文章编号:1671-4598(2023)08-0217-07 DOI:10.16526/j.cnki.11—4762/tp.2023.08.033 中图分类号:TP391 文献标识码:A

基于多智能体强化学习的微装配任务规划方法

徐兴辉[1]，唐大林[2]，顾书豪[1]，左家祺[1]，王晓东[2,3]，任同群[2,3]
(1. 大连理工大学 微纳米技术及系统辽宁省重点实验室，辽宁 大连 116024；
2. 北京航天测控技术有限公司，北京 100041；
3. 大连理工大学 高性能精密制造全国重点实验室，辽宁 大连 116024)

摘要：现有装配任务规划方式多为人工规划，存在低效、高成本、易误操作等问题，为此分析了微装配操作的任务特点，以及对微装配中多操作臂协作与竞争关系进行了详细分析，并提出多智能体强化学习中符合微装配任务特点的动作空间、状态空间以及奖励函数的构建方法；利用CoppeliaSim仿真软件构建合理的仿真模型，对已有设备进行物理建模；构建了基于多智能体深度确定性策略梯度算法的学习模型并进行训练，在仿真环境中对设计的状态、动作空间以及奖励函数进行了逐项实验验证，最终获得了稳定的路径以及完整的任务实施方案；仿真结果表明，提出的环境构建方法，更契合直角坐标运动为主要框架的微装配任务，能够克服现有规划方法的不足，能够实现可实际工程化的多臂协同操作，提高任务的效率以及规划的自动化程度。

关键词：多智能体强化学习；奖励函数；微装配；任务规划；仿真环境构建

Microassembly Task Planning Method Based on Multi-agent Reinforcement Learning

XU Xinghui[1], TANG Dalin[2], GU Shuhao[1], ZUO Jiaqi[1], WANG Xiaodong[2,3], REN Tongqun[2,3]
(1. Dalian University of Technology, State Key Laboratory of High-Performance Precision Manufacturing, Dalian 116024, China; 2. Beijing Aerospace Measurement&. Control Technology Co., Ltd., Beijing 100041, China; 3. Dalian University of Technology, Key Laboratory for Micro/Nano Technology and System of Liaoning Province, Dalian 116024, China)

Abstract: The existing planning methods mostly are manual planning, which have problems such as inefficiency, high cost, and easy misoperation. Thus, the characteristics of microassembly operationtasks, collaboration and competition relationship of micro-assembly operation are analyzed in detail, and a method for the construction of action, state and reward conditions that conforms to the characteristics of micro-assembly tasksin multi-agent reinforcement learningis proposed. Using CoppeliaSim simulation software to model existing equipment physically, a learning model is built and trained based on multi-agent deep deterministic policy gradientalgorithm, then the designed action, state and reward function are verified experimentally in simulation environment. Ultimately a stable path and complete task implementation scheme is obtained. The simulation results show that the proposed method is more suitable for the micro-assembly task with Cartesian coordinate motion as the main framework, and can overcome the shortcomings of existing planning methods. Besides, the method can realize the multi manipulator arm collaborative operation, which can be practically engineered and improve the efficiency of the task and the automation degree of planning.

Keywords: multi-agent reinforcement learning; reward function; micro-assembly; task planning; simulation environment construction

0 引言

国防高端武器中常用的传感器、惯性器件等，具有尺度小、精度高的特点，装配是其制造过程中的重要环节。装配任务就是基于视觉、力觉等传感器对作业环境的感知，控制引导多操作臂协作完成序列零件的拾取、搬运、位姿调整对齐以及释放等操作，即多操作臂的协同操作。由于任务需要综合考虑装配序列信息，操作臂运动区域等条件，导致在操作臂前往目标位置的途中，难以始终保持连续、直线的运动路径，因此通常需要对任务路径进行规划。目前，大多在调试阶段通过多次人工试验找到一个优先保证安全的控制方案。然而这种方式相当耗时，且在试验过程中由于没有明确目标，容易操作失误，造成操作臂之间的

收稿日期:2023-03-09； **修回日期**:2023-03-14。
基金项目:国家重点研发计划资助项目(2019YFB1310901)；辽宁省“兴辽英才计划”资助项目(XLYC2002020)；辽宁省自然科学基金项目(2020-MS-104)。
作者简介:徐兴辉(1998-)，男，甘肃临夏人，硕士研究生，主要从事精密微小装配技术方向的研究。
通讯作者:任同群(1980-)，男，辽宁省瓦房店人，博士研究生，副教授，主要从事精密微小装配技术、精密测试技术及信号处理等方向的研究。
引用格式:徐兴辉，唐大林，顾书豪，等.基于多智能体强化学习的微装配任务规划方法[J].计算机测量与控制，2023，31(8)：217-223.

