Modeling and Simulation 建模与仿真, 2023, 12(6), 5294-5302
Published Online November 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.126481

Hans汉斯

仓储中基于多智能体深度强化学习的多AGV路径规划

王梅芳，关　月

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2023年9月18日；录用日期：2023年11月7日；发布日期：2023年11月14日

摘　要

随着工业自动化和物流行业的迅速发展，自动引导车辆(Automated Guided Vehicle, AGV)在物流仓库中的路径规划已成为确保运输效率和准确性的关键环节。尽管近年来已经有很多策略被提出，但多AGV系统在复杂的物流环境中仍然频繁地出现碰撞、路径冲突以及控制迟延等问题。鉴于此，本研究提出了一种基于多智能体深度强化学习(Multi Agent Deep Reinforcement Learning, MADRL)的路径规划方法，以期解决多AGV之间的相互协调问题并提高其路径规划效率。为验证所提方法的有效性，我们采用了与遗传算法(Genetic Algorithm, GA)的比较实验。结果显示，基于MADRL的策略在整体运输效率上实现了28%的提升，并在碰撞事件上有了明显的减少。

关键词

路径规划，MADRL，AGV，仓储

Multi-AGV Path Planning in Warehousing Based on Multi-Agent Deep Reinforcement Learning

Meifang Wang, Yue Guan

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: Sep. 18th, 2023; accepted: Nov. 7th, 2023; published: Nov. 14th, 2023

Abstract

With the rapid advancement of industrial automation and the logistics industry, the path planning

文章引用: 王梅芳, 关月. 仓储中基于多智能体深度强化学习的多 AGV 路径规划[J]. 建模与仿真, 2023, 12(6): 5294-5302. DOI: 10.12677/mos.2023.126481

of Automated Guided Vehicles (AGV) in logistics warehouses has become a critical component to ensure transportation efficiency and accuracy. Although numerous strategies have been proposed in recent years, multi-AGV systems still frequently encounter collisions, path conflicts, and control latencies in complex logistics environments. In light of this, our study introduces a path planning approach based on Multi-Agent Deep Reinforcement Learning (MADRL) aiming to address the coordination issues among multiple AGVs and to enhance their path planning efficiency. To validate the effectiveness of the proposed method, we conducted comparative experiments with the Genetic Algorithm (GA). Results show that the MADRL-based strategy achieved a 28% improvement in overall transportation efficiency and a significant reduction in collision incidents.

Keywords

Path Planning, MADRL, AGV, Warehousing

1. 引言

自动引导车辆(Automated Guided Vehicle, AGV)已为材料搬运和仓储物流领域带来革命性变革[1] [2]。在高度自动化的现代仓库中，AGV 能够根据系统的调度，在货架之间快速移动，精确地定位并取得或存放货物，大大提高了出入库效率。其精准的导航系统和可编程路径确保了货物被高效、安全和准确地搬运。在当今工业 4.0 智能汽车制造厂中，如特斯拉、理想、比亚迪等公司广泛采用 AGV 来搬运汽车部件，从而确保了生产线的连续性和流畅性[3] [4]。此外，在智能化的物流产业中，AGV 能够为工作人员精确地搬运货物，显著提高了仓储效率[5]。最初仅用于制造车间中搬运笨重材料的 AGV，如今已经逐渐演变成现代物流仓储解决方案的核心组成部分。在当今错综复杂的物流领域，AGV 在加速流程、减少人工劳动和优化存储空间方面都显示出其不可替代的价值[6] [7] [8]。

然而在物流仓储中，多 AGV 同时进行路径规划还存在诸多挑战[9] [10] [11]。首先，AGVs 之间可能会发生路径交叉和碰撞，尤其是在空间有限且结构复杂的仓库内，当 AGV 出现碰撞或锁死将会严重影响货物搬运效率[12]。其次，多个 AGV 的实时协同和调度，以满足高效率和低延迟的要求，也是一个难以解决的问题。传统的基于遗传算法(Genetic Algorithm, GA)算法的路径规划方法很难应对这些动态和复杂的挑战[12] [13] [14]。而深度强化学习已经成为人工智能领域的一个重要技术，为智能体提供了通过与环境互动来学习最优策略的途径[15]。但是利用 DRL 进行小车路径规划任然存在状态空间和动作空间的维度灾难[16]。因此，面对多 AGV 情景，我们基于多智能体深度强化学习(Multi Agent Deep Reinforcement Learning, MADRL)来对 AGV 的状态、动作、奖励进行建模。MADRL 可以有效地整合多个 AGV 的集体潜力，使它们在一个共享的环境中进行有效的协同、规避碰撞和实现快速路径规划。

