1:M$成立，认为任务覆盖任务完成。覆盖控制场景如图 3-1 表示，算法的任务是合理的规划集群的轨迹，使得在尽可能短的时间内，每个 PoI 都能被投射足够的探测能量。

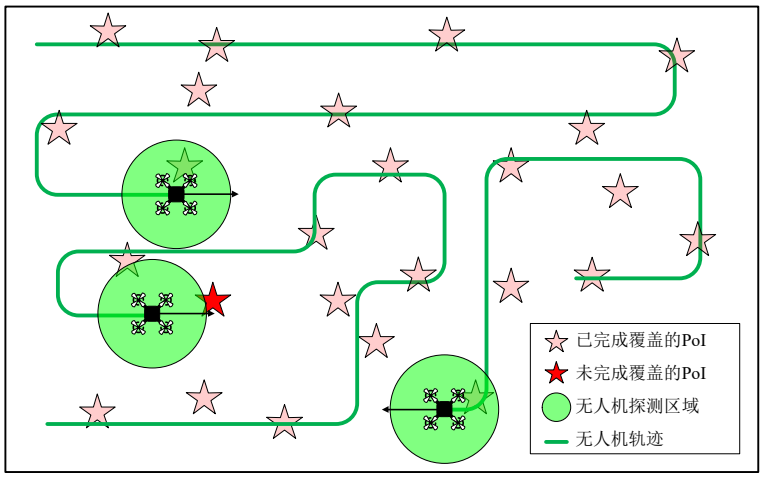


图 3-1 动态点覆盖控制场景示意图

3.2 基于 MARL 的动态覆盖控制轨迹规划算法

动态覆盖控制系统由无人机和 PoIs 两部分组成，因此描述动态覆盖控制系统的状态变量可以定义为：

$$s^t=\left[\left\{x_i^t,vel_i^t\right\}_{i=1:N},\left\{p_j,E_j^*-E_{N,j}^t\right\}_{j=1:M}\right] \tag{3-5}$$

以无人机为智能体，其观测包括自己的位置与速度、其他个体的位置、PoIs 的位置和当前探测需求。以自身为观测坐标系原点，智能体观测为：

$$o_i^t=\left[x_i^t,vel_i^t,\left\{x_j^t-x_i^t\right\}_{j=1:N,j\neq i},\left\{p_j-x_i^t,E_j^*-E_{N,j}^t\right\}_{j=1:M}\right] \tag{3-6}$$

对于覆盖控制场景，无人机的移动是连续的，基于值的 DQN 方法并不适用。相比之下，以 DDPG 为代表的策略梯度方法更适用。本文所采用的 MADDPG 算法[13]是 DDPG 在集中式训练和分散式执行上的拓展。每个智能体 v_i 都有两个主网络 Actor $\mu_i(a\,|\,o,\theta_i)$ 和 Critic $Q_i(o_i,a_i,\psi_i)$ 和对应的两个目标网络 $\hat{\mu}_i$ 和 $\hat{Q}_i$。主网络和目标网络参数周期同步。Critic 网络的训练与有监督学习类似，以 TD-target 为标签值，以 $\hat{Q}_i$ 生成的动作价值为预测值，使用均方误差为损失函数，对 batch B，Critic 的损失函数如式(3-5)所示。使用梯度上升构造 Actor 的损失函数，对 batch B，Actor 的损失函数如式(3-6)所示。

$$L(\psi_i)=\frac{1}{|\mathrm{B}|}\sum_{\{o,a,r,o'\}}\left(Q_i\left(o_i,a_i,\hat{\psi}_i\right)-y\right)^2 \tag{3-5}$$

$$y=r_i+\gamma\hat{Q}_i\left(o_i',a_i',\hat{\psi}_i\right)\Big|_{a_i'=\mu_i(o_i',\theta_i)}$$

