DOI：10.13973/j.cnki.robot.230148

基于行为的多差速机器人强化学习任务监管器设计

张祯毅[1,2]，黄　捷[1,2]

（1. 福州大学电气工程与自动化学院，福建 福州　350108；　2. 福州大学 5G+工业互联网研究院，福建 福州　350108）

摘　要：针对多差速机器人系统提出了一种基于试错学习的多智能体强化学习任务监管器。此方法解决了基于行为的多智能体系统总是依赖人的智能设计切换规则以决策行为优先级的问题。首先，在零空间行为控制框架下引入了差速模型代替质点模型，首次推导了具有非完整约束的零空间行为控制范式，从而提升了系统对最小极值状态的鲁棒性。然后，首次将行为优先级切换问题建模为协作式马尔可夫博弈问题，学习了一个最优的联合策略以动态且智能地决策行为优先级，不仅避免了人工设计切换规则，而且降低了在线计算和存储负担。仿真结果显示，所提出多智能体强化学习任务监管器具有优越的行为优先级切换性能。在 AgileX Limo 系列多差速机器人系统上的成功应用，验证了该任务监管器的实用性。

关键词：差速机器人；行为控制；强化学习；任务监管器；智能决策

Reinforcement Learning Mission Supervisor Design for Behavior-based Differential Drive Robots

ZHANG Zhenyi[1,2]，HUANG Jie[1,2]

(1. *School of Electrical Engineering and Automation*, *Fuzhou University*, *Fuzhou* 350108, *China*;
2. 5G+ *Industrial Internet Institute of Fuzhou University*, *Fuzhou* 350108, *China*)

Abstract: A multi-agent reinforcement learning mission supervisor (MARLMS) is designed for differential drive robots using trial-and-error learning. The proposed MARLMS addresses the challenge inherent in behavior-based multi-agent systems, wherein the design of switching rules to determine behavior priorities relies heavily on human intelligence. Building upon the null-space-based behavioral control (NSBC) framework, a differential model is introduced to replace the particle model. Consequently, a paradigm of NSBC with nonholonomic constraints is presented for the first time, enhancing the system robustness to the minimum extremum state. Subsequently, a joint policy is developed to dynamically and intelligently determine behavior priorities by modeling the behavior priority switching problem as a cooperative Markov game. The proposed MARLMS not only eliminates the need for manual design of switching rules but also reduces the computational and storage burdens during online operations. Simulation results demonstrate the superior behavior priority switching performance of the proposed MARLMS. Furthermore, successful implementation on AgileX Limo robots validates the practicality of the proposed MARLMS.

Keywords: differential drive robot; behavioral control; reinforcement learning; mission supervisor; intelligent decision

多差速机器人系统通过协同能够完成个体无法独立执行的任务，因此已广泛应用于物流、交通和巡检等场景[1-4]。随着工作环境的动态化和任务需求的高性能化，任务目标变得更为复杂，且多差速机器人系统不得不同时完成多个相互冲突的任务，包括局部任务和全局任务[5-6]。局部任务是指个体独立完成的任务，而全局任务是指群体协同完成的任务[7]。这种多任务冲突问题是多智能体领域的热点问题之一[8-9]。

行为控制方法是 Brooks 首次提出的[10]，通过建模和融合多个行为来解决多任务冲突问题。文 [11] 总结了几类典型的行为选择机制，包括分层、加权、模糊和零空间等。文 [12] 提出了一种分层行为控制方法，采用竞争式架构，完整执行最高层次行为，但任务执行效率低。文 [13] 和文 [14] 分别提出了加权和模糊行为控制方法，均采用协作式架构，充分利用系统冗余度执行各种行为，但每个行为都未得到完整执行。结合竞争式与协作式架构的优点，Antonelli 等[15] 提出了一种新颖的基于零空间的行为控制（NSBC）框架，不仅能完整执行最

基金项目：国家自然科学基金（92367109）.
通信作者：黄捷，jie.huang@fzu.edu.cn　收稿/录用/修回：2023-06-03/2024-02-07/2024-02-22

高优先级的行为，而且可以通过零空间执行部分低优先级行为。为了完成零空间的投影，NSBC 基于任务监管器（mission supervisor）分配行为的优先级。然而，行为优先级最初是人工提前设定的且固定不变，因而该方法执行任务时的动态性能不佳。

为克服固定行为优先级的缺陷，学者们相继提出了有限状态自动机任务监管器（FSAMS）[16-21]、模糊任务监管器（FMS）[22] 和模型预测控制任务监管器（MPCMS）[23-24]。FSAMS 将每个行为优先级隐含在一个有限状态机的状态中，通过人工设计数值化的状态转移条件，以状态转移的方式切换行为优先级，但数值条件依赖人工设计且缺乏理论依据。FMS 使用模糊逻辑表代替数值逻辑规则，大幅度降低了人工设计规则的难度，但需要人工设计模糊集合和模糊规则等。MPCMS 将行为优先级切换问题建模为一个最优模式切换问题，通过实时求解最优行为优先级来避免人工设计切换规则，但在线的计算量和存储负担非常大，且实时性不佳。

文 [25-26] 将行为优先级切换问题建模为一个序贯决策问题，提出了一种新颖的强化学习任务监管器（RLMS）。RLMS 学习一个最优的行为优先级策略，不仅避免了人工设计规则，而且降低了硬件负担。然而，RLMS 通常无法在多差速机器人系统中取得理想的任务性能，具体原因如下：1) RLMS 使用质点模型建模智能体的运动学，但差速机器人系统受到非完整约束的限制，其运动学不满足质点模型。若使用质点模型表征差速机器人系统，则参考指令难以跟踪，且多差速机器人系统易陷入极值状态。2) RLMS 只在单个学习者的环境下可保证收敛，扩展至多差速机器人系统中时，存在学习环境非平稳问题，从而强化学习算法将失去收敛保证。3) RLMS 只能实施局部行为，无法发挥多差速机器人系统的群体智能，因而降低了任务执行能力。

针对上述问题，本文提出了一种新颖的多智能体强化学习任务监管器（MARLMS），解决了多差速机器人系统的行为优先级决策问题。具体而言，MARLMS 设计的难点在于如何构建多差速机器人行为控制方法与多智能体强化学习算法之间的“桥梁”。此外，MARLMS 的行为集合将包含局部行为和全局行为，且必须克服学习环境的非平稳问题，达到群体效益最大化而非个体效益最大化。由于 NSBC 框架在任务层通常是集中式的，因此本文考虑将多差速机器人的行为优先级切换问题建模为协作式马尔可夫博弈问题，联合差速机器人的状态和行为，以最大化团队奖励为目标，学习一个最优的联合行为优先级策略。一方面，MARLMS 减少了对人工设计行为优先级切换规则的依赖，且降低了硬件平台实时计算和存储行为优先级的负担；另一方面，MARLMS 弥补了RLMS 不能实施全局行为的致命缺陷，且解决了多差速机器人学习环境的非平稳问题，从而极大地提升了 RLMS 的可扩展性。

1 建模与问题描述（Modeling and problem statement）

1.1 多差速机器人系统运动学模型

在由 N（$N>2$）个差速机器人组成的多差速机器人系统中，每个差速机器人均具有 2 个辅助轮和 2 个驱动轮，且第 i 个差速机器人的结构示意图如图 1 所示，$i=1,2,\cdots,N$。

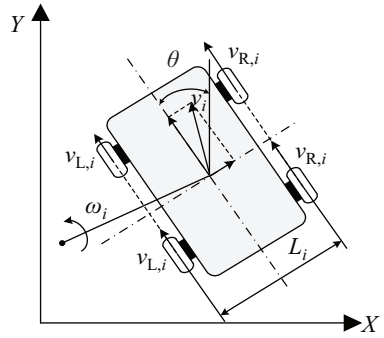


图 1 第 i 个差速机器人的结构示意图

Fig.1 The schematic diagram of the i-th differential drive robot

第 i 个差速机器人的线速度 $v_i\in\mathbb{R}$ 和角速度 $\omega_i\in\mathbb{R}$ 分别可表示为 [27]

$$v_i=(v_{\mathrm{L},i}+v_{\mathrm{R},i})/2 \tag{1}$$

$$\omega_i=(v_{\mathrm{L},i}-v_{\mathrm{R},i})/L_i \tag{2}$$

其中，$v_{\mathrm{L},i}\in\mathbb{R}$ 和 $v_{\mathrm{R},i}\in\mathbb{R}$ 分别是左右驱动轮的速度，$L_i\in\mathbb{R}$ 是左右驱动轮间的距离，$\mathbb{R}$ 为实数集合。