本文深入探讨了在物流仓储背景下利用 MADRL 进行多 AGV 路径规划的复杂性。鉴于物流行业长期面临的碰撞、路径冲突和控制迟延等挑战，我们的研究引入了一种新颖的基于 MADRL 的 AGV 路径规划策略。与经常采用预定义规则或静态算法的传统方法不同[17]，我们的方法能够动态适应实时变化，确保最佳的 AGV 协调。此外，在我们的研究中，为了验证所提出的 MADRL 方法的有效性，我们与传统的 GA 进行了比较。实验结果显示，基于 MADRL 的路径规划策略在运输效率上实现了 33%的显著提升，并且能够显著减少 AGV 之间的碰撞事件。

2. 模型

2.1. 仓储环境

下图 1 为仓库中多 AGV 的搬运场景图，本文将仓储的自动分拣系统采用栅格化进行建模，即将场景栅格化为一个个相同大小的正方形[栅格化]。其主要包括以下部分：

- 道路：在图 1 中表示为白色栅格。这些道路用于 AGV 行驶，且其中的 AGV 可以自由向四个方向行驶，即 AGV 可实时调整行驶策略，避免碰撞和锁死。
- AGV：它们以一定的速度行驶并搬运包裹。
- 自动分拣机：它依据包裹的收获地址发出运输指令，通知 AGV 将包裹运送到指定的货物缓存区。
- 货物缓存区：在图中由绿色栅格代表，基于包裹的收货地址，在地图上均匀地划分出多个投递区域。每个投递区域都配备一个竖直向上容量有限的存储空间。当该空间内的包裹数目满时，这些包裹将被从仓库中转移出，以便进行后续的装车和发货。
- 障碍：由一些固定物，如楼梯、门、工人活动区等构成，AGV 不能到达此区域。
- 充电站：它为蓝色栅格，用于提供给 AGV 进行充电，AGV 工作时不能经过占用。当 AGV 点亮到达预警值时，将自动寻找最近空充电站进行充电。

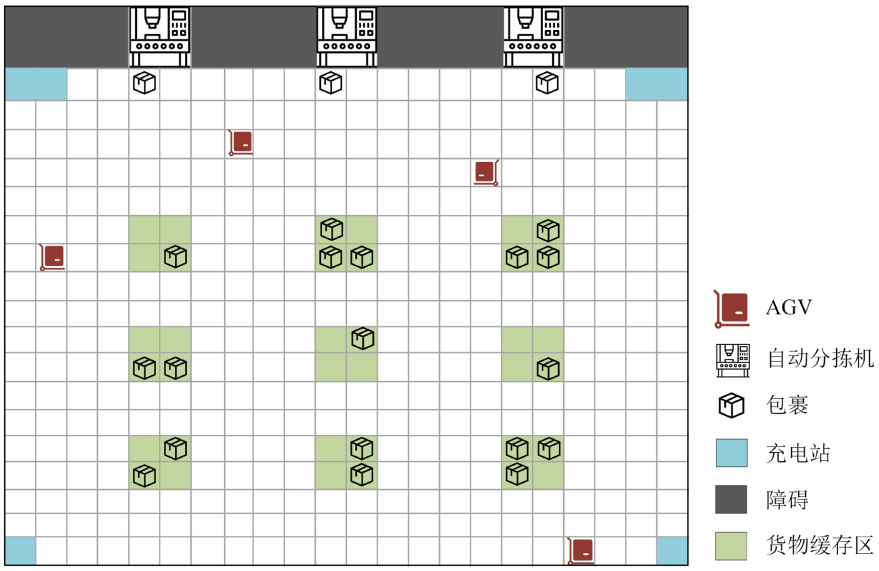

Figure 1. Multi-AGV handling scenario in the warehouse
图 1. 仓库中多 AGV 搬运场景

在不失一般性的前提下，本文假设 AGV 都是单载量的且匀速行驶，即 AGV 一次只能运输一个包裹和经过每个栅格的时间相同；假设 AGV 在工作过程中均不会发生物理损坏；假设 AGV 的位置基于 GPS 和 RFID 可实时定位；假设 AGV 基于 5G 通信实时获取自动分拣机发出的包裹运输命令；假设时间被均匀的分为一个个时长为 t 的时隙，AGV 移动一格距离的时长定义为一个时隙。如果搬运场景发生显著的改变，本文的结论可能需要重新验证。

