群智感知场景下基于深度强化学习的多无人机路径规划算法

刘鑫彬[1]，高慧[2]

[1] 北京邮电大学计算机学院，北京 100876

[2] 北京邮电大学计算机学院，北京 100876

摘要： 群智感知活动通常对需要感知的兴趣点要求较高的感知覆盖率，作为群智感知活动的重要参与者之一，无人机可以有效感知人类参与者无法感知和其他无人机还未感知的兴趣点。本文面向群智感知下的多无人机数据协同收集问题，基于深度强化学习理论模型提出了 GMRL 算法，并借助于大量模拟实验和小规模的真实实验证实本文所提供的算法的效率。

关键词： 深度强化学习；多智能体；群智感知；路径规划

中图分类号： 请查阅《中国图书馆分类法》

A Deep Reinforcement Learning Algorithm for UAVs Route Planning in Mobile Crowd Sensing

Liu Xin-bin[1], Gao Hui[2]

[1] Beijing University of Posts and Telecommunications, Beijing 100876

[2] Beijing University of Posts and Telecommunications, Beijing 100876

Abstract: Mobile crowd sensing activities usually require a high coverage of the points of interest to be sensed. As one of the important participants, UAVs can effectively sense points of interest that are not sensed by human participants and not yet sensed by other UAVs. This paper proposes a GMRL algorithm based on deep reinforcement learning to solve the cooperative data collection problem of multiple UAVs. The efficiency of this algorithm is confirmed with the help of a large number of simulations and small-scale real experiments.

Key words: Deep reinforcement learning; Multi-agent system; Mobile crowd sensing; Route planning

0 引言

群智感知是指通过利用广大群众的参与，从大规模的、分散的、异质的信息来源中汇集、加工、分析和处理信息的过程。这种信息获取方式不依赖于专业人员，而是依靠众包、志愿者和

基金项目：国家青年基金 (62002025)；河北省重点研发计划项目：围绕低碳炼钢场景的工业物联网关键技术研究及平台应用示范 (21310102B)

作者简介：刘鑫彬（1999-），男，硕士研究生，主要研究方向：物联网和群智感知。通信作者：高慧（1986-），男，副教授，主要研究方向：物联网和群智感知。E-mail: gaohui786@bupt.edu.cn

互联网等技术手段。在群智感知活动中，无人机可以在较短的时间内覆盖更大的区域，并获得更多的数据。在人类参与者无法到达或不愿意进入的区域中，无人机可以收集更准确和全面的数据。不仅如此，无人机可以搭配更先进的传感器和设备，例如高解析度相机和精度更高的气象仪器等，协助获得更精准和准确的数据。因此，无人机可以通过范围、质量和效率方面的优势，支持和改善群智感知的质量和效率，成为感知任务的重要辅助手段。

在本文，我们将无人机的任务感知过程描述为一个马卡洛夫决策过程，并通过强化学习的方式实现我们所提出的多无人机协同的路径规划算法。通过该算法，无人机之间可以协同高效地执行感知任务，从而提高群智感知任务的感知覆盖率。同时，通过大量模拟实验和小规模真实的实验，我们评估了该方法的可行性和高效性。我们将本文的主要贡献总结如下：

1. 本文提出了一个多无人机协同算法，以满足群智感知任务中的感知覆盖率要求。
2. 本文基于真实城市背景搭建仿真环境，并进行小规模的真实试验验证所提算法的可行性。

接下来，我们将在文章的第一部分讨论相关的研究工作；在第二部分描述群智感知任务的系统模型；在第三部分介绍无人机的轨迹调度算法；在第四部分呈现了模拟实验和真实实验的结果；并在第五部分总结全文。

1 相关工作

无人机上通常配备有许多传感器，能够完成对准确性和时效性要求更高的，更具挑战的感知任务。当下，无人机已经被广泛地投入到群智感知的场景中，用以执行城市的感知任务，例如监测交通状况 [1]、空气污染 [2] 和停车情况 [3]。