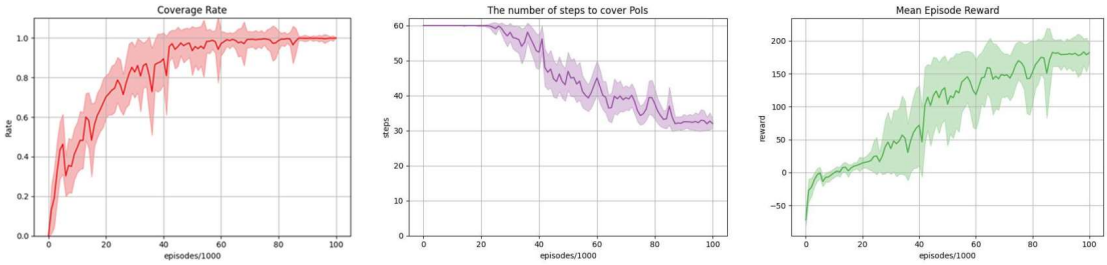
$$L(\theta_i)=-\frac{1}{|\mathrm{B}|}\sum_{o\in\mathrm{B}}Q_i\left(o_i,\hat{\mu}_i\left(o_i,\hat{\theta}_i\right),\psi_i\right) \tag{3-6}$$

动态点覆盖控制的奖励函数如式(3-7)所示。其中，第一项 R_p 表示单个 PoI 完成覆盖后给予的奖励，$M_d^t=\left\{j\middle|E_{N,j}^t\geq E_j^*\right\}$ 表示 t 时刻已经完成覆盖的 PoIs 的集合；第二项 R_d^t 表示完成全部覆盖任务的奖励，其只在 $\left|M_d^t\right|=M$ 时非零；第三项表示对所有未完成覆盖的 PoIs，计算与其最近的无人机之间距离，以 R_s 为系数后作为距离惩罚。

$$R^t=R_p\left(\left|M_d^t\right|-\left|M_d^{t-1}\right|\right)+R_d^t-R_s\sum_{j\in M_d^t}\left\|x_i^t-p_j\right\|_2 \tag{3-7}$$

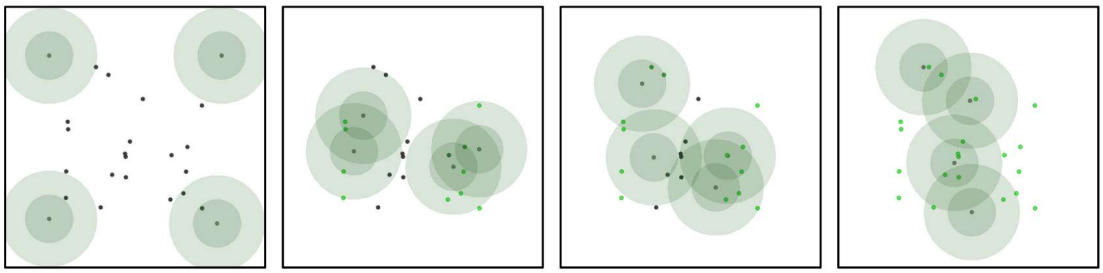
3.3 实验验证与分析

考虑一个二维点覆盖场景，其中包含 20 个 PoIs，其位置每幕随机生成。执行任务的集群包含 4 架无人机，每架无人机最大速度为 1.0。任务空间为 x 轴和 y 轴范围均是[−1,1]的 2*2 方形区域。训练时 batch size 为 1024，策略网络和价值网络中间层神经元个数为 128，学习率均为 0.005，目标网络每隔 100 步更新，更新率为 0.75。折扣比 $\gamma=0.95$，单幕最大步数为 60，重放缓存区包含 1024*60 个状态转换。算法在 100k 幕上的收敛曲线如图 3-2 所示。



(a) (b) (c)

图 3-2 基于 MARL 的覆盖控制算法训练效果 (a) 覆盖率；(b)每幕步数；(c)任务奖励



(a) (b) (c) (d)