2.2. 问题描述与建模

在仓储环境下的多 AGV 的路径规划问题可以由以下描述。设 $\mathcal{I} \triangleq \{i \mid i=1,2,3,\cdots,I\}$、$\mathcal{J} \triangleq \{j \mid j=1,2,3,\cdots,J\}$ 和 $\mathcal{K} \triangleq \{k \mid k=1,2,3,\cdots,K\}$ 分别代表仓库中的 AGVs、包裹和货物缓冲区。包裹 j 所

对应自动分拣机栅格坐标 p_j^{start} 和对应的货物缓存区坐标 p_j^{end} 已知，坐标 $p \in \mathbb{R}^2$ 是通过笛卡尔坐标系来确定的。AGV 载货和空载由 ω 表示，即令 ω_i 表示第 i 台 AGV 的载货状态：

$$\omega_i = \begin{cases} 1, & \text{如果第 } i \text{ 台 AGV 载货} \\ 0, & \text{如果第 } i \text{ 台 AGV 空货} \end{cases} \tag{1}$$

当自动分拣机分发出一个新的包裹，将基于包裹与周围的空 AGV 的曼哈顿距离进行任务派发，AGVi 与包裹 j 的曼哈顿距离给出为：

$$d_{ij} = \left|x_i - x_j\right| + \left|y_i - y_j\right| \tag{2}$$

其中，有坐标 $p = (x, y)$，x 表示网格的横坐标；y 表示网格的纵坐标。包裹 j 计算与所有空 AGV 的曼哈顿距离由下给出：

$$\min_{i \in I}\left(d_{ij} = \left|x_i - x_j\right| + \left|y_i - y_j\right|\right), \quad \text{if } \omega_i = 0 \tag{3}$$

AGV 的路径规划需要决策 AGV 行驶的路径和进行是否停车判断，其策略目标为最小化搬运包裹的总和时间 $C_{\max}$，$C_{\max}$ 为所有包裹的搬运时间和等待时间相加，具体如下：

$$C_{\max} = \sum_{j=1}^{J}\left(w_j + t_j\right) \tag{4}$$

其中 w_j 是包裹放置在分拣台的等待时间，t_j 表示包邮由 AGV 运输到指定货物缓存区的时间。因为 AGV 的速度 v 恒定，则 $t_j = d_j^{tran} \times v$，$d_j^{tran}$ 为包裹起始坐标 p_j^{start} 与包裹终点坐标 p_j^{end} 的曼哈顿距离。

2.3. 数学公式化

基于上述的假设和场景建模，针对多 AGV 的实时协作搬运包裹的路径规划问题，建立以最小化搬运完成时间为目标的优化目标，如下所示：

$$\min C_{\max} \tag{5}$$

$$\text{s.t. } \omega(t) \in (0,1) \tag{6}$$

$$d_j^{tran} \geq 0, \quad j \in J \tag{7}$$

$$p_i(t) \in [x_{\min}, x_{\max}], \quad i \in I \tag{8}$$

$$p_i(t) \in [y_{\min}, y_{\max}], \quad i \in I \tag{9}$$

$$\phi_k(t) \in (0,1,2,\cdots,N), \quad k \in K \tag{10}$$

公式(5)表示最小化包裹搬运总时长；公式(6)约束一个 AGV 只能搬运一个包裹，且一个包裹只能由一个 AGV 搬运；公式(7)约束包裹的搬运距离必须为正数；公式(8)和(9)约束 AGVs 的行驶范围，其不能超出仓库区域；公式(10)约束货物缓冲区的容量界限，单个缓冲区最多容纳 N 个包裹。