然而，无人机在执行感知任务时也存在明显的限制，例如有限的电池能量供应 [4]。因此，Dai 等人旨在最大化收集到的数据量并最小化无人机的能量消耗。他们提出了一个以分散深度强化学习为基础的框架，促进无人机节能高效地执行感知任务 [5]。为了有效利用无人机收集感知数据，Wei 提出了一种基于深度强化学习的路径规划方法 [6]。Wang 等人考虑使用卡车和无人机共同完成感知任务，在此过程中，卡车作为移动无人机的枢纽，对无人机进行电池更换并收集归纳感知数据 [7]。Xie 等人设计了一种无人机声誉激励方案，并基于该激励方案筛选具有高声誉值的无人机 [8]。

2 系统模型

在群智感知任务场景中，一块感知区域里分布着 N 个兴趣点 $\mathcal{P}=\{p|1,2,...,N\}$，$M$ 台无人机 $\mathcal{U}=\{u|1,2,...,M\}$ 需要在该区域内飞行 T 个时隙 $\mathcal{T}=\{t|1,2,...,T\}$，并在相应的兴趣点附近停留以执行感知任务。每台无人机 u_j 均配置一颗容量为 E 的电池，并在每个时隙 t 内消耗电量 e_j^t 以执行一次飞行动作 $a_j^t=(x_j^t,y_j^t)_{j\in\mathcal{U},t\in\mathcal{T}}$，其中，$x_j^t$ 表示该无人机横向移动的距离，而 y_j^t 表示该无人机在纵向移动的距离。我们假设在感知任务开始时，每台无人机都已充满电；同时，假设每个时隙足够长，所有无人机均能在一个时隙内顺利完成其飞行和感知动作。

每台无人机 u_j 在飞行周期内对兴趣点的感知情况可以用集合 $\mathcal{L}_j = \{d_{i,j}^t\}$ 表示，$d_{i,j}^t = 1$ 意味着无人机 u_j 在时隙 t 内感知了位于集合 $\mathcal{P}$ 内的兴趣点 p_i 的数据；若 $d_{i,j}^t = 0$，则意味着无人机在时隙 t 内未能感知兴趣点 p_i 的数据。

本篇文章的目标在于：在多台无人机受到电量和时间限制的情况下，为无人机群规划合理的感知路径，从而尽可能提高兴趣点的感知覆盖率，即：

$$\begin{aligned} \text{maximize: } & \frac{\sum_{t=1}^{T}\left|\bigcup_{j\in\{1,2,\ldots,M\}}\mathcal{L}_j\right|}{N*T} \\ \text{subject to: } & \sum_{t=1}^{T} e_j^t \le E \end{aligned} \tag{1}$$

其中，$\sum_{t=1}^{T}\left|\bigcup_{j\in\{1,2,\ldots,M\}}\mathcal{L}_j\right|$ 表明了 J 台无人机在 T 个时隙内总共感知的兴趣点个数。

3 预测方法

我们使用一个引入贪心思想的独创强化学习算法来解决无人机群的路径规划问题。与大多数强化学习的场景相似，每台无人机都需要在该环境下活动，基于其自身对环境的观测值来决策飞行动作，从而学会避开环境中的障碍物，并与其他无人机一起协同感知环境中的兴趣点。

通常，在多智能体的强化学习环境中，每个智能体都需要充分了解自己所处的环境，即观测到环境中所有目标物和其他智能体的所有信息。然而，在实际的群智感知环境中，无人机的数量会远远少于所需要感知的兴趣点个数，因此，引入贪心思想可以有效地降低无人机的观测空间维度。此外，贪心算法可以帮助无人机避免大量重复和低效的探索，并迅速定位到与自身最近的兴趣点。因此，我们将贪心思路与深度 Q 神经网络（DQN）进行结合，从而为无人机群进行路径规划。

