基于强化学习的机器人底盘能量管理与路径规划优化算法

李潇宇[1]，张君华[1]，郭晓光[1※]，伍 纲[2]
（1. 北京信息科技大学机电工程学院，北京 100192；2. 中国农业科学院农业环境与可持续发展研究所，北京 100081）

摘 要：为解决温室底盘传统路径规划中因忽略地面粗糙度而导致的电池寿命缩短与利用效率低下的问题，该研究探讨了 3 种融合电池能量管理与路径规划的强化学习算法。首先基于先验知识构建分级预打分奖励模型，并通过增加曼哈顿距离构建奖励函数，提高电池寿命和利用率；其次针对传统 Q-Learning(QL) 算法收敛效率低、易陷入局部最优等问题，提出了自适应变步长的优化算法（adaptive multi-step q-learning，AMQL）和基于自适应改变探索率的优化算法（adaptive ε-greedy q-learning，AEQL），以提升 Q-Learning 算法的性能。此外，为进一步提高算法的可行性，本文将 AMQL 算法和 AEQL 算法进行融合，提出了一种自适应多步长和变 ε-greedy 融合算法（adaptive multi-step and ε-greedy q-learning，AMEQL），并通过仿真对比的方式，验证了 AMQL 和 AMEQL 算法相对于传统 QL 算法在 3 个不同垄道下的性能。仿真试验结果表明：AMQL 相对于传统 QL 算法，训练平均时间降低 23.74%，收敛平均迭代次数降低 14.01%，路径平均拐点数降低 54.29%，收敛后的平均波动次数降低 18.01%；AMEQL 相对于传统 QL 算法，训练平均时间降低 34.46%，收敛平均迭代次数降低 23.68%，路径平均拐点数降低 63.13%，收敛后的平均波动次数减少 15.62%，在 400 次迭代过程中，AMEQL 到达最大奖励后平均每 7.12 次迭代波动 1 次，而 AMQL 平均每 6.68 次迭代波动 1 次。可知 AMEQL 训练时间最短，收敛最快，路径拐点数量最低，奖励波动最小，而 AMQL 次之。该算法可为温室底盘自主路径规划提供理论参考。

关键词：温室；路径规划；强化学习；能量管理；多目标优化
doi：10.11975/j.issn.1002-6819.202405192
中图分类号：S23-0 **文献标志码：**A **文章编号：**1002-6819(2024)-21-0175-09

李潇宇，张君华，郭晓光，等. 基于强化学习的机器人底盘能量管理与路径规划优化算法[J]. 农业工程学报，2024，40(21)：175-183. doi: 10.11975/j.issn.1002-6819.202405192 http://www.tcsae.org
LI Xiaoyu, ZHANG Junhua, GUO Xiaoguang, et al. Reinforcement learning-based optimization algorithm for energy management and path planning of robot chassis[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(21): 175-183. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202405192 http://www.tcsae.org

0 引 言

中国作为传统农业大国，温室大棚的出现有效解决了作物生长受外部条件限制的难题。为提升设施农业生产的效率，温室内通常种植密集，作业空间狭小，且耕、种、管、收全周期的工作任务繁重。此外，温室中的高温、潮湿和密闭环境对人体健康产生不利影响，不适宜人工长时间作业[1-4]。温室机器人的应用为解决这些问题提供了一种有效途径，然而温室内复杂且有限的空间要求机器人具备高效的路径规划能力，并能够在避免与障碍物碰撞的前提下完成任务。由于地面粗糙度变化，温室机器人的底盘（以下简称“机器人底盘”）可能会受到不同程度的垂直脉冲应力，这些应力会损坏电池的内部结构，如隔膜撕裂、电极结构受损等，进而缩短电池寿命。如果路径规划仅考虑避免大面积粗糙地面，可能会导致行驶路径过长，从而降低工作效率和能源利用率。因此，寻找兼顾能源高效利用与路径优化的规划方法，对于减少温室气体排放[5-6]和实现能源政策目标至关重要。

路径规划的研究通常分为 3 种主要方法：传统方法、深度学习方法和强化学习方法，其中传统全局路径规划算法在实际应用中非常广泛[7-11]。沈跃等[12]提出了一种基于相邻争夺（adjacent competition，AC）算法的植保无人机作业路径规划算法，该算法首先对粒子设置作业距离范围，其次在范围内对粒子作业距离初始值进行随机分配，最后通过相邻粒子争夺作业距离来搜索最优路径，此算法不仅解决了传统粒子群（particle swarm optimization，PSO）算法在规划植保作业路径时易陷入局部最优、搜索能耗最优方案能力低等问题，同时也保证了作业总距离一定、对搜索方向进行先验且不遗漏特殊点，提高了植保无人机作业效率、减少了无人机损耗。孙月平等[13]提出了一种基于改进 A*算法与人工势场法相融合的蟹塘投饵船动态路径规划算法（fusion of improved A* and artificial potential field，FIA*-APF），通过引入动态加权因子优化 A*算法评价函数、加入转折惩罚函数等方法，在静态和动态仿真环境及蟹塘试验中，FIA-APF 算法在规划时间、指令节点数量、路径长度及

收稿日期：2024-05-27 修订日期：2024-09-13
基金项目：国家自然科学基金（12272057）
作者简介：李潇宇，研究方向为强化学习和路径规划。
Email：xiao@bistu.edu.cn
※通信作者：郭晓光，博士，讲师，研究方向为多智能体协同控制。
Email：solarguo@bistu.edu.cn

转角等方面表现出色；然而，该算法的缺点在于计算量大、灵活性不足、难以适应复杂动态环境变化，并且对先验知识依赖程度高，这使得在复杂温室环境中构建先验知识模型变得困难，导致无法实现最优的路径规划。

深度学习可以用来对路径规划所涉及的环境进行建模和理解，通过对大量数据进行特征提取和数据分析，从而提高路径规划的效率和结果。DANG 等[14-15] 提出了一种基于单目摄像机的移动机器人实时避障策略，该方法使用二值语义分割单目摄像机捕获的图像来提取特征，并估计机器人环境中障碍物的位置和距离。然后，基于增强 A*算法的优化路径规划，结合碰撞、路径、平滑代价等加权因子，来提高移动机器人的路径规划性能。孙国祥等[16] 为解决传统温室导航方案的路径规划问题，提出了基于即时定位与地图构建技术的激光视觉融合式自主导航算法，利用多种设备和算法实现局部与全局定位及自主导航，满足了温室高精度建图、定位和导航需求。但是深度学习在规划效果上仍存在一些不足，其根本原因是路径规划方面存在计算量大、硬件需求高、模型设计复杂、可解释性差、对数据依赖强且易受数据质量影响等缺点，并且在温室环境中还可能面临实时性不足、难以适应动态变化环境以及训练和调试难度大等问题。