3. 算法

在本章中，我们基于先前构建的数学模型，深入探讨多 AGV 路径规划的复杂性和连续性特征。为此，我们将其路径规划过程建模为分布式部分可观测马尔可夫决策过程(Decentralized Partially Observable Markov Decision Process, Dec-POMDP) [18]，并借助 MADRL 算法进行有效求解。

3.1. 强化学习环境建模

在 AGV 研究领域中，DRL 为我们提供了一种新颖的方法来解决路径规划、任务分配和决策问题。

回顾 1954 年，当 Minsky 首次描述强化学习的概念时，他提到的是智能体如何通过与环境的试错互动来优化报酬[19]。这种理念在 AGV 系统中找到了其实际应用场景。在复杂的物流环境中，AGV 作为智能体，需要确定最佳的路径以避免障碍物、减少碰撞风险，并有效地完成货物运输任务。起初，AGV 并不知道哪条路径最优或如何避免碰撞。这时，深度强化学习的方法允许 AGV 在其操作环境中不断尝试、学习并调整其策略。通过与环境的持续互动，AGV 可以根据返回的奖励或惩罚来评估并调整其行动策略，从而找到最优的路径和决策策略。

图 2 揭示了强化学习的核心结构。在此结构中，智能体(Agent)作为决策实体，位于环境(Environment)中进行行动选择(Action)。每执行一次动作，环境的状态(State)都将改变并为智能体提供奖励(Reward)反馈，这种奖励量化了所执行动作相对于任务目标的适当性。智能体旨在最大化其长期累积奖励，因此它需要基于当前状态及其历史经验制定策略。在与环境的持续互动中，智能体不断地评估和优化其策略，以确定在给定状态下哪些动作最可能产生最大的期望奖励。随着互动的深入，智能体持续地优化其决策策略，以更好地满足任务需求。

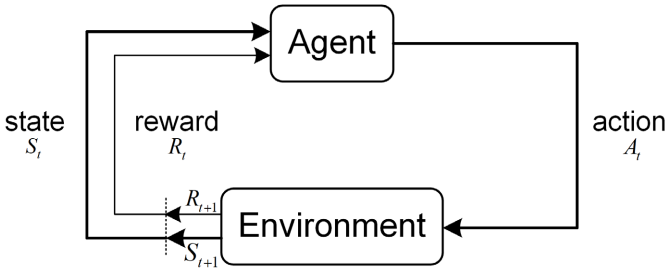

Figure 2. Core structure of reinforcement learning
图 2. 强化学习核心结构

但是，对于多 AGV 进行强化学习策略学习时，较高的状态维度和动作维度和多 AGV 之间存在的动作干扰性将使得算法难以收敛，无法找到纳什均衡。因此为了应对多 AGV 中央控制导致的动作空间维度急剧增长问题，以及单个 AGV 在获取全局信息时的局限性，我们采用了 Dec-POMDP。在此框架下，AGV 的路径规划被详细地划分为：状态空间 $S(t)$、观察集合 $O(t)$、动作集合 $A(t)$ 和奖励函数 $R(t)$。具体定义如下所示：

- 状态空间 $S(t)$：描述了 AGV 在物流环境中的所有可能位置和状态和包裹的状态，$S(t)$ 具体如下表示：

$$S(t)=\{p_1^{agv}(t),p_2^{agv}(t),\cdots,p_I^{agv}(t);p_1^{pack}(t),p_2^{pack}(t),\cdots,p_J^{pack}(t),\\ \phi_1(t),\phi_2(t),\cdots,\phi_k(t),\omega_1(t),\omega_2(t),\cdots,\omega_I(t)\} \tag{11}$$

其中 $p^{agv}(t)$ 和 $p^{pack}(t)$ 分别代表 AGV 和包裹在时隙 t 的位置。$\phi_k(t)$ 代表获取缓存区 k 在时隙 t 的容量。

- 观察集合 $O(t)$：代表了 AGV 在任意时刻所能观察到的环境信息，这包括但不限于其他 AGV 的位置和状态以及任务信息，$O(t)$ 具体如下表示：

$$O(t)=\{p_1^{agv}(t),p_2^{agv}(t),\cdots,p_I^{agv}(t);p_1^{pack}(t),p_2^{pack}(t),\cdots,p_J^{pack}(t),\\ \phi_1(t),\phi_2(t),\cdots,\phi_k(t),\omega_i(t)\} \tag{12}$$