我们将无人机群的路径规划问题描述为一个动态的马卡洛夫决策过程，以下是该马卡洛夫决策过程涉及到的强化学习概念和具体的解释：

1. **观测空间**：群智感知环境下的观测空间 $\mathcal{S}$ 包含以下三个部分：一台无人机自身的绝对位置；该无人机与最贴近自身且尚未被感知的兴趣点的相对位置；该无人机与该兴趣点最近的另一无人机的相对位置。在每一个时隙 t 开始时，每台无人机 u_j 都会观测环境，从而得到来自状态空间 $\mathcal{S}$ 的观测值 s_j^t。

2. **行为空间**：群智感知环境的行为空间 $\mathcal{A}$ 包含环境中每台无人机 u_j 在每个时隙 t 内可以执行的飞行动作 a_j^t，该动作使得无人机在其东、西、南、北、东北、西北、东南或西南方向方向做出一定量的位移。

3. **状态转移函数与概率空间**：群智感知环境的状态转移函数可以表示为 $F : \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$，该函数表明，当无人机 u_j 处于状态空间下的某个状态 s_j^t，并做出一个动作 a_j^t 时，转移到状态 s_j^{t+1} 的概率，该概率位于概率空间 $\mathcal{V} = \{s^{t+1}|s^t, a^t\}_{s\in\mathcal{S},t\in\mathcal{T}}$ 中。

4. **奖励函数**：每当一台无人机 u_j 通过在状态 s_j^t 下执行行为 a_j^t 进入状态 s_j^{t+1} 时，都会立即获得来自群智感知环境的奖励 r_j^t，该奖励函数可以由 $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 表示。好的奖励函数可以促进无人机更快的学习到如何规划正确的路径，我们定义环境对一台无人机 u_j 在状态 s_j^t 下执行行为 a_j^t 所获的奖励计算方式如下：

$$r(s_j^t, a_j^t) = c(s_j^t, a_j^t) - l(s_j^t, a_j^t), \tag{2}$$

其中，$c(s_j^t, a_j^t)$ 表示环境根据无人机 u_j 在状态 s_j^t 下执行行为 a_j^t 所感知到的兴趣点总量为其计算的奖励，当无人机感知到的兴趣点越多，则该奖励值越大；$l(s_j^t, a_j^t)$ 则表示无人机 u_j 在状态 s_j^t 下执行行为 a_j^t 所带来的惩罚：当无人机撞到某个障碍物或飞出感知区域时，环境将给予其较重的惩罚，当无人机毫无目的地漫游时，环境会根据其与自身最近一个兴趣点的相对位置，对其进行相对轻微的惩罚。

在转移函数和奖励函数给出之后，环境针对任一无人机 u_j 的优化问题可以表述如下：

$$Q^{u_j,*}(s_j^t, a_j^t) = \mathbb{E}\left[r(s_j^t, a_j^t) + \gamma \max_{a_j^{t+1}} Q^{u_j}(s_j^{t+1}, a_j^{t+1})\right], \tag{3}$$

针对该问题，我们寻求的优化策略表述如下：

$$\pi^{u_j,*}(s_j^t, a_j^t) = \arg\max_{a_j^t} \mathbb{E}\left[r(s_j^t, a_j^t) + \gamma \max_{a_j^{t+1}} Q^{u_j}(s_j^{t+1}, a_j^{t+1})\right], \tag{4}$$

其中，$\gamma \in (0, 1)$ 为衡量即时奖励和长远奖励重要性的折扣因子，当 γ 越接近 1，表示策略更倾向于获得更多的长远奖励，当 γ 越接近 0，则表示该策略更倾向于获得更多的即时奖励，接下来，我们将描述基于该优化策略所设计的强化学习算法 1：

1. 在算法的第 1 至 2 行，我们将初始化容量为 O 的经验缓冲区 B，并用随机权重初始化神经网络 Q。接下来，我们将重复 *episodes* 个训练周期以训练该神经网络。

2. 算法 3 至 16 行描述了每个训练周期内无人机的飞行过程。初始时，无人机的电量为 E 且状态为 s_begin。在每个时隙 t，无人机都会根据概率 ϵ 和 β 的大小，从而在贪婪、随机和神经网络三者之一中决策行为 a_t 并执行。