运动干涉，甚至产生硬件损坏[1-2]。除此，人工方式不能全局规划，所得运动控制方案并非最优，势必牺牲一定的装配效率。此时，微装配任务自主规划就成为了传统微装配演变成数字化智能微装配的关键手段之一[3]。

针对规划问题，典型的方法是人工势场法[4]及其衍生的算法[5-7]。其优点是收敛速度快，路径平滑，运行稳定[8]。还有学者将各类智能搜索算法如遗传算法、模拟退火以及模糊控制算法等融入人工势场法[9-11]，以此来增加算法的搜索能力，改善局部最优问题。微装配的任务由操作臂连续动作决策所构成，因此可将其归为序贯决策问题。强化学习（RL，reinforcement learning）的出现主要用于序贯决策问题，能够增加了操作臂的适应能力，且只需要操作臂在探索过程中不断地从错误中学习，即可得到稳定的路径[12]。近年来，针对大维度数据带来的迭代复杂问题，有学者将深度学习（DL，deep learning）融入 RL 中形成深度强化学习（DRL，deep reinforcement learning），如控制机械臂重新排列物体[13]，完成协作任务[14]，在非结构化区域中抓取目标对象[15]。

在装配领域，李彦江利用多智能体深度确定性策略梯度算法（MADDPG，multi-agent deep deterministic policy gradientalgorithm）算法，实现了双关节臂的协同装配并进行了仿真验证[16]。李妍等将模糊贝叶斯与深度 Q 网络结合，提高了狭窄空间中的运行效率[17]。微装配系统特有的支撑框架、直角坐标、空间密集使其对精度、空间等要求较高。而障碍物与目标点集中等特征加剧了迭代复杂、局部最优问题，因而传统的人工势场法等并不适用此类规划问题。一次任务中有不同的子任务，而各子任务的操作臂个数、障碍物布置等均不同，此类先验信息、环境信息的变化将导致使用智能搜索算法规划时需要重新构建环境，耗费了大量资源。当前 DRL 方法多用于移动机器人的路径规划[18-19]、无人机导航与任务分配[20]、以及实际生产线上的人机协作问题[21]等。然而上述工况与微装配任务特点差异较大，存在动作空间不符、奖励函数条件不合理等可能导致微装配实物训练难度大的问题，以及精确度低的仿真环境可能导致训练失败。

综上所述，提出了多智能体强化学习（MARL，multi-agent reinforcement learning）下更契合微装配任务的动作、状态、奖励条件的构建方法。同时，基于 MADDPG 算法，利用 Coppelisiam 软件对已有测量设备进行物理建模，构建了任务的深度强化学习模型并进行训练得到可工程化的路径，实现了微装配设备的自主规划。

1 多智能体强化学习算法

1.1 深度强化学习机制

深度强化学习方法将深度学习较强的感知能力与强化学习具有的决策能力相结合。主要思想是利用深度神经网络实现对原始低层特征输入的非线性变换，增强智能体的环境感知能力，并结合强化学习的探索能力，形成一种将原始环境状态输入直接映射为动作决策输出的端对端的学习方法[22]。与强化学习一样，深度强化学习框架中也由智能体（Agent）、动作、状态及奖励等要素组成。

策略（Policy），状态集（State）与动作空间（Action）：策略 π 是智能体与环境交互时的行为选择依据，$\pi(s,a)$ 智能体在状态 s 下根据 π 选择动作 a。状态空间 S 表示智能体状态信息的集合，是智能体训练过程中制定决策和获取长期收益的交互依据；动作空间 A 包含了智能体在某状态下所能做出所有可选动作的集合。$S\rightarrow A$ 表示状态空间到动作空间的映射。

奖励函数（Reward）：奖励函数 r 反映当前环境状态 $S_t\in S$ 下所选择的动作 $A_t\in A$ 对达成目标的贡献度，其将任务目标具体化和数值化，是实现操作臂与环境之间交互的载体，影响着智能体策略的选择。奖励值是否合理，决定着智能体通过训练之后选择的动作能否有利于达成目标。

价值函数（Value Function）：价值函数 $Q(s,a)$ 表示某个状态下在选择某动作时的价值，其由带有折扣因子的未来奖励组成，因此表示选择动作的潜在价值。从定义上看，价值函数就是回报的期望：

$$R_t = r_t + \lambda r_{t+1} + \lambda^2 r_{t+2} + \cdots \tag{1}$$

$$Q(s,a) = E\left[r_t + \lambda \max Q(s_{t+1},a_{t+1}) \mid S_t = s, A_t = a\right] \tag{2}$$

其中：λ 为奖励的折扣因子。在 DRL 中，使用神经网络逼近拟合价值函数。

经验回放池（ER，experience replay）：深度学习对训练数据的假设是独立同分布的，然而在 DRL 中，训练数据由高度相关的“智能体环境交互序列”组成，不符合采样数据独立性条件。智能体在探索状态空间的不同部分时是主动学习的，因此训练数据的分布是非平稳的，不符合采样数据独立同分布条件。ER 机制使 DRL 中的智能体能够记住并回放过去的经验数据，通过与最近数据混合，打破观测序列的时间相关性，有助于解决数据非独立同分布造成的问题。