强化学习是一种通过与环境交互进行自我更新的方法[17-19]，能适应复杂动态环境，注重长期回报，并具有良好的可扩展性和通用性。LAZZARONI 等[20] 探讨了一种基于深度强化学习的智能体，能够同时执行路径规划和轨迹执行，处理传感器感知信息，并像普通驾驶员一样直接控制方向盘和加速器。他们将研究对象限制在低速运动和狭窄的可行驶区域内，该智能体完全依赖传感器所捕获的实时信息作为输入参数来取消对地图的依赖。结果显示智能体在目标到达率和换挡次数方面均优于传统的混合 A*路径规划算法。但该研究仅对规划效果进行了优化，未充分考虑到地面粗糙度变化对机器人电池寿命的影响以及能源效率的需求，同时，DRL 是基于传统 QL 的贪心算法进行单步长梯度优化，易产生局部最优、收敛效率低以及训练慢等问题。

针对以上问题，本文基于强化学习对温室底盘的路径规划算法进行研究。一方面，温室环境复杂且空间有限，强化学习能够让温室底盘通过与环境不断交互自主完成路径规划的工作，并在避免与障碍物碰撞的同时提高路径规划的效率。另一方面，考虑到地面粗糙度变化对机器人底盘电池寿命以及电池利用效率的影响，强化学习可以通过定义合适的奖励函数，确保温室底盘在规划路径时会权衡因行驶路线过长导致工作效率、能源利用率的降低和因在大粗糙地面行驶导致电池寿命缩短之间的关系，并确定出最优路线。同时，针对传统 Q-Learning(QL) 的单步长梯度优化导致易陷入局部最优、收敛效率低以及训练慢等问题，本文提出并探讨了 AMQL、AEQL、AMEQL 算法的优点和适用性，以期为温室底盘自主路径规划提供理论参考。

1 传统 QL 算法

传统 QL 是一种基于价值迭代的强化学习算法，用于解决马尔可夫决策过程（markov decision process，MDP）中的最优控制问题[21-22]。与基于模型的强化学习方法不同，Q-Learning 不需要事先了解环境的完整动态模型，就能通过与环境交互的方式进行学习。同时，Q-Learning 具有一定的鲁棒性，即使目标位置存在偏差，所学习的策略仍能在一定程度上保证任务的顺利完成。该算法通过训练智能体在特定环境中选择最优行动，从而最大化累积预期奖励。作为强化学习中的一种基于价值的算法（value-based），Q-Learning 通过计算 Q 值，即在状态 s（$s \in S$）下采取动作 a（$a \in A$）时的预期收益，来指导智能体的决策。该算法的核心思想是通过计算 Q 值来构建 Q 表存储各状态-动作（S-A）对 Q 的值，并依据 Q 值选择最优动作，如表 1 所示。

表 1 部分存储 Q 值的 Q 表

Table 1 Partial Q-table for storing Q-values.

状态 Status	A1	A2
S1	Q（$s1$,$a1$）	Q（$s1$,$a2$）
S2	Q（$s2$,$a1$）	Q（$s2$,$a2$）
S3	Q（$s3$,$a1$）	Q（$s3$,$a2$）

注: Q（$s1$,$a1$）表示在状态 S1 下做动作 A1 的 Q 值，余同。

Note: Q($s1$, $a1$) represents the Q value of taking action A1 in state S1， the following as the same.

算法的核心计算式-贝尔曼方程如式（1）所示：

$$Q(s,a) \leftarrow Q(s,a) + \alpha[\gamma \cdot \max_{a'} Q(s',a') - Q(s,a)] \quad (1)$$

式中s是当前状态，a是当前动作，s'是转移到的下一个状态，a'是转移状态之后执行的动作，γ是学习率（控制 Q 值更新的步长），α是折扣因子（用于平衡当前奖励和未来奖励的重要性）。该算法利用时间差分方法，将蒙特卡罗采样与动态规划中的自举技术相结合，使其成为一种无需环境模型的算法。通过自举机制，算法利用后继状态的值函数来估计当前状态的值函数，实现单步更新。这种方法不仅避免了对完整环境模型的依赖，还显著加快了学习速度。通过更新 Q 值来优化策略，使得机器人在不断探索环境的同时最大化累积奖励，图 1 为算法流程图。

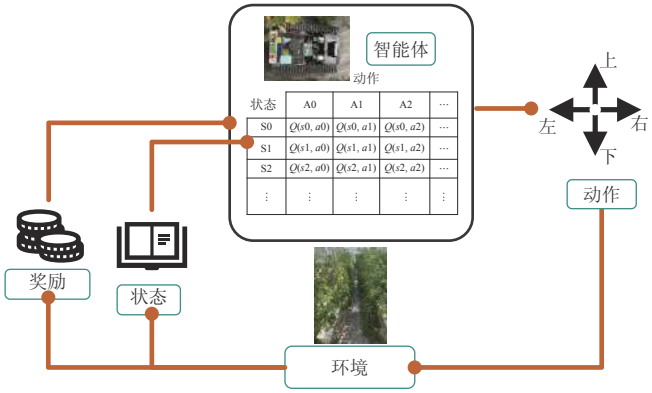


图 1 传统 QL 算法流程图

Fig.1 Flow chart of the traditional QL algorithm

2　强化学习能量管理策略

通过对实际场景的考察发现，不同垄道中泥土和沙砾导致地面粗糙度存在明显差异，如表 2 所示，不同的地面粗糙度对电池的工作效率和寿命产生了显著影响[25-28]。鉴于温室地面的复杂性，本文基于强化学习的能量管理进行路径规划。在此过程中，首先需要学习先验的专家知识，作为强化学习奖励设置的基础。该先验知识主要包括两部分：电池在不同地面粗糙度下的寿命以及利用效率的使用情况。温室底盘实车坐标系如图 2 所示。

表 2　地面粗糙度

Table 2　Ground roughness

地面覆盖物 Ground cover	粗糙度 Roughness（Z_0）/m	预打分奖励 Pre-scoring reward		颜色表示 Color indication
		电池寿命 Battery life	电池利用率 Battery utilization efficiency	
城郊房舍区（Level 1）	0.60	0.167	1.67	绿色
牧场（Level 2）	0.20	0.5	5	黄色
谷草地（Level 3）	0.10	1	10	紫色
长草地、石头摊（Level 4）	0.050	2	20	蓝色

注：通过对参考文献 [23-24] 提取 level1～level4 四个不同粗糙度地面以实现分级预打分。

Note: Four different roughness levels of ground, namely level1 to level4, were extracted from references [23-24] to achieve graded pre-scoring.