3. 算法的 17 至 23 行描述了每个训练周期内神经网络的训练过程。每次无人机做出行为后，我们都将 (s_t, a_t, r_t, s_{t+1}) 作为一条经验存入经验缓冲区内。每次训练时，我们从缓冲区抽取 b 条记录，并根据损失函数数 $(y_t - Q(s_t, a_t))^2$ 训练 Q 神经网络。

4 仿真实验

我们使用 OpenCV-Python 搭建一个类 OpenAI-Gym 风格的仿真强化学习环境来验证我们所提出的 GMRL 算法效果，实验中环境所需参数见表 1。接下来，我们将进一步介绍如何搭建我们的仿真环境：

Algorithm 1 GMRL：无人机路径规划算法

输入: 初始状态 s_{begin}, 初始电量 E, 最大时隙 T

输出: 深度 Q 网络模型 $Q(s,a)$

初始化容量为 O 的经验缓冲区 B，以及动作-价值神经网络 Q
for 训练 *episodes* 次 **do**
  初始化 $t=0, s_t=s_{begin}, e=E$
  while 当前电量 $e>0$，且当前时隙 $t<T$ **do**
    生成一个随机概率 $\epsilon \in [0,1]$
    if $\epsilon < \epsilon_0$ **then**
      生成一个随机概率 $\beta \in [0,1]$
      if $\beta < \beta_0$ **then**
        通过贪婪算法选择行为 a_g，作为当前时隙的行为 $a_t = a_g$
      else
        随机选取行为 a_r，作为当前时隙的行为 $a_t = a_r$
      end if
    else
      通过网络模型决策出行为 $a_q = \max_a Q^*(s_t, a_t)$，并作为当前时隙的行为 $a_t = a_q$
    end if
    执行行为 a_t，获得奖励 r_t，扣减电量 e_t 并进入新状态 s_{t+1}
    将 (s_t, a_t, r_t, s_{t+1}) 作为一条经验保存到缓冲区 B
    if 当前电量 $e \le 0$，或当前时隙 $t=T$ **then**
      $y_t = r_t$;
    else
      $y_t = r_t + \gamma \max_{a^*}(s_{t+1}, a_{t+1})$;
    end if
    从缓冲区 B 中随机抽样 b 条经验，并根据损失函数 $(y_t - Q(s_t,a_t))^2$ 训练 Q 神经网络
    $t = t+1$
  end while
end for

表 1: 参数设定

参数	**值**
无人机个数	1 台至 5 台，默认为 4 台
无人机感知范围	12 米至 20 米，默认为 20 米
兴趣点个数	170 个至 270 个，默认为 270 个
时隙个数	100 个

1. 我们选取意大利罗马市位于经度范围 12.45450-12.46450，纬度范围：41.90500-41.91500 内约 1000×1000 平方米的城市街道作为仿真群智感知实验背景，并在该街道上每隔 50 米布置一个兴趣点。

2. 我们使用型号为 DJI Mavic 2 的大疆无人机来作为实验无人机的数据参考。理想情况下，该无人机最大飞行速度可达 $20m/s$，在满格电量下能够飞行约 $18000m$。

3. 仿真环境运行在 Ubuntu 18.04.3 X64 服务器系统上，该服务器配置有 4 核 3.60 GHz 的 Intel(R) Xeon(R) Gold 5122 英特尔处理器，62GB 内存和 2 块英伟达 Nvidia GeForce RTX 2080Ti 显卡。为训练神经网络，实验采取 Python 3.7.9 和 Pytorch 1.7.0 环境，并在该环境上基于 OpenCV-Python 独立搭建了一个类 OpenAI-Gym 强化学习环境风格的迷你学习环境。