1.2 MADDPG 算法

由于微装配任务规划是一个典型的多操作臂问题，在利用多智能体强化学习进行求解时，存在环境不稳定问题，当前状态优化的策略在下一个变化的环境状态中可能又无效了，这就导致不能直接使用 ER 来进行训练，因此适用于单智能体的 DRL 方法需要改进。由 Lowe 等提出的 MADDPG 算法将多智能体思想融入了基于深度强化学习的演员－评论家算法[23]，MADDPG 算法采用集中学习，分散执行的方法解决了多智能体竞争、合作或者竞争合作共存的复杂环境中存在的环境不稳定问题。集中学习指的是环境中的所有智能体的信息是全局共享，分散执行指智能体在做出决策时，仅依靠自己观测得到的环境情况进行选择合适的动作，无需其他智能体的状态或动作，因此该算法解决了之前单智能体算法只能获得自己的状态动作的问题。微装配环境中，各操作臂既有多操作臂共同完成一项工作的合作关系，同时也有不同模块间避免干涉的竞争关系。因

此在使用多智能体强化学习训练多操作臂时，应具体分析微装配任务特点设计合理的环境交互机制，对动作、状态，特别是奖励函数进行针对性设计。

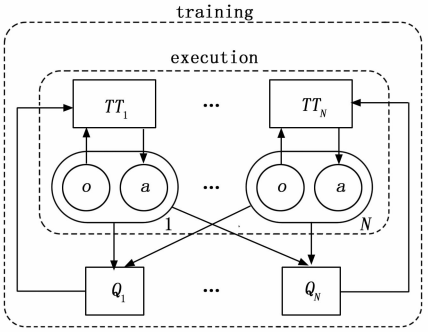


图 1　MADDPG 算法框架示意图[22]

2　微装配测量设备训练模型构建

分析微装配任务特点：其通常采用直角坐标机器人，运动方式不能脱离支撑框架，难以完成顺滑的非直线运动轨迹，因此一些高精度避障、协调的动作难以通过插补实现。同时，规避干涉时应尽量避免停顿和反向运动，以减小控制难度和装配精度损失；其次，多操作臂拥挤在狭小的操作空间内，既有协作也有竞争（运动干涉），使得规划求解空间有限；最后，操作臂的运动轨迹相对于操作空间多为长序列，且易干涉区域相对集中，使得协同运动时干涉的风险增加。动作、状态以及奖励函数构建以上述特点为依据。

2.1　状态空间与动作空间构建

状态空间的设计包括子任务划分、观测空间的建立。在微装配中，操作臂应尽量避免反向运动，但还是存在如操作臂取放工件，相机观测不同位置零件等难以避免反向运动的情况。因此可将这些装配任务中运动方向的临界位置作为各子任务的划分点，既便于进行精度控制，又便于后续动作空间设计后续路径的训练。观测空间是状态空间的基础，决定了要对哪些操作臂的状态进行观测，并将观测结果输入策略网络以及价值评价网络。在进行观测时，各模块操作臂所处位置可作为观测空间的组成部分，为了便于网络输入，对实时位置应当进行归一化处理：

$$Rel = \frac{|\ p_r - p_f\ |}{P_F} \tag{3}$$

其中：p_r 是当前操作臂位置，p_f 是子任务中操作臂需要运动的总长度。同时，对于无碰路径，还需要对实体之间的碰撞情况进行观测，最终得到的观测空间为：

$$S = \{Collision, Rle\} \tag{4}$$

动作空间决定了操作臂动作库的动作数量。动作空间的设计应具有完备、高效以及合理的特点。在进行动作空间设计时，应当对无效动作进行屏蔽。无效动作是指操作臂选择的动作违反其所处的实际环境。不同于无人机控制、关节机械臂的应用场景，微装配设备在运行时由于高精度要求，需要尽量避免停顿和反向运动的无效动作，以减少频繁启停和回程误差造成的装配精度损失，尤其在进行干涉规避时，更应避免随意反向运动等无效运动；停顿次数过多也会导致脉冲计数误差增大，增加操作臂的控制难度。因此微装配测量系统的子任务中，操作臂一旦朝向一个方向运动，原则上就不能再向反方向移动，但仍可以接受由于避碰而导致的较少次数的停顿。对于已经到达目标点的操作臂，其动作将始终为 0，因此动作空间设计为：

$$A = \begin{cases} \{v, dir\}, & \text{if} \quad \text{no treach target} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

其中：v 为操作臂所选择动作的速度，包括选择速度为 0 的停顿，dir 为操作臂的运行方向。每一个 step 开始时，都通过将操作臂当前的状态输入策略网络得到选择不同动作组合的概率，并以概率最大的动作组合作为该 step 的动作。

2.2　奖励函数设计

操作臂的训练过程就是不断试错的过程，奖励函数作为评价操作臂所选择动作的好坏评价标准，直接决定了能否训练出成功的方案。微装配任务中，各操作臂通过环境给予的奖励得知自己在某一状态下采取的动作是否合理，因此奖励函数的本质是建立状态动作对与奖励之间的映射关系，并将这种关系用于对操作臂所采取的动作的评价和约束。奖励函数直接影响了算法的收敛性，与微装配环境契合的奖励函数是实现设备无碰、高效的前提。