图 2　车辆坐标系

Fig.2　Vehicle coordinate system

2.1　电池寿命与电池利用效率先验知识

1）电池寿命

SOMERVILLE 等[29] 采用在振动测试期间对电池进行 X 射线光电子能谱分析的方法，研究了锂离子电池因振动引起性能退化的机电机制。结果表明，振动导致电池表面选择性形成的膜被去除，并由电解质分解产生的新表面膜所取代。这种由振动引发的膜层变化是电池性能退化的主要原因。

2）电池利用效率

ZHANG 等[30] 研究了振动对锂离子电池的直流电阻、1C 容量以及一致性的影响。用统计学方法研究了锂离子电池试验数据，通过对 32 个 18 650 电池单轴振动获取数据，借由大样本量剖析了测试前后电池直流电阻和容量的方法变化。结果表明，在 95% 的置信水平下，振动致使直流电阻显著上升，同时也能观察到 1C 容量有所降低。

振动导致电池内阻增加和容量减小，会对电池利用效率产生多方面的明显影响。首先，内阻增加会导致电池在放电过程中更多的电能被转化为热能而散失掉，即能够实际用于驱动设备或完成其他工作的电能减少，电池利用效率降低。其次，电池可提供的总电荷量减少，原本能够支持机器人底盘运行较长时间的电池，由于容量降低，其使用时间会大幅缩短。最后，内阻增加和容量减小还可能导致电池的输出功率不稳定。在温室环境中，这可能会引起机器人底盘性能的下降，甚至出现故障。

2.2　先验知识分级预打分机制

通过电池寿命与电池利用效率先验知识可知，地面粗糙度越大，对电池寿命以及电池利用效率的影响越明显。因此，本文通过学习相关先验知识，为不同粗糙度地面进行分级预打分，其打分机制如式（2）～（3）所示。由于车载惯性测量单元（inertial measurement unit，IMU）可以实时获取地面粗糙度数据，分级预打分可以为不同地面粗糙度数据提供相应的先验奖励，因此机器人底盘可借助 IMU 实时获取这些奖励数据，在后续的强化学习中，这些奖励将有助于寻找最优路径。

$$e_{(i)} = 1/Z_o \tag{2}$$

$$l_{(i)} = 1/Z_o \cdot H \tag{3}$$

式中$e_{(i)}$与$l_{(i)}$为不同粗糙度地面分级预打分奖励，即粗糙度越大奖励值越低。$l_{(i)}$为电池寿命在不同地面粗糙度下的分级预打分，$e_{(i)}$为电池利用效率在不同地面的分级预打分，Z_o为粗糙度，H为权重因子，由于粗糙度对电池寿命的影响更加明显[31]，本文设置H为 0.1 以减少对奖励的影响。

2.3　外部环境及底盘奖励函数模型

为了实现机器人底盘在作业时的能量管理控制，需要根据不同地面粗糙度和底盘状态进行合理的路径规划，进而使机器人底盘能够在避开高粗糙度地面的同时，提高电池的有效利用率。由于强化学习在训练的初始阶段无法直接做出正确的决策，而是通过不断试错来获取奖励值并进行优化，最终形成有效的训练模型，以确保机器人底盘的行为符合多系统动态协调控制的要求。然而，逐一将不同动作作用于车辆动力学模型并计算对应的奖励值，该过程过于繁琐，且车辆动力学模型中许多部分与奖励值计算并无直接关联。因此，需建立一个能够准确模拟机器人底盘在接收到上层控制指令后不同动作表现的环境模型，作为强化学习算法的训练环境。在实际验证中，将整车的状态作为输入，动作作为输出，具体步骤如下：

首先，设计奖励函数：1）根据机器人底盘在不同地面粗糙度环境下的电池利用效率和寿命表现，设置对应

奖励数值，鼓励机器人底盘选择更加高效的电池使用方式；2）根据机器人底盘的路径规划表现，设置对应距离的奖励函数，鼓励其选择更加高效的路径。

其次，基于以上两种补充情况，在底盘执行动作后评估其行为并进行反馈，将相应的奖励分配给机器人底盘。对于正向行为，通过给予正向奖励的方式增加这些行为的发生频率；对于负向的行为，通过给予负向奖励的方式减少这些行为的发生频率。

最后，通过机器人底盘与环境的持续交互、学习和调整，使其能够根据奖励反馈逐步优化行为策略，从而实现更高效的能量管理控制，并提升电池的使用效率和寿命。

本文将奖励函数$R_{(i)}$设计为基于不同地面粗糙度下的电池使用效率、电池寿命以及运行路径长度这 3 个因素有关的函数，具体如式 4 所示。

$$R_{(i)} = \omega \cdot e_{(i)} + \beta \cdot l_{(i)} + \mu \cdot d \tag{4}$$

其中为运行路径（用曼哈顿距离表示），ω、β、μ为权重因子，通过设置权重来优化温室底盘在规划路径时避免大粗糙度地面以提高电池寿命、增加电池利用效率，同时缩短行驶路程以减少工作时间等多个目标。

3 Q-Learning 算法优化

3.1 改进的自适应多步长 QL 算法（AMQL）

传统 QL 算法在大型状态空间中的收敛速度通常较慢，尤其当状态空间包含大量随机性因素或者存在长期延迟现象时。为解决这一问题，本文构建了 AMQL 的优化方法。AMQL 是指在 QL 算法中采用自适应的步长，以提高算法的效率和收敛性，并且大幅减少拐点数量。传统的 QL 算法使用固定单步长向周围探索，但在实际应用中，步长的选择可能会对算法的性能产生重要影响。自适应多步长的 QL 算法可以根据学习过程中的反馈动态地调整步长，以更好地适应环境的变化和学习的进展，如图 3 所示。

AMQL 算法流程：

1）初始化动作 a 与状态 s。

2）判断是否到达最大迭代次数。

3）若达到最大迭代次数，则输出训练完的最优路径；若未到达最大迭代次数的同时，未达到终止条件（最大步数），则继续根据当前的策略选择动作 a。

4）执行上一步选择的动作 a，观察环境反馈（奖励 r 和转移到的下一个状态 s'）。

5）使用式（1）更新 Q 值。

6）判断环境所给予的奖励是否为正，如果在连续两个步长中都获得正向奖励，则认为当前搜索为正确路径，并且计算当前位置与目标点的曼哈顿距离。如果满足连续正向奖励且曼哈顿距离大于 10（避免在探索终点时欠拟合），在下次探索时将步长增加为 2。若持续获得正向奖励，则步长增加，最大可至 5；如果满足连续正向奖励且曼哈顿距离小于 10，则步长会根据距离的缩小自适应减少；若不满足上述条件，重置步长为 1，如图 4 所示。