图 3-3 算法可视化效果 (a)$t=0$；(b)$t=12$；(c)$t=22$；(d)$t=32$

图 3-3 对本章算法的部署效果进行了可视化展示。图中黑色圆点为 PoIs，当 PoI 的覆盖能量上升时，其颜色会从黑到暗绿再到绿，亮绿色的 PoIs 表示已完成覆

盖。无人机在环境中使用同心圆表示，其中半径较小的深绿色表示传感器的覆盖范围，半径较大的浅绿色表示无人机的通信范围。

第四章 连通保持约束下的动态覆盖控制

4.1 连通保持约束

上一章讨论了无连通保持约束下的无人机集群动态覆盖控制问题，即假设了每个智能体都具有全局视野，可以获得集群中任意其他个体和 PoIs 的信息。但这个假设在有些情况下是不现实的。因此，本节智能体模型进行改造，引入如下功能：

1、信息的实时广播。无人机向外广播信息，包括自身位置、自身探测范围内 PoIs 的当前覆盖需求；2、基于连通的信息获取。按照 2.3 中所述模型构造通信网络，存在路径的无人机之间可以获取信息。3、信息存储与更新。观测空间在式(3-6)基础上使用所有它能获取的信息当中的最新值来替代观测。

4.2 基于规则控制器的连通保持算法

本节设计了一个基于规则的连通保持算法，其核心由一个双层控制器，主要由决策控制器、规则控制器和动作预执行器组成。每当决策控制器输出一个动作，首先进入预执行器，得到执行动作后的系统状态 s'。若 s' 失去连通，则规则控制器起作用，生成连通约束力，智能体受到的控制由驱动力和连通约束力共同作用，由此保证系统的连通性。

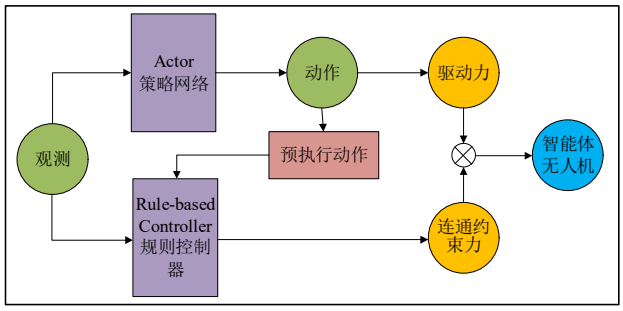


图 4-1 基于规则的双层控制器

连通保持规则设计如下：预执行决策控制器输出 a，若系统失去连通，则在每一对 v_i 与 v_j 之间产生拉力，二者满足 $(v_i,v_j)\in \mathrm{E},(v_i,v_j)\notin \mathrm{E}'$，约束力为

$$f_{r,i}=\alpha\times\ln\left(1+e^{d_{ij}^t-2R}\right)\times\frac{x_j^t-x_i^t}{d_{ij}^t} \tag{4-1}$$

其中，第一项 α 为比例系数，调节约束力大小；第二项 $\ln(1+e^x)$ 为近似线性函

数，但在 $x=0$ 附近不为 0，使用此函数保证了在即使刚刚脱离连通也会产生约束力；第三项为方向向量。当比例系数满足式(4-2)时，规则(4-1)可以保持集群连通性。

$$\alpha > {mvel_{\max}} \Big/ {\ln(1+e^{2vel_{\max}\Delta t})} \tag{4-2}$$

4.3 基于动作矫正器的连通保持算法

集中式的规则控制器其局限性在于，其一，需要更多的硬件支持和计算资源，其二，当集群失去连通，集中式的控制器将会失效，系统的鲁棒性较差。一个可行的替代方案是将连通保持的能力整合到决策控制器之中。受到CTDE框架的启发，本小节通过设计一个在训练时起作用的集中式动作矫正器来将连通保持的能力整合到智能体的策略网络之中。