各操作臂的首要任务是到达目标位置，因此奖励函数应具有吸引操作臂前往目标点的功能。以操作臂所处位置为自变量，构建主要的奖励函数为：

$$R_1 = 1 - \left| \frac{\Delta p}{p_f} \right| \tag{6}$$

其中：Δp 是操作臂当前位置与目标点之间的距离。p_f 是子任务中操作臂需要运动的总长度。该函数下，随着操作臂靠近目标点，Δp 不断减少，则正向奖励值不断增大。动作空间中包含速度为 0 的停止动作，对每个子任务而言，为了保证精度以及保证操作臂的运动可控性，频繁启停将会导致精度损失，显著增加控制难度，在规划时需要避免，因此操作臂的停顿次数应当越少越好。然而在 MADDPG 算法中，为了让操作臂保证一定的探索性，各动作的选择概率具有一定的随机性，所以对停顿次数过多的行为进行惩罚：

$$R_2 = -\alpha \frac{(s_0 - s)}{s} \tag{7}$$

其中：s_0 为子任务中操作臂选择动作为 0 的 step 个数，s 为操作臂在路径中所经历的 step 个数，α 为对应的系数。由于对于路径的要求还有无碰的要求，因此当操作臂产生碰撞时需要给予较大的惩罚，即为：

$$R_3 = -a \tag{8}$$

其中：a 为一常量，表示碰撞结果一经产生，就给予操作臂固定的惩罚。然而，微装配设备中多为狭小空间，多臂协同动作时碰撞极易发生，这导致操作臂即便通过大量

冗余试验也不能明白碰撞前的路径是低价值的，此时若仅当产生碰撞时才惩罚，则会由于相关信息不直接而导致操作臂始终选择低价值动作，避碰失败，在这种情况下，操作臂需要通过大量冗余试验才能明白碰撞前的路径也是低价值的。其中，正是因为与控制目的相关的信息并未在奖励函数的设计中得到体现，而导致此类问题的产生。此时就需要使用奖励函数进行外部干预，这意味着奖励函数不仅需要对碰撞行为进行惩罚，还需要能够对碰撞进行预测。具体到微装配中，就是以实体对象之间的距离为依据，构建奖励函数为：

$$R_4 = \omega(\Delta d - D_t) \tag{9}$$

其中：ω 为系数，用于调节该部分奖励函数对整体奖励的影响程度。D_t 为碰撞检测阈值，Δd 为实体之间的实时距离，当实体距离小于 D_t 时，则 R_4 为负值，对操作臂开始惩罚，反之进行正向奖励。综上，可得符合微装配测量设备的奖励函数为

$$R = 1 - \left|\frac{\Delta p}{p_f}\right| - \alpha\frac{(s_0 - s)}{s} - a + \omega(\Delta d - D_t) \tag{10}$$

3 微装配设备仿真

3.1 物理仿真环境搭建

采用 CoppeliaSim 软件对已有的测量设备进行物理建模，基于前述奖励函数的设计方法，在 CoppeliaSim 中以 MADDPG 算法进行训练并验证。训练过程中，利用仿真软件自身提供的实体间碰撞以及距离检测功能，实现零部件之间的距离检测和碰撞检测，操作臂实时位置监测等操作，这些信息通过软件的消息机制送给训练算法，再由训练算法处理并将得到的策略通过远程控制实现仿真运动，训练环境与仿真环境关系如图 2 所示。目标设备为课题组研制的“航天陀螺仪气浮动压马达间隙测量设备”，结构如图 3 所示。设备中，由 9 自由度操作臂协同完成马达轴向及径向 μm 级间隙的测量任务[24]。

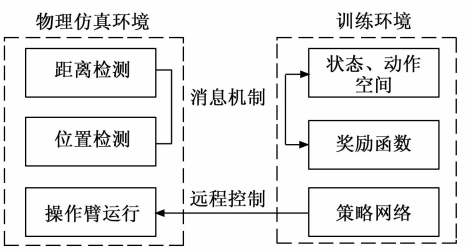


图 2 训练环境以及仿真环境关系示意图

CoppeliaSim 是一款机器人仿真平台，其具有完备的物理引擎，支持实体重构模型、距离检测、实体碰撞检测等，支持通过 Python 远程控制关节进行移动。为了节约仿真计算量，通常将复杂的模型部件进行重构，将各零部件有具有精准细节的实体转换为凸包体，因此利用开源的 V-HACD 库对操作臂组成部分进行凸分解，实现粗略建模，同时由于对夹指、工件、测头等进行较为精细的重构，使凸包体与部件形状接近一致，如图 4 所示。最终得到的物理模型如图 5 所示。