7）循环执行步骤 2）至步骤 5），直到达到指定的停止条件。

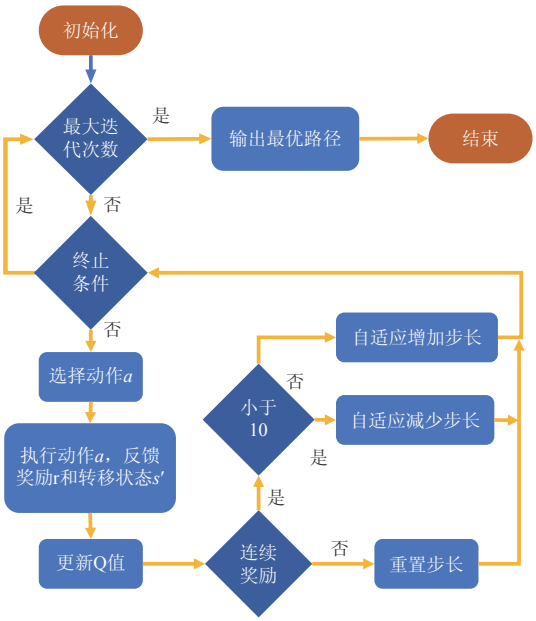
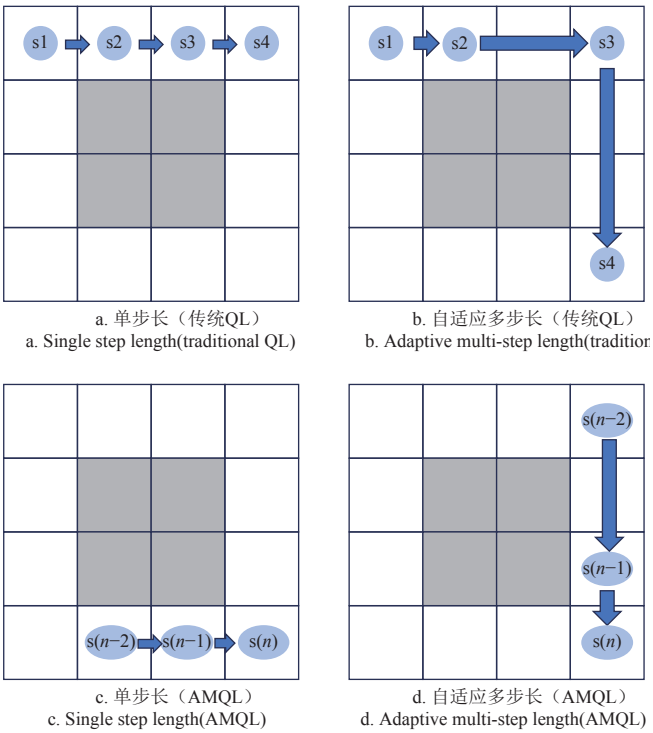


图 3 AMQL 算法流程图

Fig.3 Flow chart of the AMQL algorithm



a. 单步长（传统QL） b. 自适应多步长（传统QL）

a. Single step length(traditional QL) b. Adaptive multi-step length(traditior

c. 单步长（AMQL） d. 自适应多步长（AMQL）

c. Single step length(AMQL) d. Adaptive multi-step length(AMQL)

注：a、c 为传统 QL；b、d 为 AMQL；图中 n 为时间步数。

Note: a and c are traditional QL; b and d are AMQL; In the figure, n is the number of time steps .

图 4 传统 QL 单步长与 AMQL 对比

Fig.4 Comparison of traditional QL single step length and AMQL

3.2 改进的自适应 ε-greedy 的 QL 算法（AEQL）

ε-greedy 是强化学习中常用的一种策略，用于在探索与利用之间进行权衡，特别是在 Q-learning 等基于值的方法中常被使用。

传统 ε -greedy 策略是在每次选择动作时，以ε的概率随机选择一个动作（探索），以$1-\varepsilon$的概率选择当前估计最优的动作（利用），ε在每一轮迭代完之后会将以线性化的方式进行递减。这种贪心策略的缺点为：虽然每个动作都有被选择的概率，但选择过程过于随机，某些状态-动作对本应达到全局最优，但由于初始化的影响，其被访问的概率较低，容易陷入局部最优。此外，迭代过程中，若ε过大会导致过度探索；若ε过小则会导致难收敛。

为解决上面问题，本文提出一种自适应ε策略。通过式（5）~（8）来实现对ε的动态调整。

$$v = c_R - p_R \tag{5}$$

$$N = \frac{v_i - v_{\min}}{v_{\max} - v_{\min}} \tag{6}$$

$$\varepsilon \leftarrow \varepsilon/(1+f\cdot N) \tag{7}$$

$$\varepsilon \leftarrow \varepsilon\cdot(1+f\cdot N) \tag{8}$$

式中v表示奖励值变化率，c_R表示当前的奖励值，p_R表示上一次奖励值，N为归一化结果，v_i为当前的奖励值变化率，$v_{\min}$为最小奖励值变化率，$v_{\max}$为最大奖励值变化率，ε为探索度，f为权重。式（5）表示当前状态所获得的奖励值与上一个状态获得的奖励值之间的变化率。式（6）表示对当前获取的奖励值变化率进行归一化处理。式（7）表示当v为非负值时ε的迭代公式，v越大ε减少的幅度越大，v越小ε减少的幅度越小。式（8）表示当v为负值时ε的迭代公式，v越大ε增大的幅度越大，v越小ε增大的幅度越小。通过将ε线性化递减的方式改变为依据奖励值的变化率来自适应实现对ε的调整，更符合$\varepsilon-greedy$算法的策略初衷，如图 5 所示。

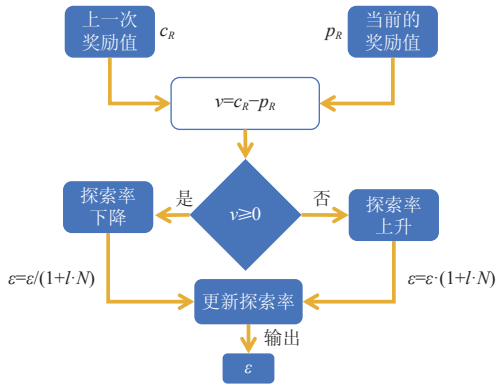


图 5　自适应 ε-greedy 的 Q-Learning 优化流程图

Fig.5　Flow chart of Q-Learning optimization with adaptive ε-greedy

3.3　算法融合（AMEQL）

AMQL 算法与 AEQL 算法各有其优点，但也都存在明显的缺陷：虽然 AMQL 提高了收敛效率和迭代速度，但由于步长变化较大，导致奖励波动幅度很高，算法稳定性较低；同时，多步长的引入对收敛效率和迭代速度的提升效果并不十分明显；另一方面，AEQL 通过动态调整ε可以提高探索效率，增强算法稳定性，但由于ε的在训练初期波动较大，增加了训练时间。为了解决这些问题，本文将 AMQL 与 AEQL 进行融合，提出融合算法 AMEQL，此算法不仅解决了上述问题，而且结合了各自的优点，能够在全局路径规划中快速选择出更优的路径。其融合算法如图 6 所示。

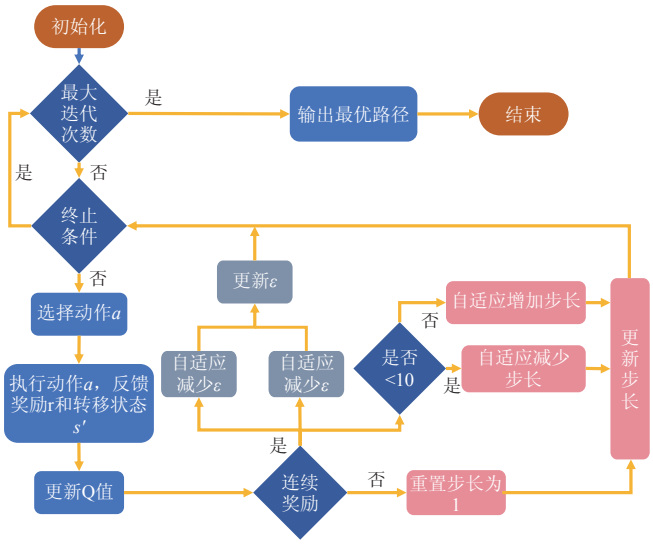


图 6　融合算法流程图

Fig.6　Flow chart of the fusion algorithm

4　仿真环境搭建

为了测试传统 QL 算法与优化 QL 算法的路径规划效果，本文在仿真试验中基于实际番茄温室农业场景进行了仿真建模，图 7 为农业环境实际取景，为仿真环境的搭建作参考。图 8 中的场景模拟采用 70 m×50 m 的栅格表示，包含 3 条垄道（种植区）与 4 条过道（工作区），其中垄道用黑色栅格表示，过道的数据通过 IMU 测量获得 4 种不同粗糙度的地面信息[32]。