定义第 i 个差速机器人的位置和偏航角分别为 $\boldsymbol{p}_i=[x_i,y_i]\in\mathbb{R}^2$ 和 $\theta_i\in\mathbb{R}$，则第 i 个差速机器人的运动学方程可以建模为 [28]

$$\dot{\boldsymbol{X}}_i=\begin{bmatrix}\dot{\boldsymbol{p}}_i\\ \dot{\theta}_i\end{bmatrix}=\begin{bmatrix}v_i\cos\theta_i\\ v_i\sin\theta_i\\ \omega_i\end{bmatrix}=\begin{bmatrix}\cos\theta_i & 0\\ \sin\theta_i & 0\\ 0 & 1\end{bmatrix}\boldsymbol{V}_i=\boldsymbol{\Theta}_i\boldsymbol{V}_i \tag{3}$$

其中，$\boldsymbol{X}_i=[\boldsymbol{p}_i,\theta_i]^{\mathrm{T}}\in\mathbb{R}^3$ 和 $\boldsymbol{V}_i=[v_i,\omega_i]^{\mathrm{T}}\in\mathbb{R}^2$ 分别是第 i 个差速机器人的广义位置和速度，$\boldsymbol{\Theta}_i\in\mathbb{R}^{3\times2}$ 是非完整约束矩阵。

假设 1：多差速机器人系统工作在一个静态的场景中，其中所有障碍物都是静态且固定的。

1.2 任务目标

多差速机器人系统的运动学方程如式 (3) 所示，其任务目标是学习一个联合的行为优先级策略，以在满足假设 1 的工作环境中动态且智能地决策它们的行为优先级，从而使得多差速机器人系统在避开障碍物的同时形成或重构队形。

2 具有非完整约束的 NSBC 范式设计（Paradigm design of NSBC with nonholonomic constraint）

2.1 基本行为设计

假设每个差速机器人均有 M 个基本行为，其中第 i 个差速机器人的第 j 个基本行为可采用一个任务变量 $\boldsymbol{\sigma}_{i,j}\in\mathbb{R}^{m_j}$ （$m_j\leqslant 3$，$j=1,\cdots,M$）来表示：

$$\boldsymbol{\sigma}_{i,j}=\boldsymbol{g}_{i,j}(\boldsymbol{X}_i) \tag{4}$$

其中，$\boldsymbol{g}_{i,j}(\cdot):\mathbb{R}^3\to\mathbb{R}^{m_j}$ 为任务函数。

然后，任务变量 $\boldsymbol{\sigma}_{i,j}$ 的微分形式推导为

$$\dot{\boldsymbol{\sigma}}_{i,j}=\frac{\partial \boldsymbol{g}_{i,j}(\boldsymbol{X}_i)}{\partial \boldsymbol{X}_i}\dot{\boldsymbol{X}}_i=\boldsymbol{J}_{i,j}\dot{\boldsymbol{X}}_i=\boldsymbol{J}_{i,j}\boldsymbol{\Theta}_i\boldsymbol{V}_i \tag{5}$$

其中，$\boldsymbol{J}_{i,j}\in\mathbb{R}^{m_j\times 3}$ 表示任务的雅可比矩阵。

最后，基于闭环逆运动学方法[11]，第 i 个差速机器人的第 j 个基本行为的参考速度指令推导为

$$\boldsymbol{V}_{i,j}=\boldsymbol{\Theta}_{i,j}^{\ddagger}\boldsymbol{J}_{i,j}^{\dagger}(\dot{\boldsymbol{\sigma}}_{\mathrm{d},i,j}+\boldsymbol{\Lambda}_{i,j}\tilde{\boldsymbol{\sigma}}_{i,j}) \tag{6}$$

其中，$\boldsymbol{\Theta}_{i,j}^{\ddagger}=\boldsymbol{\Theta}_{i,j}^{\mathrm{T}}(\boldsymbol{\Theta}_{i,j}\boldsymbol{\Theta}_{i,j}^{\mathrm{T}})^{-1}\in\mathbb{R}^{2\times 3}$ 表示 $\boldsymbol{\Theta}_{i,j}$ 的左伪逆矩阵，$\boldsymbol{J}_{i,j}^{\dagger}=\boldsymbol{J}_{i,j}^{\mathrm{T}}(\boldsymbol{J}_{i,j}\boldsymbol{J}_{i,j}^{\mathrm{T}})^{-1}\in\mathbb{R}^{3\times m_j}$ 表示 $\boldsymbol{J}_{i,j}$ 的右伪逆矩阵，$\boldsymbol{\sigma}_{\mathrm{d},i,j}\in\mathbb{R}^{m_j}$ 是期望的任务函数，$\dot{\boldsymbol{\sigma}}_{\mathrm{d},i,j}\in\mathbb{R}^{m_j}$ 是 $\boldsymbol{\sigma}_{\mathrm{d},i,j}$ 的微分形式，$\boldsymbol{\Lambda}_{i,j}\in\mathbb{R}^{m_j\times m_j}$ 是任务的增益矩阵，$\tilde{\boldsymbol{\sigma}}_{i,j}=\boldsymbol{\sigma}_{\mathrm{d},i,j}-\boldsymbol{\sigma}_{i,j}\in\mathbb{R}^{m_j}$ 是任务的误差。

在不失一般性的前提下，编队保持、重构和避障行为设计如下：

编队保持行为（FM）：是一个全局行为，旨在驱使多差速机器人系统形成一个期望的队形，相应的任务函数、期望任务和任务雅可比矩阵可分别表示为

$$\boldsymbol{\sigma}_{\mathrm{FM},i}=\left[(\boldsymbol{p}_i-\boldsymbol{p}_{\mathrm{c}}-\boldsymbol{p}_i^{\mathrm{c}})^{\mathrm{T}}\ \ \theta_i\right]^{\mathrm{T}}\in\mathbb{R}^3 \tag{7}$$

$$\boldsymbol{\sigma}_{\mathrm{FM,d},i}=\left[(\boldsymbol{p}_{\mathrm{c,d}}-\boldsymbol{p}_{\mathrm{c}})^{\mathrm{T}}\ \ \theta_{\mathrm{d}}\right]^{\mathrm{T}}\in\mathbb{R}^3 \tag{8}$$

$$\boldsymbol{J}_{\mathrm{FM},i}=\begin{bmatrix}\dfrac{N-1}{N}\boldsymbol{I}_2 & \boldsymbol{0}_{2\times 1}\\ \boldsymbol{0}_{1\times 2} & 1\end{bmatrix}\in\mathbb{R}^{3\times 3} \tag{9}$$

其中，$\boldsymbol{p}_{\mathrm{c}}=\sum_{i=1}^{N}\boldsymbol{p}_i\in\mathbb{R}^2$ 是编队质心的位置，$\boldsymbol{p}_i^{\mathrm{c}}\in\mathbb{R}^2$ 是编队质心与第 i 个差速机器人的相对位置，$\boldsymbol{p}_{\mathrm{c,d}}\in\mathbb{R}^2$ 是编队质心的期望位置，$\theta_{\mathrm{d}}=\arctan\|\boldsymbol{p}_{\mathrm{c,d}}-\boldsymbol{p}_{\mathrm{c}}\|\in\mathbb{R}$ 是编队的期望方向，$\boldsymbol{I}$ 表示单位矩阵，$\boldsymbol{0}$ 表示零矩阵。

编队重构行为（FR）：类似于编队保持行为，亦是全局行为，旨在驱使多差速机器人系统重构一个期望的队形，相应的任务函数、期望任务和任务雅可比矩阵分别表示为

$$\boldsymbol{\sigma}_{\mathrm{FR},i}=\left[(\boldsymbol{p}_i-\boldsymbol{p}_{\mathrm{c}}-\boldsymbol{\Gamma}_{\mathrm{FR},i}\boldsymbol{p}_i^{\mathrm{c}})^{\mathrm{T}}\ \ \theta_i\right]^{\mathrm{T}}\in\mathbb{R}^3 \tag{10}$$

$$\boldsymbol{\sigma}_{\mathrm{FR,d},i}=\left[(\boldsymbol{p}_{\mathrm{c,d}}-\boldsymbol{p}_{\mathrm{c}})^{\mathrm{T}}\ \ \theta_{\mathrm{d}}\right]^{\mathrm{T}}\in\mathbb{R}^3 \tag{11}$$

$$\boldsymbol{J}_{\mathrm{FR},i}=\begin{bmatrix}\dfrac{N-1}{N}\boldsymbol{I}_2 & \boldsymbol{0}_{2\times 1}\\ \boldsymbol{0}_{1\times 2} & 1\end{bmatrix}\in\mathbb{R}^{3\times 3} \tag{12}$$

其中，$\boldsymbol{\Gamma}_{\mathrm{FR},i}\in\mathbb{R}^{2\times 2}$ 是第 i 个差速机器人的编队重构矩阵。

避障行为（OA）：避障行为是一个局部行为，旨在驱使多差速机器人系统避开路径附近的障碍物，相应的任务函数、期望任务和任务雅可比矩阵分别表示为

$$\boldsymbol{\sigma}_{\mathrm{OA},i}=\left[\min\{d_i^{\mathrm{o}}\}\ \ \theta_i\right]^{\mathrm{T}}\in\mathbb{R}^2 \tag{13}$$

$$\boldsymbol{\sigma}_{\mathrm{OA,d},i}=\left[d_{\mathrm{OA}}\ \ \theta_{\mathrm{OA},i}\right]^{\mathrm{T}}\in\mathbb{R}^2 \tag{14}$$

$$\boldsymbol{J}_{\mathrm{OA},i}=\begin{bmatrix}\boldsymbol{\Gamma}_{\mathrm{OA},i}^{\mathrm{T}} & 0\\ \boldsymbol{0}_{1\times 2} & 1\end{bmatrix}\in\mathbb{R}^{2\times 3} \tag{15}$$