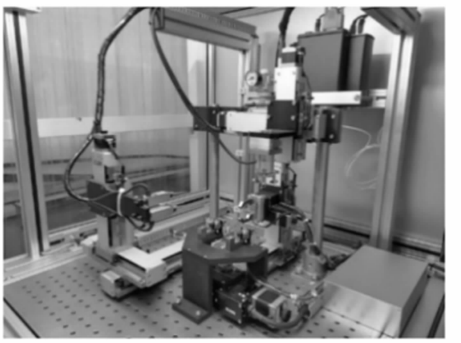
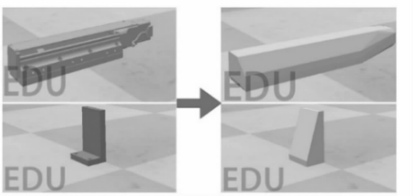
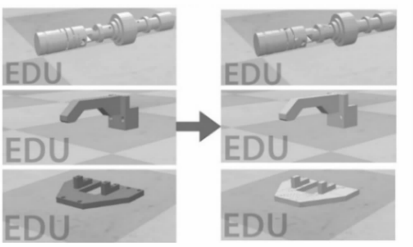


图 3 气浮轴承间隙测量设备实物图[24]



(a) 凸包体精细重构



(b) 凸包体粗略建模

图 4 凸分解构建凸包体示意图

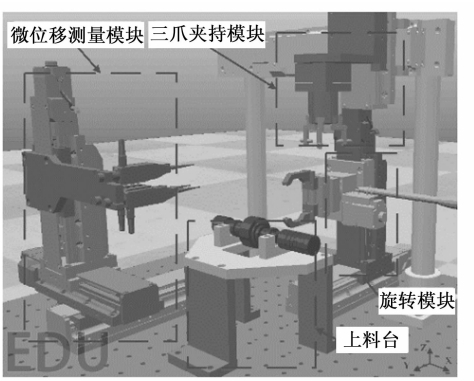


图 5 待规划设备在 CoppeliaSim 中的物理模型

虽然操作臂的运动是一个连续的动作，但连续运动时，持续性的碰撞检测将使得训练时间成本过高，如平均两个零部件之间的碰撞检测就需要 1.5 s 的时间，导致操作臂连续运动时进行训练、处理仿真软件信息流的延迟过大。因此，将操作臂的连续运动简化为离散匀速运动，即在仿真环境中以操作臂在固定时间的位移表示其在实际设备中的速度。则在间隔时间确定的情况下建立操作臂在实际设备

中与仿真模型中运动的模型映射关系为：

$$s = vT \tag{11}$$

其中：T 为确定的间隔时间，v 为操作臂运行的实际速度。Coppeliasim 机械臂控制器中的 Background Task 通过 UDP 向外部程序每隔 50 ms 循环发送当前各轴位置，确定 T 为 50 ms 的整倍数即可。基于 MADDPG 算法构建设备操作臂的训练算法。对动作、状态空间以及奖励函数进行重构，各 step 之间的间隔为 T。

3.2　仿真实验

以图 6 中的工件姿态旋转子任务为例，该子任务包含多臂协同、避障两个动作要求。旋转模块上的夹指沿 Y 轴运动并夹住工件后，操作臂的具体目标位置为：旋转模块 Y 轴执行后退动作至限位位置使零件退出上料区域，在 Y 轴退出的同时，旋转模块 R 轴旋转 90°使马达由水平转至竖直状态。同时，三爪夹持模块的 X 轴与 Z 轴也同时向目标位置运动，以准备零件的对接工作。

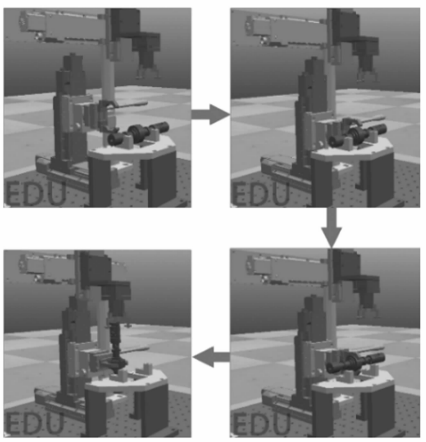


图 6　工件姿态旋转子任务实验对象

在 Pycharm 中搭建好基于 Python 的训练所需环境，然后使用基于 MADDPG 算法的训练算法对设备整体进行多个 episode 的训练。训练算法输出的策略通过封装好的函数对仿真软件中的物理模型进行远程控制。物理仿真环境中实时交互产生的操作臂位置、零部件间的距离、碰撞检测结果等作为观测值返回到训练算法。

首先构建好训练所需的动作，状态空间。然后根据式（6）设计仅以目标点吸引作为正向奖励的奖励函数，通过多次训练，得到两个操作臂位移变化曲线、动作选择曲线以及所经历的各位置处的奖励值如图 7 所示。

虽然位移的实际运动曲线显示操作臂成功避障并到达了目标位置，但由动作选择曲线可看出，操作臂运行时存在断断续续反复启停的问题，选择了过多无必要的零动作，无法保证连续稳定的运动，这违反了微装配任务规划的准则，如此频繁的启停增加了控制难度，不能契合微装配的操作特点，使得训练结果无法实际工程化。同时，该路径花费时长较大，使用了近 100 多个 step，观察实时奖励值可看到当操作臂选择了朝向目标位置的运动动作时，环境给