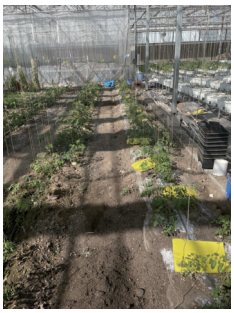
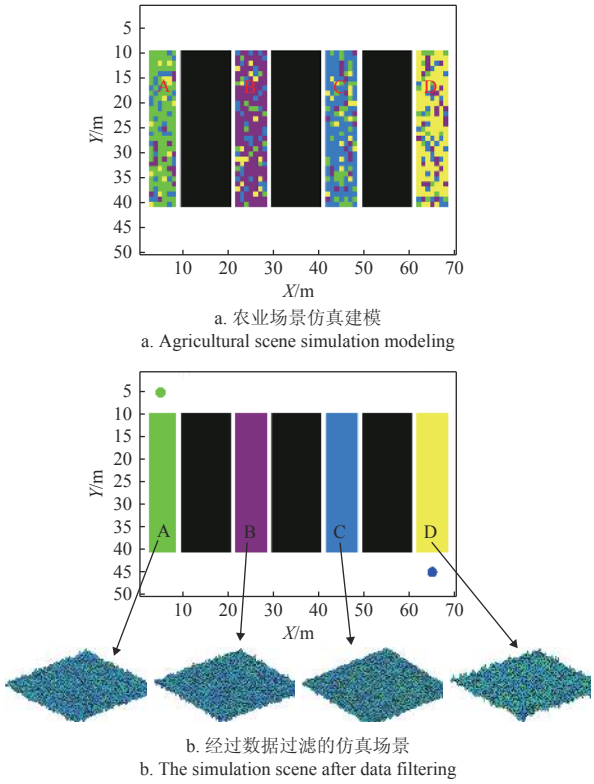


图 7　农业环境实际取景

Fig.7　Actual scene capture of the agricultural environment

由于地形复杂，本文对 IMU 数据进行了滤波、去噪处理，仅保留 4 种粗糙度颜色（表 2），白色区域表示无粗糙度地面。在仿真试验中，为了减少计算量，对图 8 中采集到的过道数据进行过滤处理，仅考虑影响最大的因素，其中“绿点”代表始点，“蓝点”代表终点，A～D 为不同地面粗糙度，底盘在路径规划过程中，需要避免触碰垄道，并对地面粗糙度进行能量管理分析，使得强化学习算法在路径规划时自动权衡多种因素（如式（2）所示）以找到最优路径。



a. 农业场景仿真建模
a. Agricultural scene simulation modeling

b. 经过数据过滤的仿真场景
b. The simulation scene after data filtering

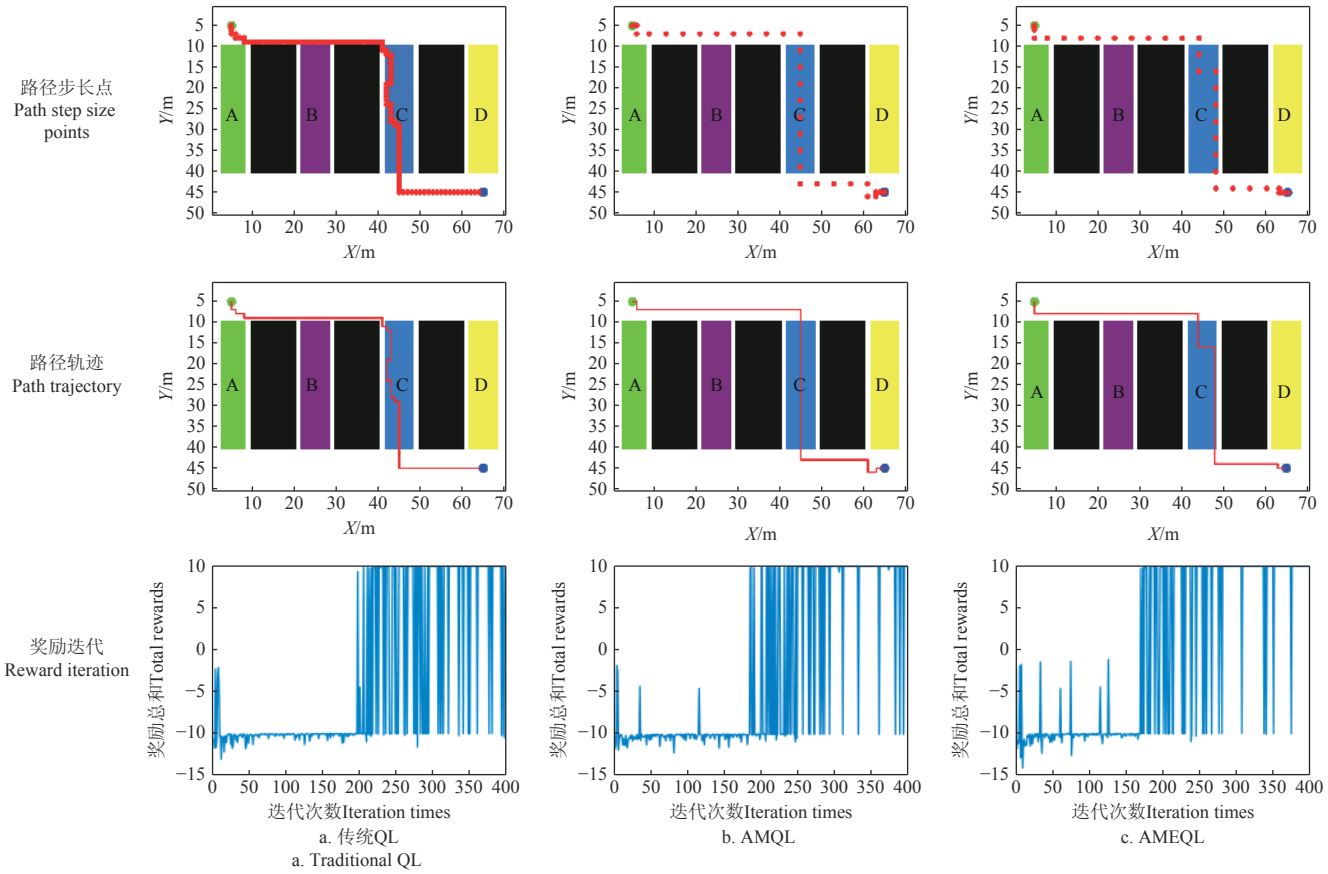
注：图中 A~D 分别为表 2 中绿色、紫色、蓝色、黄色对应的粗糙度等级。
Note: In the figure, A to D represent the roughness levels corresponding to green, purple, blue, and yellow in table 2.