其中，$\min\{d_i^{\mathrm{o}}\}\in\mathbb{R}$ 是第 i 个差速机器人与障碍物的最小距离，$d_{\mathrm{OA}}\in\mathbb{R}$ 是避障的安全距离，$\theta_{\mathrm{OA},i}=\arctan\|\boldsymbol{p}_{i,\min}^{\mathrm{o}}\|\pm\dfrac{\pi}{2}\in\mathbb{R}$ 是避障的期望方向，$+$ 和 $-$ 分别表示障碍物在第 i 个差速机器人的左侧和右侧，$\boldsymbol{\Gamma}_{\mathrm{OA},i}=\dfrac{\boldsymbol{p}_{i,\min}^{\mathrm{o}}}{\min\{d_i^{\mathrm{o}}\}}\in\mathbb{R}^{1\times 2}$，$\boldsymbol{p}_{i,\min}^{\mathrm{o}}\in\mathbb{R}^2$ 是第 i 个差速机器人与障碍物之间最小距离对应的位置向量差。

2.2 复合行为设计

复合行为是多个基本行为按照一定行为优先级顺序零空间投影的组合。定义 $\hat{j}\in N_M$ 表示行为优先级顺序，$N_M=\{1,\cdots,M\}$。定义一个与时间相关的行为优先级顺序函数 $h_i(\hat{j},t):N_M\times[0,\infty]\to N_M$。此外，基本行为满足行为分层规则如下：

1) 一个具有 $h_i(\hat{j}_\alpha,t)$ 行为优先级的基本行为不能干扰一个具有 $h_i(\hat{j}_\beta,t)$ 行为优先级的基本行为，如果 $h_i(\hat{j}_\alpha,t)\geqslant h_i(\hat{j}_\beta,t)$，对于 $\forall\hat{j}_\alpha,\hat{j}_\beta\in N_M$，$\hat{j}_\alpha\neq\hat{j}_\beta$。

2) 系统速度到任务速度的映射关系可由任务雅可比矩阵 $\boldsymbol{J}_{h_i(\hat{j},t)}\in\mathbb{R}^{m_j\times n}$ 表示。

3) 具有最低行为优先级的基本行为维度 m_M 可能大于 $m_{\mathrm{total}}-\sum_{j=1}^{M-1}m_j$，因此期望维度 m_{total} 大于所有基本行为的总维度。

4) $h_i(\hat{j},t)$ 的数值由任务监管器根据任务需求和传感器信息进行决策与分配。

在每个采样周期，一旦基本行为的行为优先级确定和分配完成，复合行为的参考速度指令可通过一个递归方案计算：

$$\boldsymbol{V}_{\mathrm{r},i}=\boldsymbol{\Theta}_i^{\ddagger}\left(\boldsymbol{X}_{i,1}+\sum_{\hat{j}=2}^{M}\bar{\boldsymbol{J}}_{i,1,\hat{j}-1}\boldsymbol{X}_{i,\hat{j}}\right) \tag{16}$$

$$\bar{\boldsymbol{J}}_{i,1,\hat{j}-1}=\boldsymbol{I}_3-\boldsymbol{J}_{i,1,\hat{j}}^{\dagger}\boldsymbol{J}_{i,1,\hat{j}} \tag{17}$$

$$\boldsymbol{J}_{i,1,\hat{j}}=[\boldsymbol{J}_{i,1}^{\mathrm{T}},\boldsymbol{J}_{i,2}^{\mathrm{T}},\cdots,\boldsymbol{J}_{i,\hat{j}}^{\mathrm{T}}]^{\mathrm{T}} \tag{18}$$

其中，下标 $\hat{j}$ 是行为优先级顺序，$\bar{\boldsymbol{J}}_{i,1,\hat{j}-1}\in\mathbb{R}^{3\times3}$ 是增广雅可比矩阵的零空间投影算子，$\boldsymbol{J}_{i,1,\hat{j}}\in\mathbb{R}^{\left(\sum_{t=1}^{\hat{k}}m_t\right)\times3}$ 表示增广雅可比矩阵。

经典 NSBC 方法使用质点模型，形如 $\dot{\boldsymbol{X}}_i=\boldsymbol{V}_i$，其中 $\boldsymbol{X}_i=[p_{x,i},p_{y,i},\theta_i]^{\mathrm{T}}\in\mathbb{R}^3$ 表示广义位置，$\boldsymbol{V}_i=[v_{x,i},v_{y,i},\dot{\theta}_i]^{\mathrm{T}}\in\mathbb{R}^3$ 表示广义速度，且它们的维度是相同的[11]。基于质点模型，经典 NSBC 方法的基本行为和复合行为分别计算如下：

$$\boldsymbol{V}_{i,j}=\boldsymbol{J}_{i,j}^{\dagger}(\dot{\boldsymbol{\sigma}}_{\mathrm{d},i,j}+\boldsymbol{\Lambda}_{i,j}\tilde{\boldsymbol{\sigma}}_{i,j}) \tag{19}$$

$$\boldsymbol{V}_{\mathrm{r},i}=\boldsymbol{X}_{i,1}+\sum_{\hat{j}=2}^{M}\bar{\boldsymbol{J}}_{i,1,\hat{j}-1}\boldsymbol{X}_{i,\hat{j}} \tag{20}$$

其中，式 (19)(20) 均不包含非完整约束矩阵，因此经典 NSBC 方法的基本行为和复合行为指令均不满足非完整约束，且不符合多差速机器人的运动学方程。为此，本文首次将欠驱动模型式 (4) 引入 NSBC 框架中，代替质点模型以改进任务设计，并推导了具有非完整约束的 NSBC 范式，其中基本行为和复合行为的参考速度指令分别如式 (6)(16) 所示。由于所提出的非完整约束的 NSBC 框架考虑了非完整约束矩阵，因此基本行为和复合行为的参考指令均满足非完整约束，且符合多差速机器人的运动学方程。首次体现在改进了 NSBC 的任务设计范式，从而使得基本行为和复合行为指令均满足非完整约束，且适用于多差速机器人。

3 多智能体强化学习任务监管器设计（The design of multi-agent reinforcement learning mission supervisor）

由于 NSBC 方法通常在任务层是集中式的，因此可将行为优先级切换问题建模为一个协作式的马尔可夫博弈问题，其中所有差速机器人共享一个团队奖励。MARLMS 是基于宽松 Q 学习（Leinent DQN）算法进行设计的[29]，其整体框图如图 2 所示，且伪代码如算法 1 所示，其中 λ 表示先前学习 Q 值的个数。

定义联合状态集合和联合行为集合分别为 $S=\{\boldsymbol{s}_t\}$ 和 $B=\{\boldsymbol{b}_t\}$，其中 $\boldsymbol{s}_t=[\bar{\boldsymbol{X}}_t^{\mathrm{T}},\bar{\boldsymbol{P}}_t^{\mathrm{T}},\mathcal{G}_t]\in\mathbb{R}^{4N+1}$，$\bar{\boldsymbol{X}}_t=[\bar{\boldsymbol{X}}_1^{\mathrm{T}},\bar{\boldsymbol{X}}_2^{\mathrm{T}},\cdots,\bar{\boldsymbol{X}}_N^{\mathrm{T}}]^{\mathrm{T}}\in\mathbb{R}^{3N}$ 表示多差速机器人系统