- 动作集合 $A(t)$：表示 AGV 可以执行的所有潜在动作，例如前进、后退、转弯或停止，此外还有装卸货动作，$A(t)$ 具体如下表示：

$$A(t)=\{\text{上、下、左、右、停、装货、卸货}\} \tag{13}$$

- 奖励函数 R：定义了基于 AGV 的动作和其结果对系统整体效益的量化评价。它考虑了路径长度、碰撞风险、任务完成度等多个因素，旨在指导 AGV 作出能够最大化系统效益的决策。R 具体如下表示：

$$R=\beta\times J-\sum_{j=1}^{J}w_j-\sum_{j=1}^{J}\partial\times t_j \tag{14}$$

其中 β 代表完成一次运输所获得的奖励；∂ 代表 AGV 运输时间的惩罚因子，引入 ∂ 是为了惩罚 AGV 移动行为，使其趋向于快速完成包裹运输任务，减少不必要的运行。

3.2. MADRL 算法

MADRL 是一种深度增强学习方法，专门设计来处理多智能体环境中的学习任务。在 MADRL 中，每个智能体都使用深度神经网络来表示其策略，并与其他智能体同时学习和交互。由于多智能体环境的动态性和非静态性，MADRL 需要考虑智能体之间的策略交互和可能的非平稳分布。因此，MADRL 算法经常集成技术，如中央化学习与去中央化执行、多智能体信用分配等，以有效地促进多智能体之间的协作或竞争学习。

本研究提出一种多智能体近端策略优化(Multi-Agent Proximal Policy Optimization, MA-PPO)算法旨在中心化训练 AGV，同时使 AGV 能去中心化自主执行策略。在仓库中 AGVs 不断与环境进行交互使其获得一段经验序列 τ，依据 τ 计算的奖励总和如下：

$$r_t(\theta)=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_n}R\left(\tau^i\right)\pi_\theta\left(\alpha_t^i\mid s_t^i\right) \tag{15}$$

为了找出最优策略使 AGV 获得最大化累积奖励，需要通过最大化奖励目标函数来更新策略网络参数 θ：

$$\nabla\hat{r}_\theta=E_{\tau\,\pi_\theta(\tau)}\left[A^\theta\left(s_t,\alpha_t\right)\nabla\log\pi_\theta\left(\alpha_t^i\mid s_t^i\right)\right] \tag{16}$$

其中 $A^\theta\left(s_t,\alpha_t\right)$ 代表优势函数[20]，接着通过在线学习方法，提示算法样本效率，公式变为：

$$\nabla\hat{r}_\theta=E_{\tau\,\pi_{\theta'}(\tau)}\left[\frac{\nabla\pi_\theta\left(\alpha_t^i\mid s_t^i\right)}{\pi_{\theta'}\left(\alpha_t^i\mid s_t^i\right)}A^{\theta'}\left(s_t,\alpha_t\right)\right] \tag{17}$$

最终引入 KL 散度对 $\hat{r}_\theta$ 进行裁剪：

$$\hat{r}_{clip}^{\theta}(\theta)=E_t\left[min\left(r_t(\theta)A^\theta,clip\left(r_t(\theta),1-\epsilon,1+\epsilon\right)A^\theta\right)\right] \tag{18}$$

4. 仿真实验

4.1. 实验参数

在本论文的仿真实验部分，着重模拟一个基于多模态深度强化学习策略的多 AGV 仓库搬运场景。实验的计算任务部署在一个专业的计算环境中，该环境配备了 AMD 5600X 作为中央处理器(Central Processing Unit, CPU)和 RTX 3080 作为图形处理器(Graphics Processing Unit, GPU)，确保了仿真的计算效率与实时性。神经网络模型的核心是一个六层全连接层，按层的节点数目分别为[64, 128, 128, 256, 128, 64]。这种设计旨在捕捉仓库环境中的复杂特征并高效地为 AGVs 制定策略。每一层都采用了 ReLU 激活函数[21]，以增强模型的非线性表示能力。场景方面，我们构建了一个面积为 50 m × 50 m 的虚拟仓库，进一步栅格化为 0.5 m × 0.5 m 的单元，得到一个细致的 100 × 100 网格地图。这种精细的栅格化可以为仿真提供足够的空间分辨率，确保 AGV 的移动策略与实际仓储操作紧密相符。