图 8 仿真环境搭建
Fig.8 Simulation environment construction

5 仿真验证

为验证所提出方法的有效性，本文将试验分为 3 种：传统 Q-Learning 算法（QL）、基于自适应多步长的优化算法（AMQL）以及融合了 AMQL 与自适应变 ε-greedy 的优化算法（AEQL）的优化算法（AMEQL），用以实现对底盘的能量管理控制策略进行仿真测试。在仿真中，3 种强化学习算法参数：学习率为 0.1，折扣率为 0.9，初始探索率为 1，探索衰减率为 0.99/0.99/式（5）-（8），最小探索率为 0.01，最大迭代次数为 400，最大探索步数为 300，初始探索步长为 1/自适应多步长/自适应多步长。图 9 为 3 种算法在 400 次迭代之后的路径规划对比，其中包含所走的路径步长点（可观测其自适应步长变化）、训练输出最优路径以及 3 种算法的奖励迭代，图中“红色圆点”表示路径点，相邻路径点的距离表示步长，“红线”表示所走路径。由图可知，3 种算法所走路径均为第三垄道，符合预期的分析。

通过对传统 QL 算法、AMQL 算法以及 AMEQL 算法进行对比分析可知：传统 QL 算法收敛速度慢，稳定性差，且拐点的数量较多，在 300 次迭代之后，奖励值波动仍旧较大，在实际进行路径规划时并不可取；相对于传统 QL 算法，AMQL 的迭代速度提升相对明显，稳定性更强，在 300 次迭代之后，奖励值的波动相对有所降低；AMEQL 算法的收敛速度最快，收敛效率最高，在迭代 280 次时，奖励值就趋于稳定，算法稳定性最高。



a. 传统QL
a. Traditional QL

b. AMQL

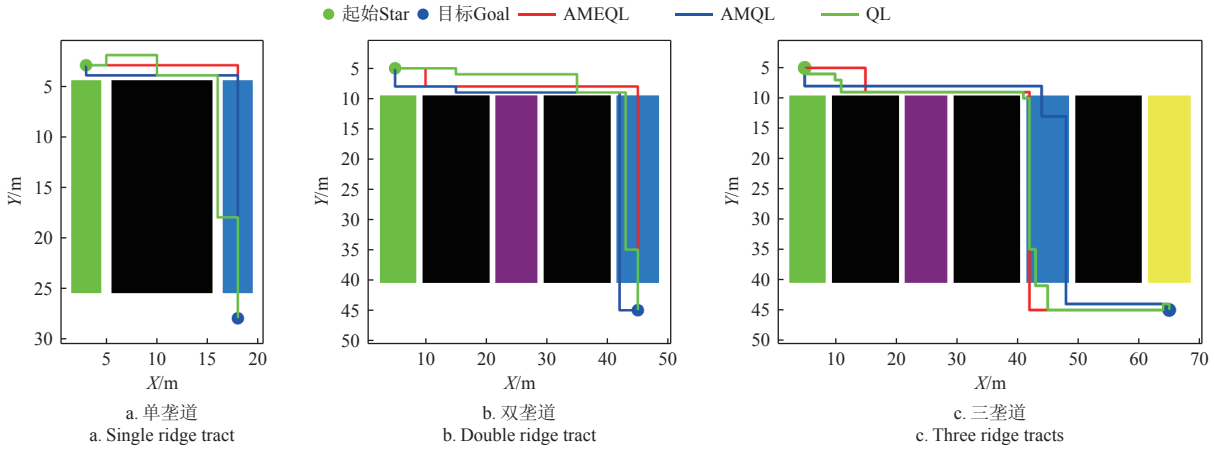
c. AMEQL

图 9 3 种试验方法在 400 次迭代之后的路径规划对比
Fig.9 Comparison of path planning after 400 iterations among the three experimental methods

为了体现改进算法在解决问题方面的有效性，本文通过 300 轮仿真试验，其相关强化学习参数与分别对传统 QL 算法、AMQL 算法和 AMEQL 算法在单垄道（30 m×20 m），双垄道（50 m×50 m）以及三垄道（70 m×50 m）下进行路径规划对比，如图 10 所示，数据结果如表 3 所示，在进行完整 400 次迭代时，在单垄道里面，AMQL 算法相比于传统 QL 算法，训练平均时间降低 4.21%，收敛平均迭代次数降低 4.92%，路径平均拐点数降低 66.67%，收敛后的平均波动次数减少 31.25%；AMEQL 算法迭代的平均时间最短，相比于传统 QL 算法，训练平均时间降低 23.18%，收敛平均迭代次数降低 42.16%，路径平均拐点数降低 66.67%，收敛后的平均波动次数降低 10.42%，到达最大奖励后每 7.93 次迭代波动 1 次，而 AMQL 每 7.97 次迭代波动 1 次。

在双垄道里面，AMQL 算法相比于传统 QL 算法，训练平均时间降低 32.54%，收敛平均迭代次数降低 13.38%，路径平均拐点数降低 54.54%，收敛后的平均波动次数减少 10.87%；AMEQL 算法相比于传统 QL 算法，训练平均时间降低 36.09%，收敛平均迭代次数降低 17.83%，路径平均拐点数降低 72.73%，收敛后的平均波动次数降低 17.39%，到达最大奖励后每 7.13 次迭代波动 1 次，而 AMQL 每 6.43 次迭代波动 1 次。

在三垄道里面，AMQL 算法相比于传统 QL 算法，训练平均时间降低 34.46%，收敛平均迭代次数降低 23.68%，路径平均拐点数降低 41.67%，收敛后的平均波动次数降低 11.91%；AMEQL 算法相比于传统 QL 算法，训练平均时间降低 34.46%，收敛平均迭代次数降低 23.68%，路径平均拐点数降低 63.13%，收敛后的平均波动次数降低 15.62%，到达最大奖励后每 6.32 次迭代波动 1 次，而 AMQL 每 5.64 次迭代波动 1 次。



a. 单垄道
a. Single ridge tract

b. 双垄道
b. Double ridge tract

c. 三垄道
c. Three ridge tracts

图 10　3 种仿真环境

Fig.10　Three kinds of simulation environments

表 3　3 种算法仿真数据对比

Table 3　Comparison of simulation data of the three algorithms

垄道大小 Ridge size	收敛平均迭代次数 Convergence average number of iterations			训练平均使用时间 Average training usage time/s			路径平均拐点数 Average number of turning points on the path			收敛后的平均波动次数 Average number of fluctuations after convergence		
	QL	AMQL	AMEQL	QL	AMQL	AMEQL	QL	AMQL	AMEQL	QL	AMQL	AMEQL
单垄道	102	97	59	95	91	73	9	3	3	48	38	43
双垄道	157	136	129	169	114	108	11	5	3	46	41	38
三垄道	208	191	185	195	121	109	12	7	6	42	37	34