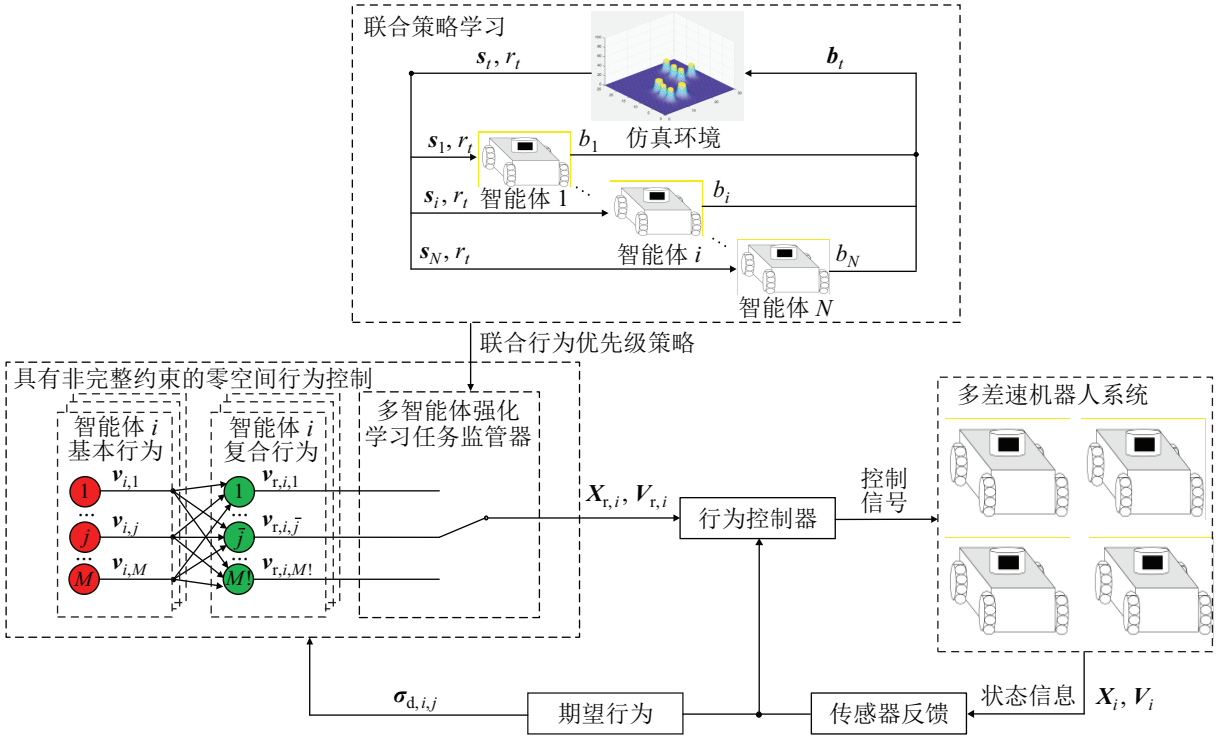


图 2 多智能体强化学习监管器的整体框图

Fig.2 Overall diagram of MARLMS

算法 1　多智能体强化学习任务监管器

输入： 训练的总回合数 T_{e}，一个回合的总时间步长 T_{s}

初始化 $Q(\boldsymbol{s}_t,\boldsymbol{b}_t;\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{V}},\boldsymbol{W}_{\mathrm{B}})=V(\boldsymbol{s}_t;\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{V}})+B(\boldsymbol{s}_t,\boldsymbol{b}_t;\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{B}})$，并给予初始化的网络权重 $\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{V}},\boldsymbol{W}_{\mathrm{B}}$

初始化经验池 $\mathcal{D}$

初始化贪婪探索策略 $\bar{T}(\phi(\boldsymbol{s}_t))$ 和宽松值 $\mathcal{L}_t$

for 回合 $=1,2,\cdots,T_{\mathrm{e}}$ **do**

　重置联合状态 $\boldsymbol{s}_t$ 至初始状态 $\boldsymbol{s}_0$

　for $t=1,2,\cdots,T_{\mathrm{s}}$ **do**

$$Q^{\mathrm{B}}_{t-1}(\boldsymbol{s}_t,\boldsymbol{b}_t)=\frac{1}{\lambda}\sum_{\iota=1}^{\lambda}Q(\boldsymbol{s}_t,\boldsymbol{b}_t;\boldsymbol{W}_{\mathrm{Q}_{t-\iota}},\boldsymbol{W}_{\mathrm{V}_{t-\iota}},\boldsymbol{W}_{\mathrm{B}_{t-\iota}})$$

$$y_{\boldsymbol{s}_t,\boldsymbol{b}_t}=E_{\mathrm{D}}\left[r+\gamma\max\nolimits_{\boldsymbol{b}_{t+1}}Q^{\mathrm{B}}_{t-1}(\boldsymbol{s}_{t+1},\boldsymbol{b}_{t+1})\mid\boldsymbol{s}_t,\boldsymbol{b}_t\right]$$

$$\boldsymbol{W}_{\mathrm{Q}_t},\boldsymbol{W}_{\mathrm{V}_t},\boldsymbol{W}_{\mathrm{B}_t}\approx\underset{\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{V}},\boldsymbol{W}_{\mathrm{B}}}{\arg\min}\,E_{\mathrm{D}}\left[(y_{\boldsymbol{s}_t,\boldsymbol{b}_t}-Q(\boldsymbol{s}_t,\boldsymbol{b}_t;\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{V}},\boldsymbol{W}_{\mathrm{B}}))^2\right]$$

　end for

end for

输出： $Q^{\mathrm{B}}_{(T_{\mathrm{e}}\cdot T_{\mathrm{s}})}(\boldsymbol{s}_t,\boldsymbol{b}_t)=\dfrac{1}{\lambda}\displaystyle\sum_{\bar{\iota}=0}^{\lambda-1}Q(\boldsymbol{s}_t,\boldsymbol{b}_t;\boldsymbol{W}_{\mathrm{Q}_{T_{\mathrm{e}}\cdot T_{\mathrm{s}}-\bar{\iota}}},\boldsymbol{W}_{\mathrm{V}_{T_{\mathrm{e}}\cdot T_{\mathrm{s}}-\bar{\iota}}},\boldsymbol{W}_{\mathrm{B}_{T_{\mathrm{e}}\cdot T_{\mathrm{s}}-\bar{\iota}}})$

的联合广义位置，$\bar{\boldsymbol{P}}_t=[\bar{P}_1,\bar{P}_2,\cdots,\bar{P}_N]^{\mathrm{T}}\in\mathbb{R}^N$ 表示联合行为优先级标识，$\bar{P}_i\in\mathbb{R}$ 表示第 i 个差速机器人的行为优先级标识，每一个行为优先级标识对应一个行为优先级的顺序，$\mathcal{G}_t\in\mathbb{R}$ 表示编队标识，$\mathcal{G}_t=0$ 表示当前多差速机器人系统未形成编队，$\mathcal{G}_t=1$ 表示当前多差速机器人系统正在重构队形，$\mathcal{G}_t=2$ 表示当前多差速机器人系统正在形成编队，$\mathcal{G}_t$ 的数值可根据多差速机器人系统与质心的相对位置判断，$\boldsymbol{b}_t=[\boldsymbol{V}^{\mathrm{T}}_{\mathrm{r},1},\boldsymbol{V}^{\mathrm{T}}_{\mathrm{r},2},\cdots,\boldsymbol{V}^{\mathrm{T}}_{\mathrm{r},N}]\in\mathbb{R}^{2N}$。然后，MARLMS 的奖励函数设计如下：

$$r_t=r_1+r_2 \tag{21}$$

$$r_1=\begin{cases}-10, & \mathcal{G}_t=0,\ \min\{d^{\mathrm{o}}_i\}\leqslant d_{\mathrm{OA}},\ \exists i=1,\cdots,N\\ 0, & \mathcal{G}_t=0,\ \min\{d^{\mathrm{o}}_i\}>d_{\mathrm{OA}},\ \forall i=1,\cdots,N\\ -10, & \mathcal{G}_t=1,\ \min\{d^{\mathrm{o}}_i\}\leqslant d_{\mathrm{OA}},\ \exists i=1,\cdots,N\\ +5, & \mathcal{G}_t=1,\ \min\{d^{\mathrm{o}}_i\}>d_{\mathrm{OA}},\ \forall i=1,\cdots,N\\ -10, & \mathcal{G}_t=2,\ \min\{d^{\mathrm{o}}_i\}\leqslant d_{\mathrm{OA}},\ \exists i=1,\cdots,N\\ +10, & \mathcal{G}_t=2,\ \min\{d^{\mathrm{o}}_i\}>d_{\mathrm{OA}},\ \forall i=1,\cdots,N\end{cases} \tag{22}$$

$$r_2=\begin{cases}0, & \bar{\boldsymbol{P}}_{t+1}=\bar{\boldsymbol{P}}_t\\ -3, & \bar{\boldsymbol{P}}_{t+1}\neq\bar{\boldsymbol{P}}_t\end{cases} \tag{23}$$

其中，奖励函数 r_t 由 r_1 和 r_2 两部分组成，r_1 的设计是以实现任务为目标，r_2 的设计是为了减少行为优先级切换次数。具体而言，由于任务目标包含形成期望的编队和避开路径上的障碍物，因此 r_1 的设计与是否形成期望的队形、是否避开障碍物和是否形成临时的重构队形相关。因为智能体的安全性在任务执行过程中是最重要的，所以只要有智能体违反安全约束，无论它们是否形成编队，团队就会得到一个 -10 的奖励。-10 的奖励旨在驱使多差速机器人优先选择避障。当多差速机器人未违反安全约束时，奖励应该聚焦于驱使多差速机器人形成编队：若多差速机器人形成了期望的队形，那么团队将接收到 $+10$ 的奖励；若多差速机器人形成了临时的重构队形，那么团队将接收到 $+5$ 的奖励；否则，团队将接收零奖励。$+10$ 的奖励旨在鼓励多差速机器人形成期望队形以实现任务目标，而 $+5$ 的奖励旨在鼓励多差速机器人在无法同时形成期望队形和避开障碍物的情况下，探索形成其他可能的队形以完成避障。r_2 的设计比较简单，其旨在减少行为优先级切换次数，若当前行为优先级与先前采样的一致，那么团队将收到 -3 的奖励；否则，团队将接收零奖励。下文将分析奖励参数的选取对任务需求的影响。首先，避障奖励范数值应该要求是最大的，否则多差速机器人可能为了实现编队而违反安全约束；其次，形成期望队形的奖励范数值应该大于重构队形，否则多差速机器人总是偏向于实现重构队形；最后，切换优先级的奖励范数值应该最小，减少行为优先级切换次数是理想情况，但不是任务目标之一。无论是暂态性能还是动态性能，都与行为优先级切换相关。若不设置 r_2，那么行为优先级切换的次数将增加显著，将导致任务误差的超调量、峰值时间、上升时间和稳态误差均增大。