图 1展示了数量分别为 2、3、4 和 5 台无人机的飞行轨迹，可以观察到所有无人机在飞行过程中都成功地避开了障碍物，并最终停留在感知区间内。同时，飞行轨迹直观地表明每台无人机都能与其他无人机协同合作，避免重复感知兴趣点。为了验证算法的表现程度，我们使用四个基准算法与本文提出的 GMRL 算法进行比较：

1. DRL-PP 算法：一个基于 PyBullet 实现的用于竞争和合作环境中的多智能体深度强化学习算法，该算法考虑了多个无人机的能量消耗，并拥有较高的兴趣点感知覆盖率。

2. Reward—Max 算法：一个基于即时奖励最大的贪婪算法，该算法以追求奖励值最大化为目标，促使无人机在每次做决策的时候都做即时奖励最大的选择。

3. Poi-Max 算法：一个基于兴趣点相对距离的贪婪算法，该算法以追求感知最近的兴趣点为目标，促使无人机在每次做决策的时候尽可能逼近最近的兴趣点。

4. Random 算法：一个随机算法，该算法下，无人机将随机决策自身的行为。

图 2展示了不同实验条件下对感知覆盖率的影响，如图所示，在不同的实验条件下，而 GMRL 算法在任何情况下的表现都优于其他基准算法。如，当感知距离的分别为 12 米和 20 米的时候，GMRL 算法采集了 165 和 181 个兴趣点，相对于采集 96 个兴趣点和 153 个兴趣点的 PoI-Max 算法而言，分别提高了约 25% 和 10% 的感知覆盖率；当无人机个数为 5 时，GMRL 相对于 DRL-PP 算法多采集了 23 个兴趣点，整体提高了约 9% 的感知覆盖率，而当无人机个数为 3 的时候，GMRL 的感知覆盖率相较于 DRL-PP 提高了 14%。

为了验证我们算法的真实可行性，我们在一块 8×8 平方米的场地进行了一次真实的无人机飞行实验，图 3中为我们实验所使用的无人机。在该实验中，我们使用 4 台无人机执行群智感知任务，红色的塑料桶和塑料椅代表无人机飞行过程中可能遇到的障碍物，矿泉水瓶表示无人机所需要感知的兴趣点。无人机通过 WiFi 连接到路由器，接收来自计算机通过 GMRL 算法计算出的飞行路径，并遵照该路径进行飞行。最终，所有兴趣点都被成功感知，每台无人机都