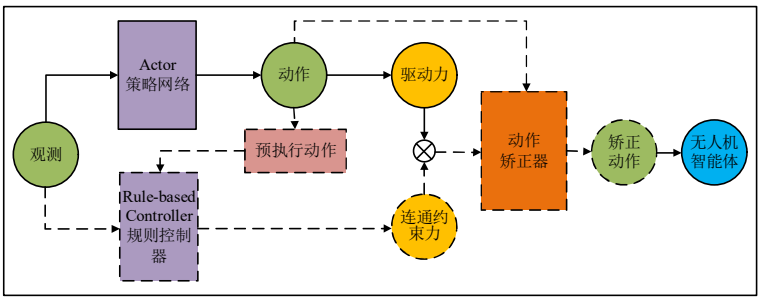
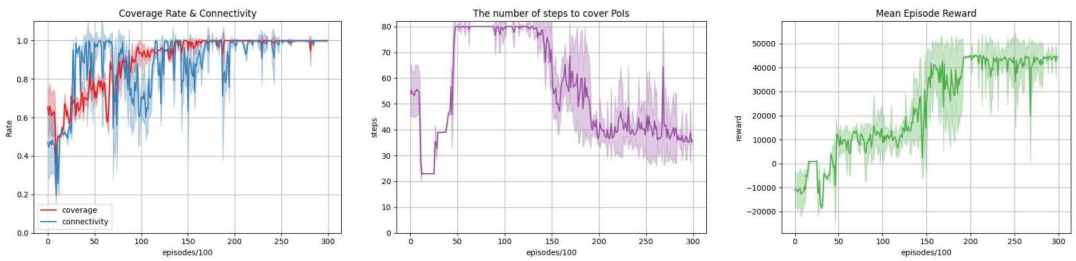


图 4-2 动作矫正器示意图

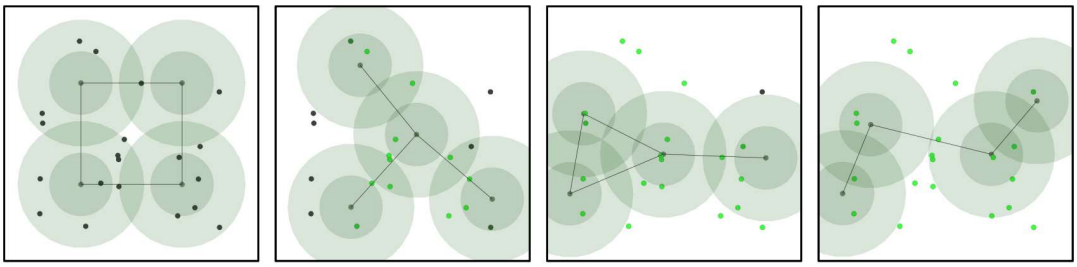
规则控制器将不再生成约束力直接作用于智能体，而是增加了一个动作矫正器。其输入为原始动作 a^0、策略驱动力 f_a 和规则约束力 f_r，输出为矫正后的动作 a^c。每当动作矫正器发生作用，将产生两个奖励不同的状态转换。未矫正的动作会导致失去连通的危险，从而获得负奖励；矫正后的动作继续保持了连通，奖励正常。动作矫正器同时生成正负样本，扩充了样本库，提高了算法的收敛性能。

将 MLP 隐层数量修改为 400，算法在 40k 幕上的收敛效果如图 4-3 所示，可视化效果如图 4-4 所示，集群覆盖轨迹如图 4-5 所示。



(a) (b) (c)

图 4-3 动作矫正器算法收敛效果 (a) 覆盖率与连通率；(b)每幕步数；(c)任务奖励



(a) (b) (c) (d)