多差速机器人系统与环境在 t 时间步交互，它们观测到联合状态 $\boldsymbol{s}_t$，基于一个 $\bar{T}(\phi(\boldsymbol{s}_t))$ 贪婪策略选取联合行为 $\boldsymbol{b}_t$，接收一个团队奖励 r_t，且转移至下一个联合状态 $\boldsymbol{s}_{t+1}$。$\bar{T}(\phi(\boldsymbol{s}_t))$ 贪婪策略是指多差速机器人系统以一个 $\bar{T}^{\zeta}(\phi(\boldsymbol{s}_t))$ 的概率选取一个随机的联合行为 $\boldsymbol{b}_t$，并以一个 $1-\bar{T}^{\zeta}(\phi(\boldsymbol{s}_t))$ 的概率选取 Q 值最大的联合行为 $\boldsymbol{b}_t=\arg\max_{\boldsymbol{b}}Q^{\mathrm{B}}_{t-1}(\boldsymbol{s}_t,\boldsymbol{b}_t)$，$\zeta$ 是一个指数。然后，该经历会存储到经验池 $\mathcal{D}$ 中，并使用一个宽松值 $\mathcal{L}(\boldsymbol{s}_t,\boldsymbol{b}_t)\in\mathbb{R}$ 标记如下

$$\mathcal{L}(\boldsymbol{s}_t,\boldsymbol{b}_t)=1-\mathrm{e}^{-\kappa_{\mathcal{L}}T_t(\phi(\boldsymbol{s}_t),\boldsymbol{b}_t)} \tag{24}$$

$$T_{t+1}(\phi(\boldsymbol{s}_t),\boldsymbol{b}_t)=\gamma_{\mathcal{L}}T_t(\phi(\boldsymbol{s}_t),\boldsymbol{b}_t) \tag{25}$$

$$\gamma_{\mathcal{L}}=\mathrm{e}^{\rho_\gamma d^{\mathrm{t}}_\gamma} \tag{26}$$

其中，$\kappa_{\mathcal{L}}$ 是宽松值的适度因子，$T_t(\phi(\boldsymbol{s}_t),\boldsymbol{b}_t)$ 是衰减温度，$\phi(\cdot)$ 是哈希自动编码函数，$\gamma_{\mathcal{L}}$ 是折扣系数，ρ_γ 是温度指数，d^{t}_γ 是衰减率。

由于 Q 值的估计过高会破坏正确的学习，因此

引入 Dueling 网络结构和平均 Q 值思想提升 Q 值的估计精度和学习的稳定性，根据宽松值 $\mathcal{L}_t$ 计算 Q 值：

$$Q(\boldsymbol{s}_t,\boldsymbol{b}_t)=\begin{cases}Q(\boldsymbol{s}_t,\boldsymbol{b}_t)+\alpha_t\delta_t, & \delta_t>0 \text{ 或 } \vartheta>\mathcal{L}_t\\ Q(\boldsymbol{s}_t,\boldsymbol{b}_t), & \delta_t\leqslant 0 \text{ 且 } \vartheta\leqslant\mathcal{L}_t\end{cases} \tag{27}$$

其中，$\alpha_t\in(0,1)$ 是学习率，$\vartheta\sim U(0,1)$ 表示一个随机变量，$\delta_t=y_{\boldsymbol{s}_t,\boldsymbol{b}_t}-Q^{\mathrm{B}}_{t-1}(\boldsymbol{s}_t,\boldsymbol{b}_t)$ 是时序差分误差，$y_{\boldsymbol{s}_t,\boldsymbol{b}_t}=E_{\mathrm{D}}[r+\gamma\max_{\boldsymbol{b}_{t+1}}Q^{\mathrm{B}}_{t-1}(\boldsymbol{s}_{t+1},\boldsymbol{b}_{t+1})\mid\boldsymbol{s}_t,\boldsymbol{b}_t]$。

MARLMS 的离线训练会在所有回合结束后停止。最后，所学习的联合策略指导多差速机器人系统在实际场景中选择最优的联合行为优先级。在每个采样周期，当联合行为优先级确定后，多差速机器人系统的参考速度指令可根据式 (16)～式 (18) 计算。

在 NSBC 框架中，FSAMS[16]、MPCMS[23] 和 RLMS[25] 是主流的任务监管器。FSAMS 将复合行为隐藏于有限状态机的状态中，通过设计状态转移规则实现行为优先级切换，其易于实施，但依赖人类智能。MPCMS 将行为优先级的切换问题描述为模式切换最优控制问题，且通过混合整数优化控制算法求解该问题，其降低了对人类智能的依赖，但对高性能硬件计算平台依赖强。文 [25-26] 提出的 RLMS 将行为优先级切换问题建模为马尔可夫决策过程，降低了对人类智能和计算平台的依赖，但是最大化了个体性能，且无法实施局部行为，具有很强的局限性。本文在 RLMS 的基础上，首次将优先级切换问题建模为马尔可夫博弈问题，不仅能学习最优的联合行为优先级策略，而且能最大化团队性能，克服了 RLMS 无法实施全局行为的致命缺陷。首次体现在行为优先级切换问题的建模上，以及联合策略学习的思想，从而最大化团体性能而非个体性能，且允许实施全局行为，达到群体协作。

4 数值仿真（Numerical simulation）

4.1 仿真配置

本节设置了一个数值仿真案例，其中控制对象为 3 个多差速机器人系统，其运动学方程如式 (2) 所示，控制目标为 3 个机器人以编队的形式移动至目标位置同时避开路径上的障碍物。所有差速机器人均假设具有探测工作环境的能力。环境和 MARLMS 所使用的仿真参数分别如表 1 和表 2 所示。为了验证所提出方法的有效性和优越性，进行了 3 组对比仿真测试，包括 MARLMS 学习前后对比、所提出具有非完整约束的 NSBC 与经典 NSBC 方法对比，以及 MARLMS 与现有主流任务监管器的对比分析。MARLMS 的网络结构如图 3 所示，其中网络的输入为联合状态 $\boldsymbol{s}_t$，输出为所有联合行为的 Q 值。为了提升 Q 值的估计精度，MARLMS 的网络结构内嵌了 Dueling 网络结构，即 Q 值网络分解为一个状态值函数网络和一个行为优势网络。

表 1 环境的仿真参数值

Tab.1 Simulation parameter values of the environment

参数名称	参数符号	参数数值
障碍物 1 位置	$\boldsymbol{p}_{\mathrm{O1}}$	$(x-25)^2+y^2=1^2$
障碍物 2 位置	$\boldsymbol{p}_{\mathrm{O2}}$	$(x-65)^2+(y-8)^2=5^2$
障碍物 3 位置	$\boldsymbol{p}_{\mathrm{O3}}$	$(x-75)^2+(y-8)^2=5^2$
障碍物 4 位置	$\boldsymbol{p}_{\mathrm{O4}}$	$(x-65)^2+(y+8)^2=5^2$
障碍物 5 位置	$\boldsymbol{p}_{\mathrm{O5}}$	$(x-75)^2+(y+8)^2=5^2$
编队质心期望轨迹	$\boldsymbol{p}_{\mathrm{c,d}}$	$[-4+t,0]$ m
编队相对位置	$\boldsymbol{p}^{\mathrm{c}}_1,\boldsymbol{p}^{\mathrm{c}}_2,\boldsymbol{p}^{\mathrm{c}}_3$	$[-4,0],[-2,6],[-2,-6]$ m
编队重构矩阵	$\boldsymbol{\Gamma}_{\mathrm{FR},1},\boldsymbol{\Gamma}_{\mathrm{FR},2},\boldsymbol{\Gamma}_{\mathrm{FR},3}$	$\begin{bmatrix}-\frac{1}{2}&0\\0&0\end{bmatrix},\boldsymbol{0}_2,\begin{bmatrix}1&0\\0&0\end{bmatrix}$
安全距离	d_{OA}	1 m
任务增益	$\boldsymbol{\Lambda}_{\mathrm{FM}},\boldsymbol{\Lambda}_{\mathrm{FR}},\boldsymbol{\Lambda}_{\mathrm{OA}}$	$9\boldsymbol{I}_3,9\boldsymbol{I}_3,20\boldsymbol{I}_2$
起始位置	$\boldsymbol{p}_{1,\mathrm{o}},\boldsymbol{p}_{2,\mathrm{o}},\boldsymbol{p}_{3,\mathrm{o}}$	$[0,0],[-7,7],[-7,-7]$ m
目标位置	$\boldsymbol{p}_{1,\mathrm{g}},\boldsymbol{p}_{2,\mathrm{g}},\boldsymbol{p}_{3,\mathrm{g}}$	$[100,0],[94,6],[94,-6]$ m