通过对 3 种垄道的数据进行统计分析：AMQL 相对于传统 QL 算法，训练平均时间降低 23.74%，收敛平均迭代次数降低 14.01%，路径平均拐点数降低 54.29%，收敛后的平均波动次数降低 18.01%；AMEQL 相对于传统 QL 算法，训练平均时间降低 34.46%，收敛平均迭代次数降低 23.68%，路径平均拐点数降低 63.13%，收敛后的平均波动次数减少 15.62%，在 400 次迭代过程中，AMEQL 到达最大奖励后平均每 7.12 次迭代波动 1 次，而 AMQL 平均每 6.68 次迭代波动 1 次。可知 AMEQL 训练时间最短，收敛最快，路径拐点数量最低，奖励波动最小，而 AMQL 次之。

6　结　论

本文针对温室底盘在行驶过程中地面粗糙度对电池寿命、利用率的影响以及传统路径规划、深度学习路径规划的问题，提出了基于先验知识构建分级预打分奖励模型的方法，并通过增加曼哈顿距离构建奖励函数，提高了电池寿命和利用率；同时，针对传统 Q-Learning 算法的迭代时间长、收敛效率低、易陷入局部最优及拐点多等问题，本文提出了自适应变步长的优化算法（adaptive multi-step Q-learning，AMQL）和基于自适应改变探索率的优化算法（adaptive ε -greedy Q-learning，AEQL），并将两者进行融合，提出了一种自适应多步长和变 ε -greedy 融合算法（adaptive Multi-step and ε-greedy Q-learning，AMEQL），通过仿真对比的方式，验证了 AMEQL 与 AMQL 算法的性能表现。最后通过对 3 种垄道的数据进行数据统计。结果表明，AMEQL 训练平均使用时间最短，收敛速度最快，路径平均拐点

数缩减最多，训练效果最稳定，AMQL 次之：

1）AMEQL 相对于传统 QL 算法，训练平均时间降低 34.46%，收敛平均迭代次数降低 23.68%，路径平均拐点数降低 63.13%，收敛后的平均波动次数减少 15.62%；

2）AMQL 相对于传统 QL 算法，训练平均时间降低 23.74%，收敛平均迭代次数降低 14.01%，路径平均拐点数降低 54.29%，收敛后的平均波动次数降低 18.01%；

3）在 400 次迭代过程中，AMEQL 到达最大奖励后平均每 7.12 次迭代波动 1 次，而 AMQL 平均每 6.68 次迭代波动 1 次，可知 AMEQL 训练时间最短，收敛最快，路径拐点数量最低，奖励波动最小，而 AMQL 次之。

[参 考 文 献]

[1] 宋成宝，柳平增，刘兴华，等. 基于温湿度异布的日光温室冬季主动通风策略设计与验证[J]. 农业工程学报，2024，40(10)：228-238.
SONG Chengbao, LIU Pingzeng, LIU Xinghua, et al. Design and verification of the active ventilation strategy in solar greenhouse in winter based on the heterogeneous distribution of temperature and humidity[J]. Transactions of the Chinese Society of Agricultural Engineering(Transactions of the CSAE, 2024, 40(10): 228-238. (in Chinese with English abstract)

[2] 张观山，丁小明，何芬，等. 基于 LSTM-AT 的温室空气温度预测模型构建[J]. 农业工程学报，2024，40(18)：194-201.
ZHANG Guanshan, DING Xiaoming, HE Fen, et al. Construction of greenhouse air temperature prediction model based on LSTM-AT[J]. Transactions of the CSAE, 2024, 40(18): 194-201. (in Chinese with English abstract)

[3] 孙颖. 日光温室老山芹生长发育模型研究[D]. 哈尔滨：东北农业大学，2023.
SUN Ying. Study on the Growth and Development Model of Old Celery inSolar Greenhouse[D]. Harbin: Northeast Agricultural University, 2023. (in Chinese with English abstract)

[4] MALINAUSKAITE J, JOUHARA H, AHMAD L, et al. Energy efficiency in industry: EU and national policies in Italy and the UK[J]. Energy, 2019, 172: 255-269.

[5] 秦硕璞，李婷，张雅京，等. 中国农业源非 CO2 温室气体排 放核算[J/OL]. 生态学报，2024，（17）：1-16
QIN Shuopu, LI Ting, ZHANG Yajing, et al. Accounting of non-CO_2 greenhouse gas emissions from agricultural sources in China[J]. Acta Ecologica Sinica, 2024, (17): 1-16. (in Chinese with English abstract)

[6] TAYLOR R I. Energy efficiency, emissions, tribological challenges and fluid requirements of electrified passenger car vehicles[J]. Lubricants, 2021, 9(7): 66.

[7] 万俊，孙薇，葛敏，等. 基于含避障角人工势场法的机器人路径规划[J]. 农业机械学报，2024，55(1)：409-418.
WAN Jun, SUN Wei, GE Min, et al. Robot path planning based on artificial potential field method with obstacle avoidance angles[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(1): 409-418. (in Chinese with English abstract)

[8] AGGARWAL S, KUMAR N. Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges[J]. Computer communications, 2020, 149: 270-299.

[9] 时维国，宁宁，宋存利，等. 基于蚁群算法与人工势场法的移动机器人路径规划[J]. 农业机械学报，2023，54(12)：407-416.
SHI Weiguo, NING Ning, SONG Cunli, et al. Path planning of mobile robots based on ant colony algorithm and artificial potential field algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(12): 407-416. (in Chinese with English abstract)

[10] CABREIRA T M, BRISOLARA L B, PAULO R F J. Survey on coverage path planning with unmanned aerial vehicles[J]. Drones, 2019, 3(1): 4.

[11] Boryga M, Kolodziej P, Golacki K. Application of polynomial transition curves for trajectory planning on the headlands[J]. Agriculture, 2020, 10(5): 1-16.

[12] 沈跃，张凌飞，沈亚运，等. 基于相邻争夺算法的无人机多架次植保作业路径规划[J]. 农业工程学报，2024，40(16)：44-51.
SHEN Yue, ZHANG Lingfei, SHEN Yayun, et al. Path planning of multi-round plant protection operations of unmanned aerial vehicles based on adjacent competition algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(16): 44-51. (in Chinese with English abstract)

[13] 孙月平，方正，袁必康，等. 基于 FIA*-APF 算法的蟹塘投饵船动态路径规划[J]. 农业工程学报，2024，40(9)：137-145.
SUN Yueping, FANG Zheng, YUAN Bikang, et al. Dynamic path planning of crab pond baiting vessels based on FIA*-APF algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(9): 137-145. (in Chinese with English abstract)

[14] DANG T V, BUI N T. Obstacle avoidance strategy for mobile robot based on monocular camera[J]. Electronics, 2023, 12(8): 1932.

[15] MABOUDI M, HOMAEI M R, SONG S, et al. A review on viewpoints and path planning for UAV-based 3D reconstruction[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 5026-5048.