图 4-4 动作矫正器算法可视化 (a) $t=0$；(b) $t=14$；(c)$t=28$；(d) $t=37$

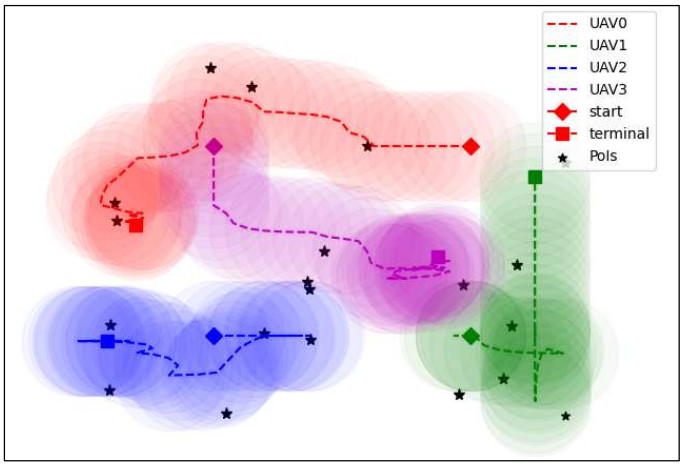


图 4-5 集群覆盖轨迹

第五章 全文总结与展望(500)

5.1 研究工作总结

纵览全文，本文主要工作如下：

1) 建立了基于覆盖能量和覆盖功率的无人机集群动态点覆盖任务模型；
2) 基于上述模型，使用 MADDPG 方法解决了无连通约束下的轨迹规划；
3) 基于 CTDE 框架提出了动作矫正器的连通保持算法；

5.2 未来工作展望

本文提出的方法已经能够解决不考虑连通保持和考虑连通保持两种情况下对无人机集群的轨迹规划，但仍有一些工作可以继续进行：

1) 无人机集群，包括动力学模型、覆盖范围等参数可以拓展至异质；
2) 可以考虑使用视觉信息替代数据式的观测，依此做到可变数量 PoIs；
3) 对于部分可观测问题，可以使用 RNN 等提高对轨迹的信息提取；

参考文献

[1] 谷旭平,唐大全,唐管政. 无人机集群关键技术研究综述[J]. 自动化与仪器仪表,2021(4):21-26,30.

[2] Gupta N, Kumar N, Jain S. Coverage Problem in Wireless Sensor Networks: A Survey[C]//2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). IEEE, 2016: 1742-1749.

[3] Cortes J, Martinez S, Karatas T, et al. Coverage Control for Mobile Sensing Networks[J]. IEEE Transactions on robotics and Automation, 2004, 20(2): 243-255.

[4] Hasanbeig M. Multi-Agent Learning in Coverage Control Games[D]. University of Toronto (Canada), 2016.

[5] 杨旭, 王锐, 张涛. 面向无人机集群路径规划的智能优化算法综述[J]. 控制理论与应用, 2020, 37(11): 2291-2302.

[6] Liu B, Dousse O, Nain P, et al. Dynamic Coverage of Mobile Sensor Networks[J]. IEEE Transactions on Parallel and Distributed systems, 2012, 24(2): 301-311.

[7] Bellman R. Dynamic Programming[J]. Science, 1966, 153(3731): 34-37.

[8] Sutton R S, Barto A G. Reinforcement Learning: An Introduction[M]. MIT press, 2018.

[9] Sutton R S, McAllester D, Singh S, et al. Policy Gradient Methods for Reinforcement Learning with Function Approximation[J]. Advances in Neural Information Processing Systems, 1999, 12.

[10] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous Control with Deep Reinforcement Learning[C]//ICLR (Poster). 2016.

[11] Horn R A, Johnson C R. Matrix analysis[M]. Cambridge university press, 2012.

[12] 李晓宇. 无线传感器网络覆盖控制优化策略研究[D]. 内蒙古:内蒙古科技大学,2018. DOI:10.7666/d.D01523036.

[13] Lowe R, Wu Y I, Tamar A, et al. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments[J]. Advances in Neural Information Processing Systems, 2017, 30.