表 2 MARLMS 的仿真参数值

Tab.2 Simulation parameter values of the MARLMS

参数名称	参数符号	参数数值
训练的总回合数	T_{e}	100 000
一个回合的总时间步长	T_{s}	1 000
宽松值的适度因子	$\kappa_{\mathcal{L}}$	2
温度指数	ρ_γ	−0.01
衰减率	d_γ	0.95
学习率	α_t	0.000 1
探索指数	ζ	0.999
经验池数量	$\mathcal{D}$	50 000
采样时间	Δt	0.01 s

4.2 具有非完整约束的 NSBC 方法与经典 NSBC 方法对比

本节对比了所提出的具有非完整约束的 NSBC 方法与经典 NSBC 方法的仿真结果，如图 4 所示。由于经典 NSBC 方法以质点模型建模智能体的运动学，因此忽略了位置和方向间的耦合，智能体可以在不改变角度的情况下到达任意位置。将经典 NSBC 方法直接应用于多差速机器人系统时，智能

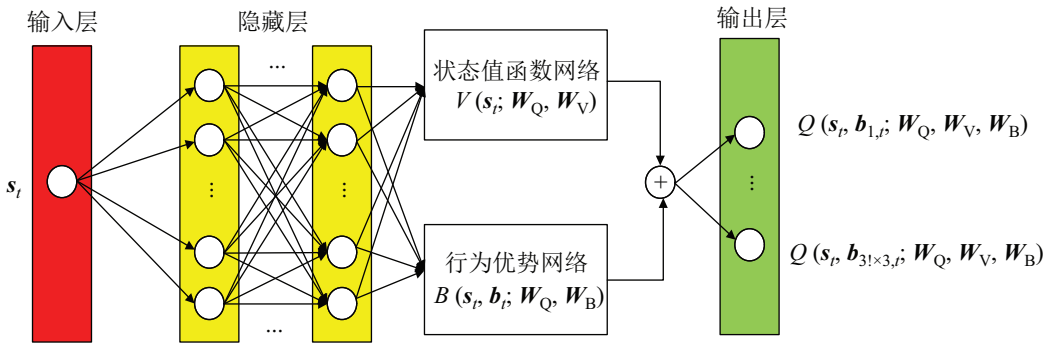
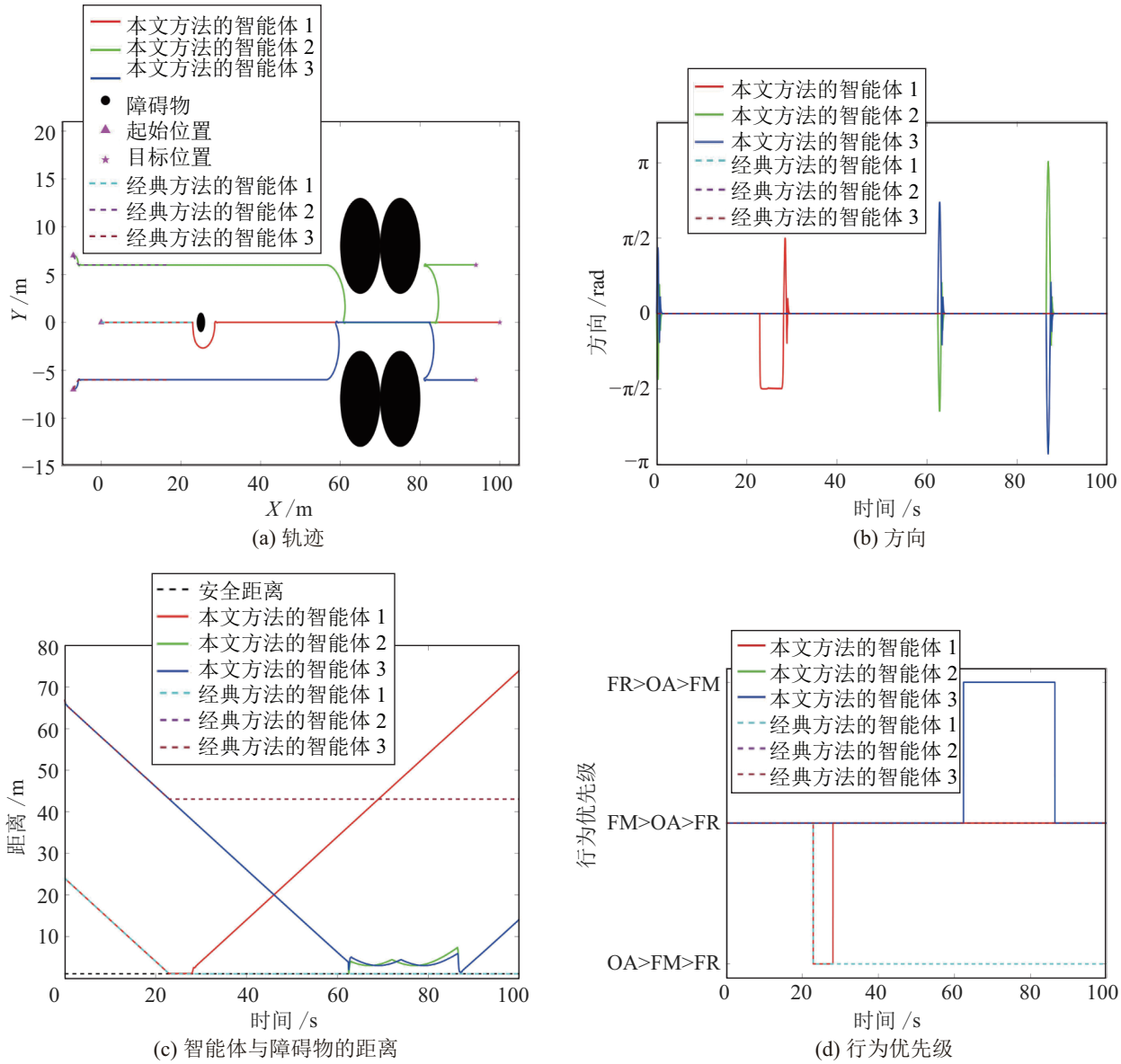


图 3　MARLMS 的网络结构图

Fig.3　Network structure diagram of MARLMS



(a) 轨迹　(b) 方向

(c) 智能体与障碍物的距离　(d) 行为优先级

图 4　本文 NSBC 方法与经典 NSBC 方法的仿真对比结果图

Fig.4　Simulation comparison results of the proposed NSBC method and the traditional NSBC method

体容易陷入最小极值状态。图 4 显示了当障碍物处于智能体的正前方时，经典 NSBC 方法不会改变智能体的方向，从而导致智能体进入最小极值状态且无法摆脱，进而造成任务目标的失败。所提出的具有非完整约束的 NSBC 方法采用差速模型建模智能体的运动学，该模型考虑了智能体位置和方向间的

耦合，能远比质点模型更为精确地描述多差速机器人系统。图 4 显示了当智能体遭遇障碍物时，本文提出的具有非完整约束的 NSBC 方法会给出改变智能体方向的参考指令，从而从最小极值状态中快速离开，进而完成预定的任务目标。该仿真对比证明了本文方法比经典 NSBC 方法更适用于多差速机器人系统，且对最小极值状态的鲁棒性更强。