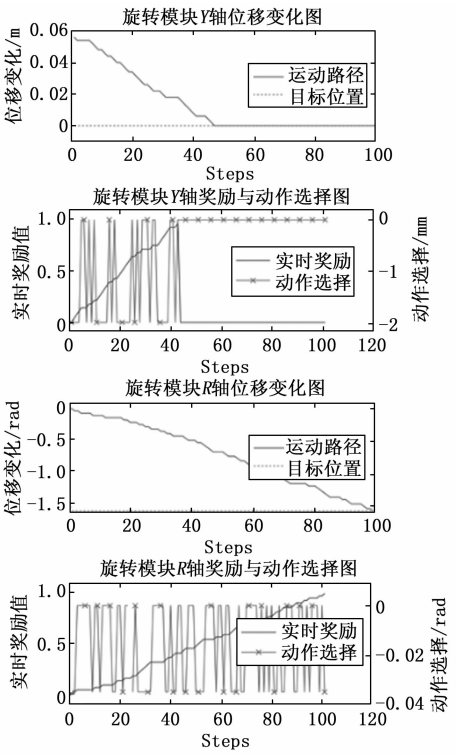


图 7　仅以目标点吸引的正向奖励为奖励函数

予一定的奖励值，然而当操作臂选择停止时，并无任何奖惩。因此，根据式（7）增加对停顿次数有约束效果的奖励函数，再次训练，可得到相关曲线如图 8 所示。

可以看到，动作选择曲线中再未出现操作臂频繁启停的现象，但由于缺少对碰撞的约束，操作臂始终在刚开始就产生碰撞，从图中可看出操作臂在第 5 个 step 就产生了碰撞。因此依据式（8）增加碰撞约束的奖励函数得到图 9 路径曲线，经过训练后可以由运动曲线图和动作选择曲线看到，旋转模块 R 轴的积极性被打消，不再进行任何的旋转动作，这是由于操作臂运动初期在狭窄的空间中较易产生碰撞，进而很容易就得到很大的惩罚，导致探索难度增大，这种惩罚使得操作臂产生“积极性丧失”的现象。为了降低操作臂初期的探索难度，根据式（10）对奖励函数进行改进，提升对碰撞情况的预测能力，得到各操作臂的相关曲线如图 10 所示。可以看到，旋转模块 R 轴在运动初期静止不动，直到 Y 轴操作臂运动了一定的距离后 R 轴的旋转运动才开始，最终实现避障，且为了避碰而产生的停顿集中，路径效率较高，在目标点吸引的正向奖励中做出更加有利于路径整体规划的动作选择。

可以看出在该稳定路径下的一个 episode 中，当操作臂未到达目标位置时，奖励值随 step 的增加而稳定增加。从动作选择曲线中可看出，受奖励函数约束，动作选择能够保持稳定，符合微装配设备操作臂运行特点。