我们具体模拟了一个小时的物流货仓搬运场景。AGV 移动速度假设为 50 m/min，其中设置 AGV 为 10 个，自动分拣机为 3 个。假设一个货物缓冲区最大缓存 30 个包裹，在缓存区满载后需等待 2 分钟使得小车运出包裹。更多实验参数和超参数如表 1 所示。基于相关参考文献和初步实验，为 AGV 的有效搬运设定了包裹运输完成奖励。同时，为优化 AGV 的搬运效率，经多轮试验迭代，确定了移动惩罚因子。货物缓存区的最大缓存则考虑了实际场景与仓库规模。其它如学习率、折扣因子、PPO 裁剪系数和广义优势估计器等参数，皆是依据相关文献与初探实验选定。经过多轮实验和敏感性分析，确保了表 1 中参数的可靠性和准确性。

Table 1. Experimental simulation parameters
表 1. 实验模拟参数

参数/超参数	值
包裹运输完成奖励(β)	100
AGV 移动惩罚因子(∂)	2
货物缓存区最大缓存(N)	30
学习率(κ)	0.001
折扣因子(μ)	0.98
PPO 裁剪系数(ϵ)	0.1
广义优势估计器(g^{λ})	0.95

4.2. 结果分析

4.2.1. 收敛性分析

如图 3 所示，为了深入研究学习率对模型收敛性的影响，我们设计了系列实验，并监测了 0.1、0.01、0.001、0.0005 四种学习率下平均系统奖励随实验次数(Episodes)的动态变化，从而研究 MADRL 算法的收敛性。很明显当学习率设定为 0.1 时，系统奖励的增长显著且迅速，且在较少的迭代次数中便趋于稳定。这表明此学习率下，模型具有较快的收敛速度。但需要注意的是，过高的学习率可能会在模型训练初期导致收敛后的性能不好。相比之下，较低的学习率，例如 0.001 和 0.0005，在训练的初阶段显示出递增速度较缓。但在长时间的迭代中，它们展现出了较高的稳定奖励值，暗示这些学习率可能提供了更加稳健的模型泛化。进一步观察，尽管各个学习率均展现出了收敛的趋势，但其最终收敛值有所区别。这进一步突显了选择合适学习率的关键性。从图 3 中可以看出，模型在不同的学习率下都能够收敛，并且在某些学习率下具有更快的收敛速度。这明模型具有较强的稳定性和适应性，能够在不同的参数设置下都达到良好的性能。

4.2.2. 性能分析

如图 4 所示，可以清晰地观察到 MADRL 和 GA 两种算法在不同 AGV 数量情境下的性能表现。首先，不论是 3 个、5 个还是 10 个 AGV，MADRL 算法的单 AGV 每分钟平均运输包裹量都明显超过了 GA 算法。这意味着在这些测试场景中，MADRL 算法具有更高的效率和更好的性能。具体来说，当 AGV 数量为 3 时，MADRL 和 GA 的性能差距达到了 0.44 个包裹/分钟，而当 AGV 数量增加到 10 时，这一差距进一步扩大到了 0.63 个包裹/分钟。这表明，随着 AGV 数量的增加，MADRL 算法的优势更为明显，表

现出更强的扩展性。此外，从 3 到 10 个 AGV，两种算法的性能都有所提升，但 MADRL 的增长速度明显快于 GA，这进一步印证了其在更复杂场景下的优越性能。最终基于上述实验结果，可以推断出 MADRL 在 AGV 调度问题上，无论在小规模还是大规模场景，都展现出了相较于 GA 更为优越的性能。