4.3 本文 MARLMS 与现有主流任务监管器对比

本节对比了本文 MARLMS 与现有主流任务监管器的仿真结果，包括 FSAMS [16]、MPCMS [23] 和 RLMS [25]。对于 FSAMS，第 i 个差速机器人的切换规则设计如下：如果满足 $d_{\mathrm{OA}} < \min\{d_i^{\mathrm{o}}\} \leqslant 3d_{\mathrm{OA}}$，那么切换 FR 行为为最高行为优先级；如果满足 $\min\{d_i^{\mathrm{o}}\} \leqslant d_{\mathrm{OA}}$，那么切换 OA 行为为最高优先级；否则，切换 FM 行为为最高行为优先级。对于 MPCMS，代价函数为编队误差和重构误差的加权和，约束为智能体与障碍物的距离大于安全距离。对于 RLMS，因为全局行为无法实施，FM 行为和 FR 行为均视为运动行为。MARLMS、FSAMS、MPCMS 和 RLMS 的仿真对比结果如图 5～图 9 所示。由图 5 可知，因为 MPCMS 在每个采样周期内均需要在线求解最优行为优先级，所以 MPCMS 的平均迭代时间远高于其他任务监管器。由于 FSAMS 只使用多差速机器人系统的当前状态信息且在状态转移阈值附近存在开关效应，因此图 7(a)(d) 中多差速机器人系统的轨迹存在抖振且行为优先级存在不理想的切换，这将导致多差速机器人系统有时会违反安全约束。相较于 FSAMS，MARLMS 学习一个联合的行为优先级策略，在任务执行过程中能智能且动态地切换行为优先级，因而轨迹是光滑的，且行为优先级切换结果是理想的。由图 8 可知，MPCMS 和 MARLMS 均实现了预定的任务目标，且二者的行为优先级切换性能十分接近。根本原因是 MPCMS 和 MARLMS 均使用了多差速机器人系统的未来状态，其中 MPCMS 在每一个采样周期计算预测时域内的状态信息时，考虑了未来的折扣奖励以保证累积奖励的最大化。由图 9 可知，RLMS 只适用于单智能体系统，虽然能够避开路径附近的障碍物，但是无法形成编队和实现任务目标。为了更好地体现 MARLMS 任务性能的优越性，表 3 统计了不同任务监管器的平均行为优先级切换次数、平均安全约束违反次数、平均在线迭代时间和任务目标完成情况。由表 3 可知，相较于 FSAMS，MARLMS 的平均行为优先级切换次数和平均安全约束违反次数更少，即行为优先级切换的性能更佳。相较于 MPCMS，MARLMS 的平均在线迭代时间更短，即实时性更佳。相较于 RLMS，MARLMS 成功完成了任务目标，但是 RLMS 失败了，因此 MARLMS 的群体性能更佳。该仿真对比结果证明了本文 MARLMS 的优越性，其不仅避免了人工设计优先级切换规则，而且大幅度降低了在线计算量和保证了实时性。

表 3 不同任务监管器的任务性能对比

Tab.3 Comparison of mission performance among different mission supervisors

性能指标	MARLMS	FSAMS	MPCMS	RLMS
平均行为优先级切换次数	2	101	2	2
平均安全约束违反次数	0	3	0	10
平均在线迭代时间	0.55 ms	0.54 ms	200 ms	0.51 ms
任务目标完成情况	成功	成功	成功	失败

MARLMS 的运行时间可分为离线训练阶段和在线执行阶段。在离线训练阶段，MARLMS 需要完成 100 000 回合的训练。本文使用 core-i7 的惠普笔记本电脑，一回合的训练用时大约在 1 s 左右，且总训练时长大约在 27 h 左右。离线训练阶段可以使用高性能电脑或云端计算来加快训练速度，也可以使用并行计算框架协同计算来减少每台电脑的训练总回合数。在线执行阶段，MARLMS 只需要调用离线学习到的策略完成行为优先级切换，每次采用的平均迭代时间为 0.55 ms，足够保证行为优先级决策的实时性。实验结果表明了多差速机器人未遭遇决策时延的问题，从而验证了 MARLMS 的实时性能够满足需求。

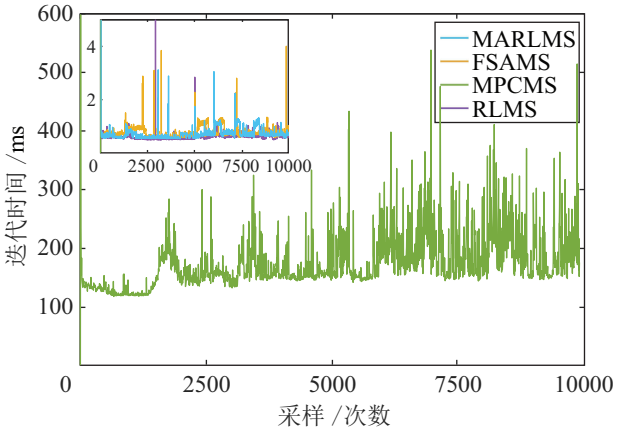


图 5 MARLMS、FSAMS、MPCMS 和 RLMS 的平均迭代时间对比结果

Fig.5 Comparison results of the average iteration time among the MARLMS, FSAMS, MPCMS and RLMS