4　结束语

分析了微装配的任务特点，提出了微装配任务动作、

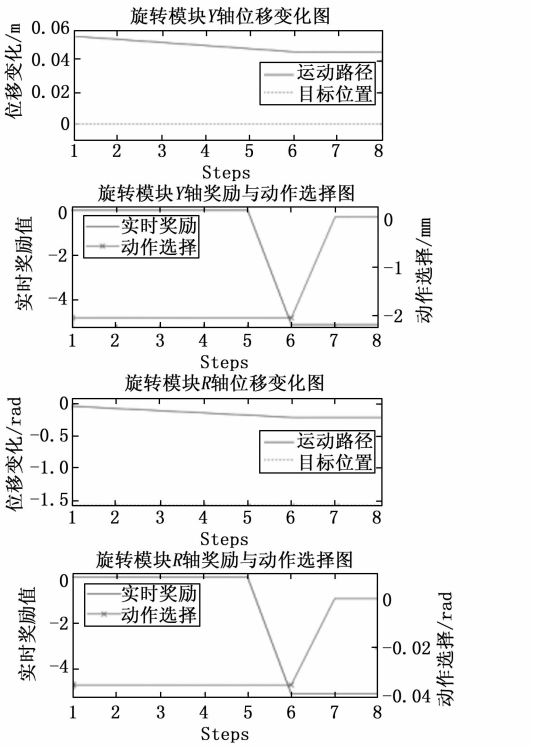


图 8 增加停顿次数奖惩的路径曲线

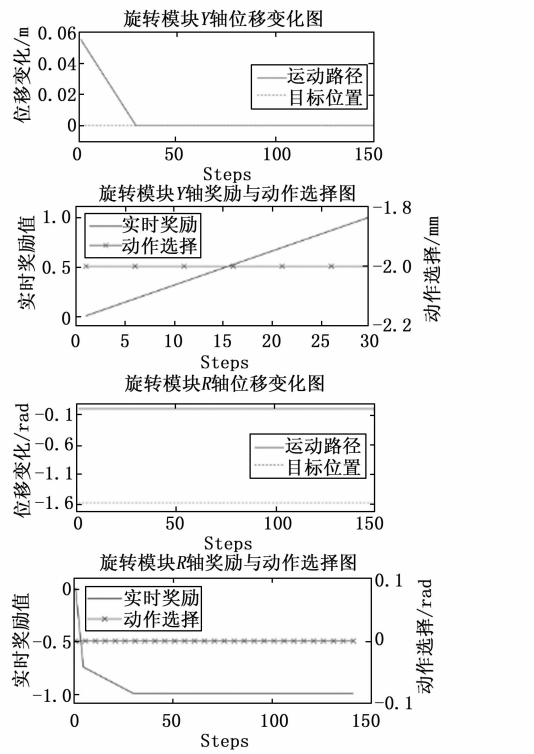


图 9 增加碰撞约束奖励函数

状态、奖励条件设计准则以及构建方法，使得多智能体深度强化学习与直角坐标下的微装配任务更契合，克服了现有环境不符合微装配特点的问题。在仿真软件中对已有的

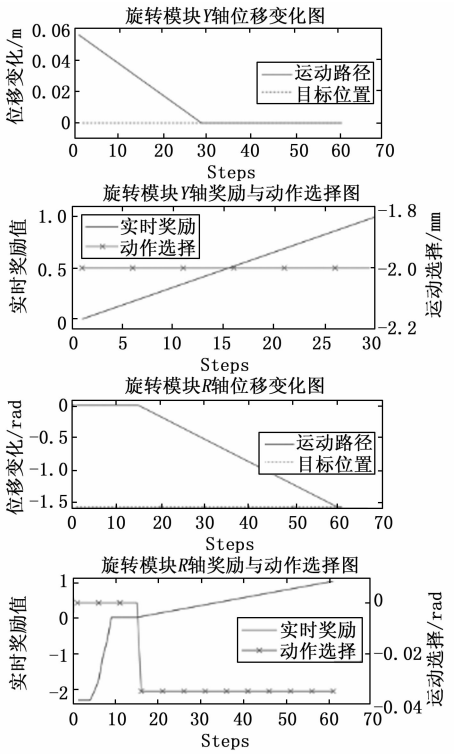


图 10 根据式（10）训练后的曲线图

测量设备各零部件进行了凸分解。最后，基于 MADDPG 算法模型进行训练，得到了完整的测量方案，并通过试验证明设计的奖励函数能够使得路径更加符合微装配的实际工况，为微装配自主规划提供了理论、技术支撑。

参考文献：

[1] 于忠洋．微小磁性零件的自动化特征定位与装配控制［D］．大连：大连理工大学，2022.

[2] 王乾州．磁钢组件自动装配设备及自动组装夹具设计［D］．大连：大连理工大学，2021.

[3] 刘 翊．刚柔混合产品的装配工艺规划技术与应用研究［D］．武汉：华中科技大学，2014.

[4] PAN ZH，LI DF，YANG K，et al. Multi-robot obstacle avoidance based on the improved artificial potential field and PID adaptive tracking control algorithm［J］. Robotica，2019，37（11）：1883－1903.

[5] FAN XJ，GUO YJ，LIU H，et al. Improved artificial potential field method applied for AUV path planning［J］. Mathematical Problems in Engineering，2020，4：1155－1175.

[6] 马庆禄，黄光浩．基于改进人工势场法的自动驾驶路径规划方法［J］．计算机仿真，2022，39（8）：160－165.

[7] MILAD N，ESMAEEL K，SAMIRA D. Multi-objective multi-robot path planning in continuous environment using an enhanced genetic algorithm［J］. Expert Systems with Applications，2019，115：106－120.

[8] 张殿富，刘 福．基于人工势场法的路径规划方法研究及展望［J］．计算机工程与科学，2013，35（6）：88－95.

[9] OROZCO-ROSAS U，MONTIEL O，ROBERTO S. Mobile ro-

bot path planning using membrane evolutionary artificial potential field [J]. Applied Soft Computing, 2019, 77: 236 - 251.

[10] 朱　颖，李元鹏，张亚婉，等. 基于改进人工势场法的搬运机器人路径规划 [J]. 电子测量技术，2020，43 (17)：101 - 104.

[11] SONG Q, LIU L. Mobile robot path planning based on dynamic fuzzy artificial potential field method [J]. Journal of Information and Computational Science, 2012, 9 (17): 5233 - 5240.

[12] SUTTON RS, BARTO AG. Reinforcement learning: an introduction [J]. IEEE Transactions on Neural Networks, 1998, 9 (5): 1054.

[13] YUAN W, STORKJA, KRAGIC D, et al. Rearrangement with nonprehensile manipulation using deep reinforcement leaning [C] //2018 International Conference on Robotics and Automation, 2018, 270 - 277.

[14] 刘钱源. 基于深度强化学习的双臂机器人物体抓取 [D]. 济南：山东大学，2019.

[15] IRIONDOA, LAZKANOE, SUSPERREGIL, et al. Pick and place operations in logistics using a mobile manipulator controlled with deep reinforcement learning [J]. Applied Sciences - basel, 2019, 9 (2): 348.

[16] 李彦江. 一种基于多智能体强化学习的智能装配系统 [D]. 北京：北京邮电大学，2020.

[17] 李　妍，甄成刚. 基于深度 Q 网络的虚拟装配路径规划 [J]. 计算机工程与设计，2019，40 (7)：2032 - 2038.