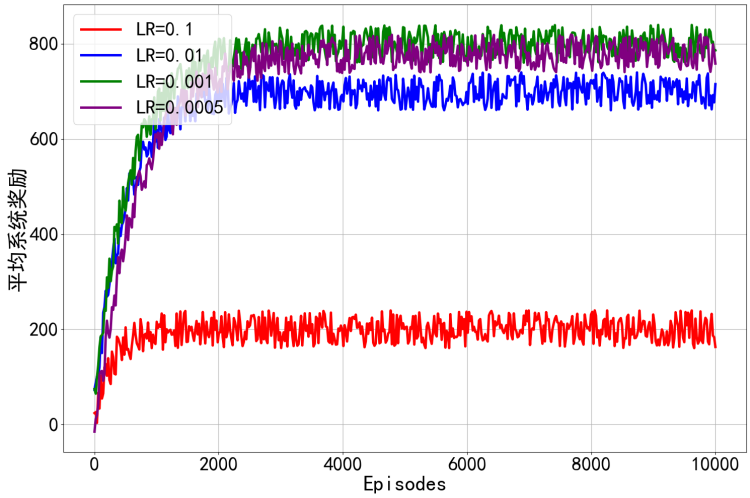

Figure 3. Convergence status of average system reward under different learning rates
图 3. 平均系统奖励在不同学习率下的收敛状况

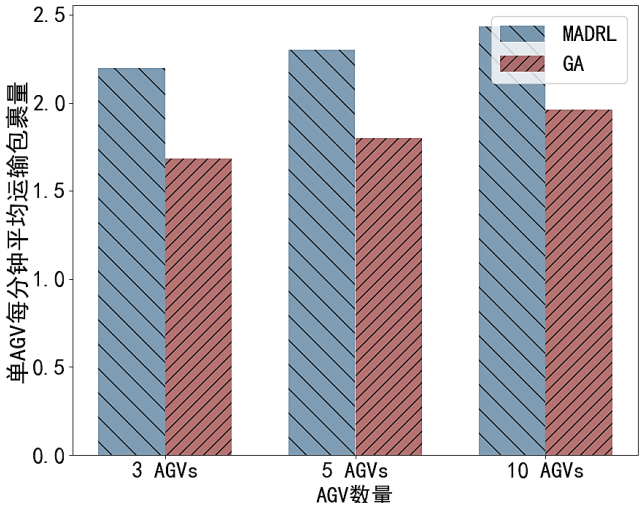

Figure 4. Convergence status of average system reward under different learning rates
图 4. 平均系统奖励在不同学习率下的收敛状况

5. 总结

在本研究中，针对物流仓库中多 AGV 的路径规划问题建立了多 AGV 搬运货物的数学模型，并提出了基于 MADRL 的解决方案。本研究中使用的 MADRL 策略基于特定的环境建模，模拟了真实的物流仓储条件。通过对智能体与环境的深度交互，可以有效地训练 AGVs 进行高效路径规划。经过与 GA 算法的实验比较，证明了 MADRL 在提高运输效率上具有显著优势。综合来看，MADRL 为物流仓储中的多 AGV 协同路径规划提供了一个有效且实用的方法。在未来的研究中，我们期望进一步探索 MADRL 在更复杂的物流场景中的应用，特别是考虑到变化的仓库布局和动态的任务需求。

参考文献

[1] Li, Z., Sang, H., Pan, Q., *et al*. (2022) Dynamic AGV Scheduling Model with Special Cases in Matrix Production Workshop. *IEEE Transactions on Industrial Informatics*, **19**, 7762-7770. https://doi.org/10.1109/TII.2022.3211507

[2] Maximilian, M. (2022) Mensch-KI-Kollaboration in der Smart Factory. *Maschinenbau*, **2**, 1-9. https://doi.org/10.1007/s44029-022-0718-z

[3] Fouad, B. and Dirk, R. (2022) A Review of the Applications of Multi-Agent Reinforcement Learning in Smart Factories. *Frontiers in Robotics and AI*, **9**, Article 1027340. https://doi.org/10.3389/frobt.2022.1027340

[4] Michel, R. (2022) Smart Factory Gets Efficiency Boost from AGV Lift Trucks. *Modern Materials Handling*, **77**, page.