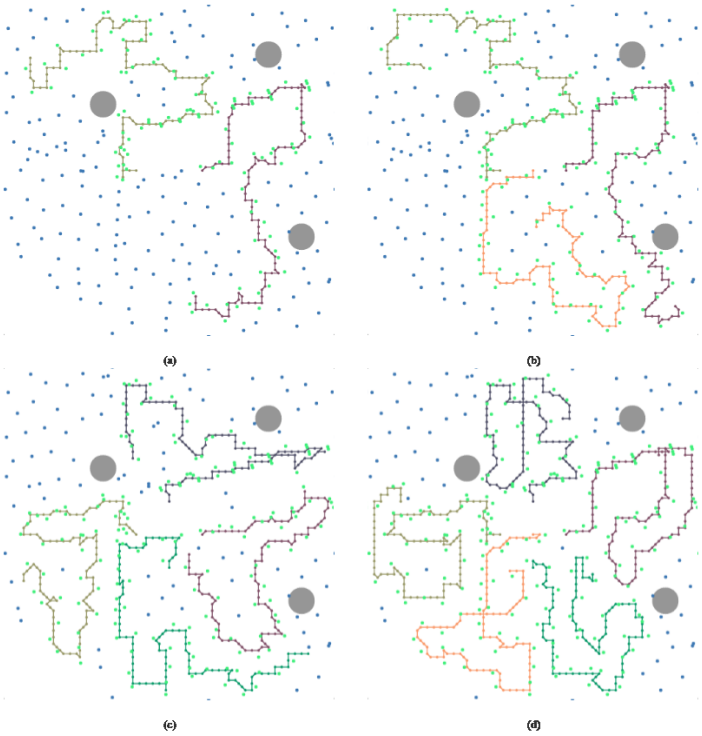

图 1: 不同个数的无人机的路径（线条表示无人机的路径，灰色方块表示障碍物，蓝色的点为还未采集的兴趣点，绿色的点表示已采集的兴趣点）。

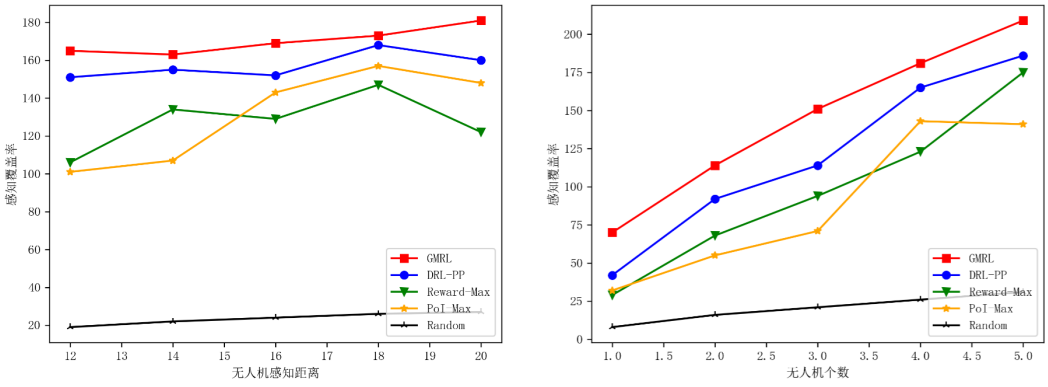

图 2: 不同个数的无人机和不同的感知距离对感知覆盖率的影响。

图 3: 用于执行该实验的四台 DJI RoboMaster TT 无人机。

成功避开了障碍物与其他无人机。我们拍摄了该实验的全部过程，完整视频可通过百度网盘[1]获取。

5 结论

为了最大化群智感知任务场景中兴趣点的感知覆盖率，本文提出了 GMRL 多无人机协同算法。通过该算法，无人机之间可以高效协同感知兴趣点，并有效避开场景中可能存在的障碍物。仿真实验和真实场景下的实验结果都证实了该算法的有效性和高效性。

参考文献（References）

[1] Huang H, Savkin A V, Huang C. Decentralized autonomous navigation of a UAV network for road traffic monitoring[J]. IEEE Transactions on Aerospace and Electronic Systems, 2021, 57(4): 2558-2564.

[2] Liu Y, Nie J, Li X, et al. Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms[J]. IEEE Internet of Things Journal, 2020, 8(12): 9827-9837.

[3] Gogoi P, Dutta J, Matam R, et al. An UAV assisted multi-sensor based smart parking system[C]//IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2020: 1225-1230.

[1]链接：https://pan.baidu.com/s/1r861pPmS8MoJEcfM2qoduQ?pwd=gmrl，提取码：gmrl

[4] Ding L, Zhao D, Cao M, et al. When crowdsourcing meets unmanned vehicles: Toward cost-effective collaborative urban sensing via deep reinforcement learning[J]. IEEE Internet of Things Journal, 2021, 8(15): 12150-12162.

[5] Dai Z, Liu C H, Han R, et al. Delay-sensitive energy-efficient uav crowdsensing by deep reinforcement learning[J]. IEEE Transactions on Mobile Computing, 2021.

[6] Wei K, Huang K, Wu Y, et al. High-performance UAV crowdsensing: A deep reinforcement learning approach[J]. IEEE Internet of Things Journal, 2022, 9(19): 18487-18499.

[7] Wang Z, Zhang B, Li C. Joint Path Planning of Truck and Drones for Mobile Crowdsensing: Model and Algorithm[C]//2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021: 1-6.

[8] Xie L, Su Z, Chen N, et al. Secure data sharing in UAV-assisted crowdsensing: Integration of blockchain and reputation incentive[C]//2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021: 1-6.