[16] 孙国祥，黄银锋，汪小昆，等. 基于 LIO-SAM 建图和激光视觉融合定位的温室自主行走系统[J]. 农业工程学报，2024，40(3)：227-239.
SUN Guoxiang, HUANG Yinfeng, WANG Xiaochan, et al. Greenhouse autonomous walking system based on LIO-SAM mapping and laser-vision fusion positioning[J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(3): 227-239. (in Chinese with English abstract)

[17] Padakandla S. A survey of reinforcement learning algorithms for dynamically varying environments[J]. ACM Computing Surveys, 2021, 54(6): 127

[18] 熊俊涛，李中行，陈淑绵，等. 基于深度强化学习的虚拟机器人采摘路径避障规划[J]. 农业机械学报，2020，51(s2)：1-10.
XIONG Juntao, LI Zhonghang, CHEN Shumian, ZHENG Zhenhui. Obstacle avoidance planning of virtual robot picking path based on deep reinforcement learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(s2): 1-10. (in Chinese with English abstract)

[19] GAO J, YE W, GUO J, et al. Deep reinforcement learning for indoor mobile robot path planning[J]. Sensors, 2020, 20(19): 5493.

[20] LAZZARONI L, BELLOTTI F, CAPELLO A, et al. Deep reinforcement learning for automated car parking[C]//International Conference on Applications in Electronics Pervading Industry, Environment and Society. Cham: Springer Nature Switzerland, 2022: 125-130.

[21] FAN J, WANG Z, XIE Y, et al. A theoretical analysis of deep Q-learning[C]//Learning for dynamics and control. PMLR,

2020: 486-489.
[22] XU S, GU Y, LI X, et al. Indoor emergency path planning based on the Q-learning optimization algorithm[J]. ISPRS International Journal of Geo-Information, 2022, 11(1): 66.
[23] 吕悦来，李广毅. 地表粗糙度与土壤风蚀[J]. 土壤学进展，1992，20(6)：38-42.
LYU Yuelai, LI Guangyi. Surface roughness and soil wind erosion[J]. Advances in Soil Science, 1992, 20(6): 38-42. (in Chinese with English abstract)
[24] 顾钧禧. 大气科学辞典[M]. 北京：气象出版社，1994.
[25] LI W, JIAO Z, XIAO Q, et al. A study on performance characterization considering six-degree-of-freedom vibration stress and aging stress for electric vehicle battery under driving conditions[J]. IEEE Access, 2019, 7: 112180-112190.
[26] HUA X, THOMAS A. Effect of dynamic loads and vibrations on lithium-ion batteries[J]. Journal of Low Frequency Noise, Vibration and Active Control, 2021, 40(4): 1927-1934.
[27] ZHANG W, LI X, WU W, et al. Influence of mechanical vibration on composite phase change material based thermal management system for lithium-ion battery[J]. Journal of Energy Storage, 2022, 54: 105237.
[28] LEE P Y, PARK S, CHO I, et al. Vibration-based degradation effect in rechargeable lithium ion batteries having different cathode materials for railway vehicle application[J]. Engineering Failure Analysis, 2021, 124: 105334.
[29] SOMERVILLE L, HOOPER J M, MARCO J, et al. Impact of vibration on the surface film of lithium-ion cells[J]. Energies. 2017, 10: 741.
[30] ZHANG L, NING Z, PENG H, et al. Effects of vibration on the electrical performance of lithium-ion cells based on mathe matical statistics. Appl Sci, 2017, 7: 802.
[31] HOOPER J M, MARCO J, Chouchelamane G H, et al. Vibration durability testing of nickel cobalt aluminum oxide (NCA) lithium-ion 18650 battery cells[J]. Energies, 2016, 9(4): 281.
[32] GIM J, AHN C. IMU-based virtual road profile sensor for vehicle localization[J]. Sensors, 2018, 18(10): 3344.

Reinforcement learning-based optimization algorithm for energy management and path planning of robot chassis

LI Xiaoyu[1] , ZHANG Junhua[1] , GUO Xiaoguang[1※] , WU Gang[2]

(1. *School of Mechanical and Electrical Engineering*, *Beijing Information Science and Technology University*, *Beijing* 100192, *China*;　2. *Institute of Agricultural Environment and Sustainable Development*, *Chinese Academy of Agricultural Sciences*, *Beijing* 100081, *China*)

Abstract: Ground roughness can significantly impact the battery performance in greenhouse environments. In this study, battery energy management was integrated with path planning to address this challenge. A systematic investigation was also implemented to explore the effects of ground roughness on the battery life and utilization efficiency of greenhouse vehicle platforms. A graded pre-scoring model was constructed using prior knowledge. Additionally, the Manhattan distance between the vehicle's current position and the target point was incorporated into the reinforcement learning reward function, thus linking travel distance with battery life to optimize both battery utilization efficiency and life during path planning. An Adaptive Multi-step Q-learning algorithm (AMQL) with adaptive step sizes and an Adaptive ε-greedy Q-learning algorithm (AEQL) with an adaptive exploration rate was proposed to enhance the performance of the Q-learning algorithm. The traditional Q-learning algorithms were associated with some issues, such as long iteration times, low convergence efficiency, susceptibility to local optima, and excessive path turns. The AMQL algorithm was used to adjust the step size, according to the forward reward assessment—if the reward at the current position increased corresponding to the previous reward, the step size increased. The step size gradually decreased to prevent suboptimal path optimization, as the current position approached the endpoint. The AEQL algorithm was used to adaptively adjust the exploration rate ε using the difference between adjacent reward values—ε increased when the adjacent reward value increased, and ε decreased when the reward value decreased. Although AMQL improved the convergence efficiency and iteration speed, the variations in the step size caused significant fluctuations in rewards, resulting in lower algorithm stability. Additionally, there was no outstanding impact of multi-step length on the convergence efficiency and iteration speed. Furthermore, the AEQL enhanced the exploration efficiency and algorithm stability through dynamic adjustments. But its fluctuating rise during the initial training phase also increased the training time. Therefore, the AMQL and AEQL algorithms were combined to develop an Adaptive Multi-step and ε-greedy Q-learning algorithm (AMEQL), in order to ensure faster and more optimal global path selection during path planning. In a simulated environment, the models were first used to simulate a realistic greenhouse tomato scenario. Then, an Inertial Measurement Unit (IMU) was used to record the changes in the aisle roughness in real time. This data was then incorporated into the simulation model. Finally, 300 rounds of simulation experiments were carried out to test the traditional Q-learning, AMQL, and AMEQL algorithm for path planning in the single-row (30 m×20 m), double-row (50 m×50 m), and triple-row (70 m×50 m) environments. Simulation results show that the AMEQL algorithm reduced the average training time by 34.46%, the average number of iterations required for convergence by 23.68%, the number of path turns by 63.13%, and the post-convergence average fluctuation by 15.62%, compared with the traditional Q-learning. Due to its higher convergence speed in 400 iterations, the AMEQL algorithm averaged 14 fluctuations per 100 iterations after reaching the maximum reward, while the AMQL algorithm averaged 15 fluctuations. This algorithm can provide a theoretical reference for the autonomous path planning of greenhouse platforms

Keywords: greenhouse; path planning; reinforcement learning; energy management; multi-objective optimization