[5] Xu, Y.X., Qi, L., Luan, W.J., Guo, X.W. and Ma, H.J. (2020) Load-In-Load-Out AGV Route Planning in Automatic Container Terminal. *IEEE Access*, **8**, 157081-157088. https://doi.org/10.1109/ACCESS.2020.3019703

[6] Yu, R.R., Zhao, H., Zhen, S.C., *et al*. (2017) A Novel Trajectory Tracking Control of AGV Based on Udwadia-Kalaba Approach. *IEEE/CAA Journal of Automatica Sinica*, 1-13. https://ieeexplore.ieee.org/document/7738999

[7] Digani, V., Sabattini, L. and Secchi, C. (2016) A Probabilistic Eulerian Traffic Model for the Coordination of Multiple AGVs in Automatic Warehouses. *IEEE Robotics and Automation Letters*, **1**, 26-32. https://doi.org/10.1109/LRA.2015.2505646

[8] Zhao, Y., Liu, X., Wang, G., *et al*. (2020) Dynamic Resource Reservation Based Collision and Deadlock Prevention for Multi-AGVs. *IEEE Access*, **8**, 82120-82130. https://doi.org/10.1109/ACCESS.2020.2991190

[9] Han, Y., Cheng, Y. and Xu, G. (2019) Trajectory Tracking Control of AGV Based on Sliding Mode Control with the Improved Reaching Law. IEEE Access, 7, 20748-20755. https://doi.org/10.1109/ACCESS.2019.2897985

[10] Zheng, Z., Qing, G., Juan, C., *et al*. (2018) Collision-Free Route Planning for Multiple AGVs in an Automated Warehouse Based on Collision Classification. *IEEE Access*, **6**, 26022-26035. https://doi.org/10.1109/ACCESS.2018.2819199

[11] Digani, V., Sabattini, L., Secchi, C. and Fantuzzi, C. (2015) Ensemble Coordination Approach in Multi-AGV Systems Applied to Industrial Warehouses. *IEEE Transactions on Automation Science and Engineering*, **12**, 922-934. https://doi.org/10.1109/TASE.2015.2446614

[12] Hu, H., Jia, X.L., Liu, K. and Sun, B.Y. (2021) Self-Adaptive Traffic Control Model with Behavior Trees and Reinforcement Learning for AGV in Industry 4.0. *IEEE Transactions on Industrial Informatics*, **17**, 7968-7979. https://doi.org/10.1109/TII.2021.3059676

[13] Tang, H.T., Cheng, X.Y., Jiang, W.G. and Chen, S.W. (2021) Research on Equipment Configuration Optimization of AGV Unmanned Warehouse. *IEEE Access*, **9**, 47946-47959. https://doi.org/10.1109/ACCESS.2021.3066622

[14] Tao, Q.Y., Sang, H.Y., Guo, H.W. and Wang, P. (2021) Improved Particle Swarm Optimization Algorithm for AGV Path Planning. *IEEE Access*, **9**, 33522-33531. https://doi.org/10.1109/ACCESS.2021.3061288

[15] Arulkumaran, K., Deisenroth, M.P., Brundage, M. and Bharath, A.A. (2017) Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, **34**, 26-38. https://doi.org/10.1109/MSP.2017.2743240

[16] Du, W. and Ding, S. (2021) A Survey on Multi-Agent Deep Reinforcement Learning: From the Perspective of Challenges and Applications. *Artificial Intelligence Review*, **54**, 3215-3238. https://doi.org/10.1007/s10462-020-09938-y

[17] Qiu, L., Hsu, W.J., Huang, S.Y. and Wang, H. (2002) Scheduling and Routing Algorithms for AGVs: A Survey. *International Journal of Production Research*, **40**, 745-760. https://doi.org/10.1080/00207540110091712

[18] Pajarinen, J. and Peltonen, J. (2011) Periodic Finite State Controllers for Efficient POMDP and DEC-POMDP Planning. *Advances in Neural Information Processing Systems*, **24**, 1-9.

[19] Minsky, M. (1961) Steps toward Artificial Intelligence. *Proceedings of the IRE*, **49**, 8-30. https://doi.org/10.1109/JRPROC.1961.287775

[20] Schulman, J., Moritz, P., Levine, S., *et al*. (2015) High-Dimensional Continuous Control Using Generalized Advantage Estimation. arXiv: 1506.02438.

[21] He, J., Li, L., Xu, J., *et al*. (2018) ReLU Deep Neural Networks and Linear Finite Elements. arXiv: 1807.03973.