[18] 张柏鑫，杨毅镔，朱华中，等. 基于深度强化学习的移动机器人动态路径规划算法 [J]. 计算机测量与控制，2023，31 (1)：153 - 159，166.

[19] 张浩杰，苏治宝，苏　波. 基于深度 Q 网络学习的机器人端到端控制方法 [J]. 仪器仪表学报，2018，39 (10)：36 - 43.

[20] 张　堃，李　珂，时昊天，等. 基于深度强化学习的 UAV 航路自主引导机动控制决策算法 [J]. 系统工程与电子技术，2020，42 (7)：1567 - 1574.

[21] 金哲豪，刘安东，俞　立. 基于 GPR 和深度强化学习的分层人机协作控制 [J]. 自动化学报，2022，48 (9)：2352 - 2360.

[22] 赵立阳，常天庆，褚凯轩，等. 完全合作类多智能体深度强化学习综述 [J]. 计算机工程与应用，2023，59 (12)：14 - 27.

[23] LOWE R, WU Y, TAMAR A, et al. Multi - agent actor - critic for mixed cooperative - competitive environments [C] // 31st Conference on Neural Information Processing Systems, 2017.

[24] 王晓飞. 气浮陀螺仪轴承间隙测量系统研制 [D]. 大连：大连理工大学，2021.

(上接第 196 页)

[4] 仝兆景，郑　权，韩耀飞，等. 基于新滑模观测器的永磁同步电机无传感器控制 [J]. 电子科技，2021，34 (12)：1 - 5.

[5] 崔　波，方玲利，蒋　全，等. 表贴式永磁同步电机高速无位置传感器控制技术比较研究 [J]. 电子科技，2020，34 (10)：8 - 15.

[6] JEONG Y S, LORENZ R D, JAHNS T M, et al. Initial rotor position estimation of an interior permanent magnet synchronous machine using carrier-frequency injection methods [J]. IEEE Transaction on Industry Applications, 2005, 41 (1): 38 - 44.

[7] 陈小玲. 基于高频注入法的永磁同步电机转子磁极位置估计 [D]. 成都：电子科技大学，2019.

[8] XUN Q, WANG P L, CAI Z D, et al. Hall rotor position estimation method and its error compensation [J]. Transactions of China Electrotechnical Society, 2017, 36 (6): 145 - 155.

[9] SCELBA G, DEDONATO G, PULVIRENTI M, et al. Hall-effect sensor fault detection, identification, and compensation in brush-less DC drives [J]. IEEE Transactions on Industry Applications, 2016, 52 (2): 1542 - 1554.

[10] 王　凯，王之赍，宗兆伦，等. 基于霍尔位置传感器的永磁同步电机速度估计方法研究 [J]. 电机与控制学报，2019，23 (7)：4 - 6.

[11] XU Z, LI T C, LU Y P, et al. Position - measuring error analysis and solution of hall sensor in pseudo - sensor less PMSM driving system [J]. IEEE Conf. on Industrial Electronics Society, 2003 (2): 1337 - 1342.

[12] 姜铁征，万秋华，于　海，等. 小型绝对式光电编码器精度自动检测装备 [J]. 仪表技术与传感器，2019，3：1 - 5.

[13] 王建鹏. 旋转变压器在伺服系统中的应用 [J]. 数字化用户，2018，24 (35)：162 - 163.

[14] 索晓杰，马小博，周　勇. 正余弦旋转变压器与线性旋转变压器的对比分析 [J]. 信息通信，2019 (2)：14 - 16.

[15] 邱美涵，王晓琳，卞　皓. 基于 DSA/DC 的旋变位置解码系统设计与研究 [J]. 微特电机，2019，47 (10)：6 - 9.

[16] KHABURI, ARAB D. Software-based resolver-to digital converter for DSP-based drives using an improved angle-tracking observer [J]. IEEE Transactions on Instrumentation and Measurement (S0018-9456), 2012, 61 (4): 922 - 929.

[17] 桑徐阳. 基于旋转变压器位置解码的开关磁阻电机控制系统设计 [D]. 杭州：浙江大学，2019.

[18] 刘继磊，杨　毅，高志民. 基于新型磁阻式旋转变压器解码问题研究 [J]. 驱动控制，2021，49 (7)：50 - 56.

[19] 杨瑞峰，张伟鹏，郭晨霞，等. 旋转变压器误差抑制与解码技术的研究 [J]. 微电机，2020，53 (2)：56 - 60.

[20] 李　明，安书董，段宇博. 一种基于 A/D2S1210 的旋转变压器位置解码及监控方法 [J]. 电子与通信技术，2021 (4)：138 - 140.

[21] 陈　亮，王秋瑶. 双通道旋转变压器解码系统设计 [J]. 光电技术应用，2016，31 (3)：50 - 53.

[22] 庞岳峰，陈建友，樊全鑫，等. 双通道旋转变压器解码算法改进 [J]. 电子科技，2019，32 (8)：51 - 55，65.

[23] 王志宏，宦　昱，俞　华. 双通道旋转变压器接线模式分析 [J]. 机电工程技术，2020，49 (7)：232 - 234.