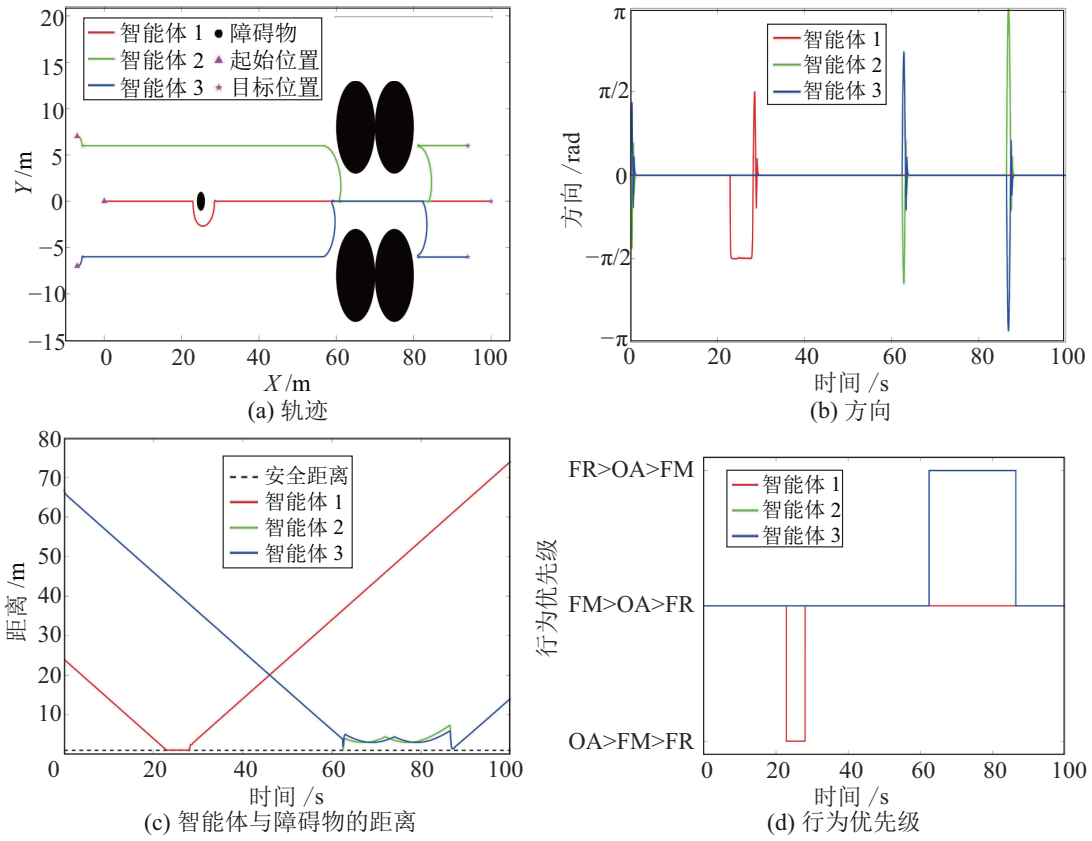


图 6　MARLMS 的仿真结果

Fig.6　Simulation results of the MARLMS

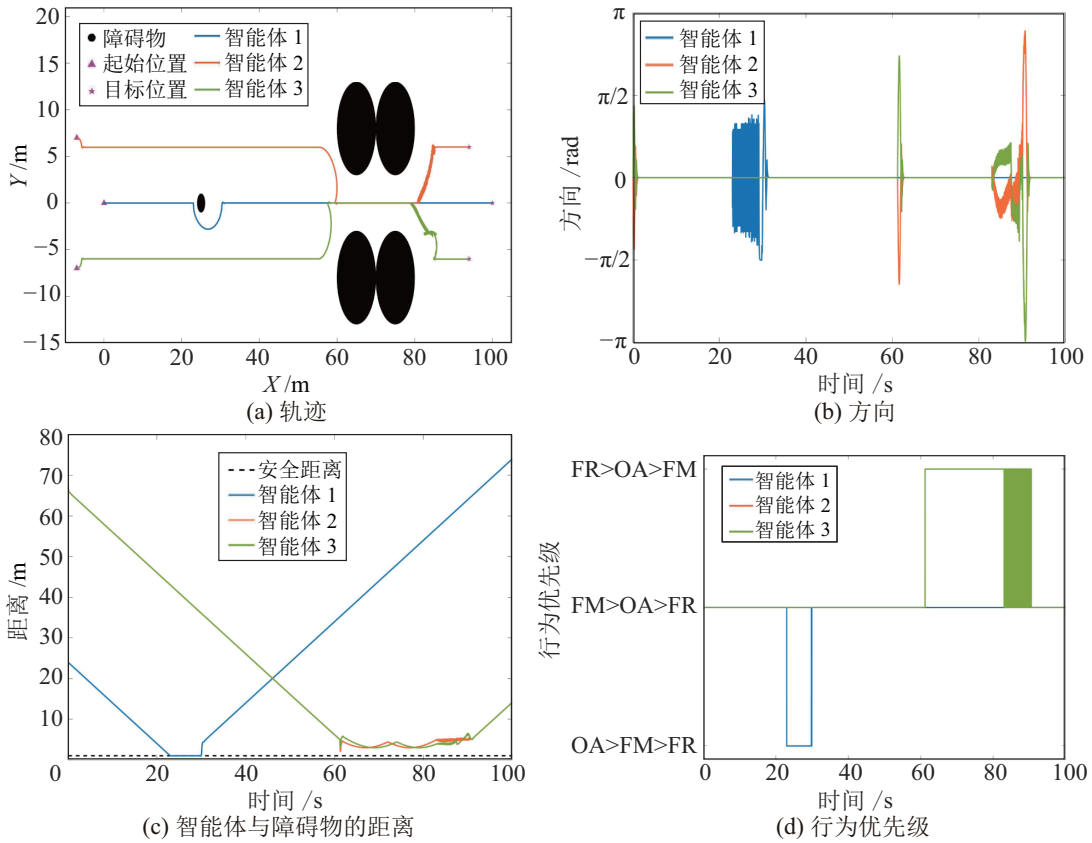


图 7　FSAMS 的仿真结果

Fig.7　Simulation results of the FSAMS

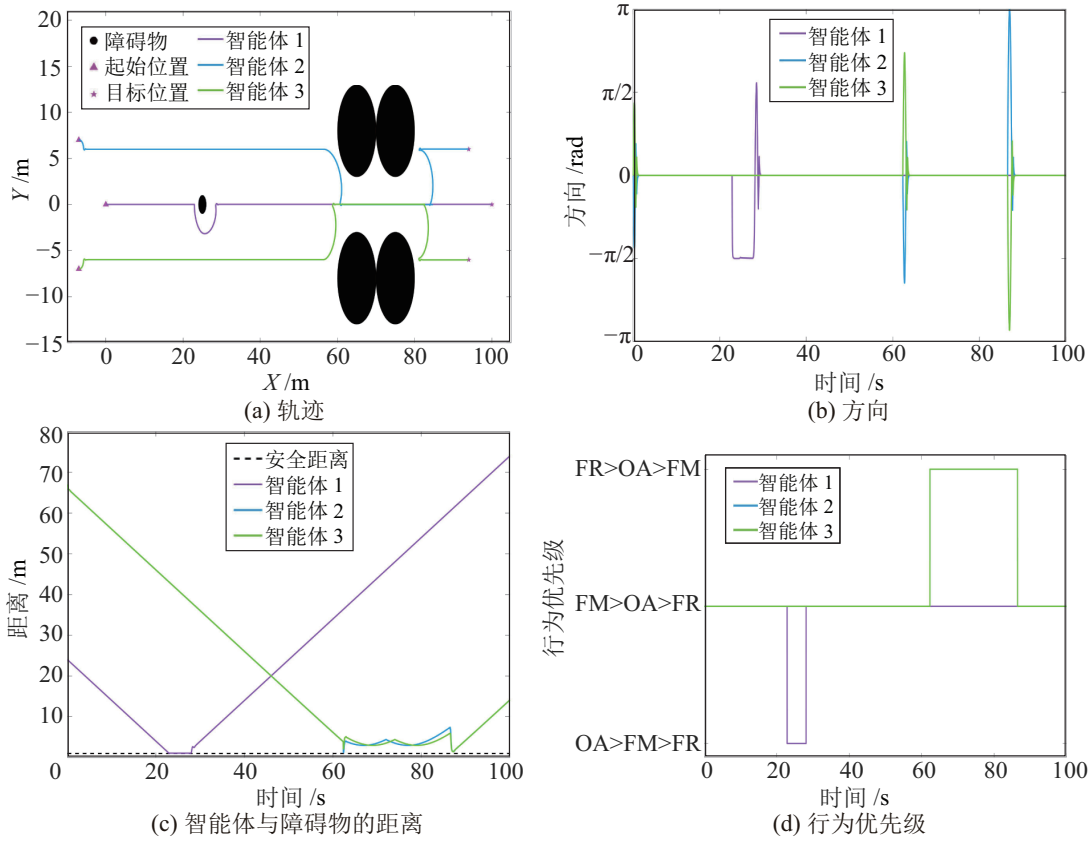


图 8 MPCMS 的仿真结果

Fig.8 Simulation results of the MPCMS

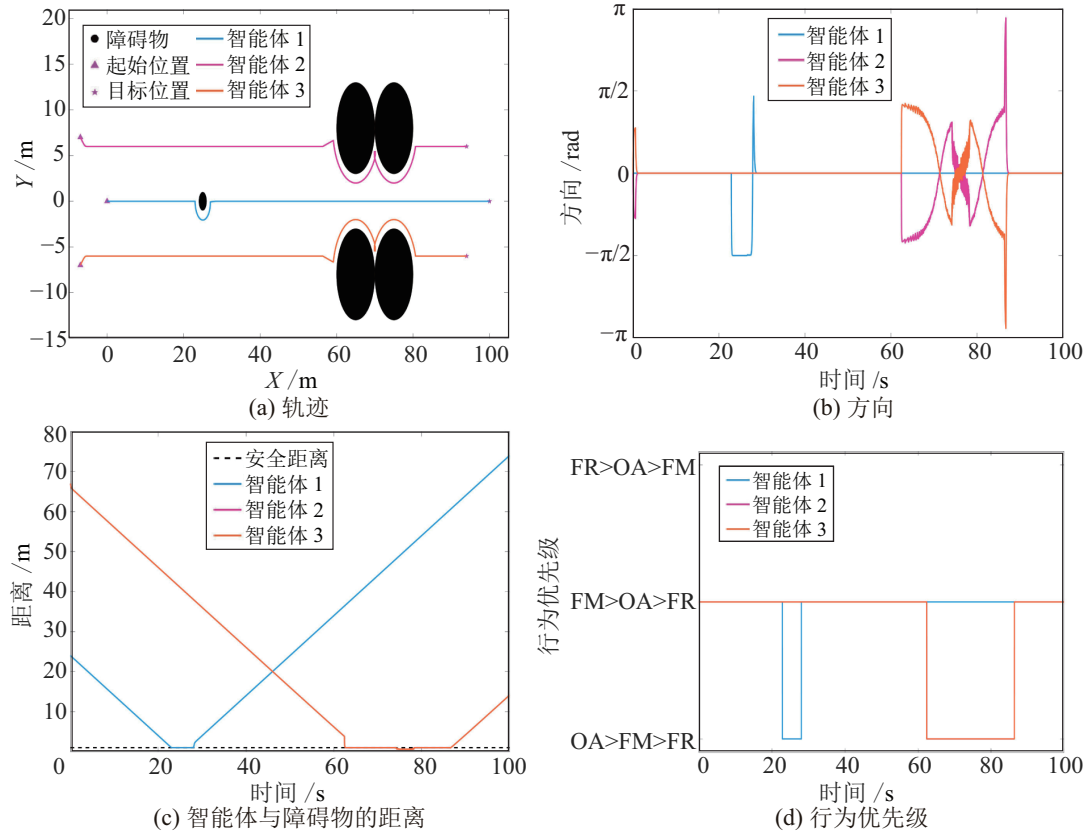


图 9 RLMS 的仿真结果

Fig.9 Simulation results of the RLMS

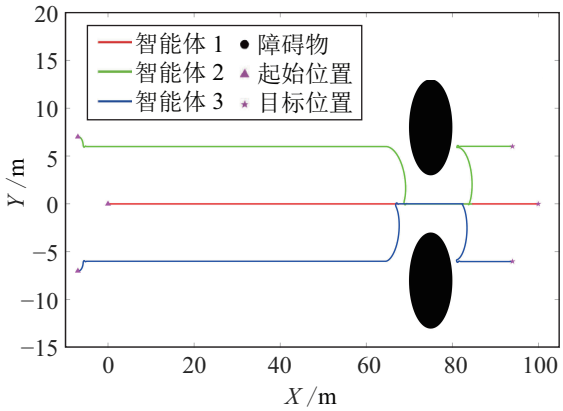


图 10　在部分障碍物未知环境的离线训练轨迹图

Fig.10　Trajectories of off-line training in the environment with some unknown obstacles

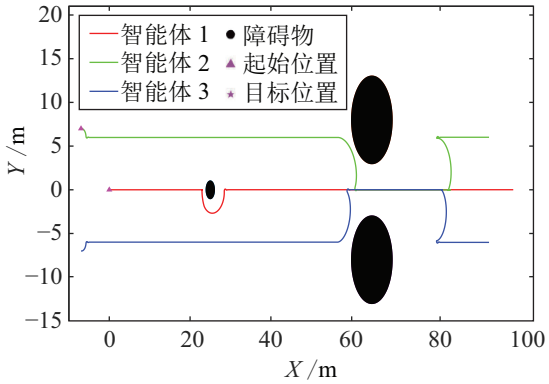


图 11　在部分障碍物未知环境的多差速机器人任务执行轨迹图

Fig.11　Mission execution trajectories of differential drive robots in the environment with some unknown obstacles

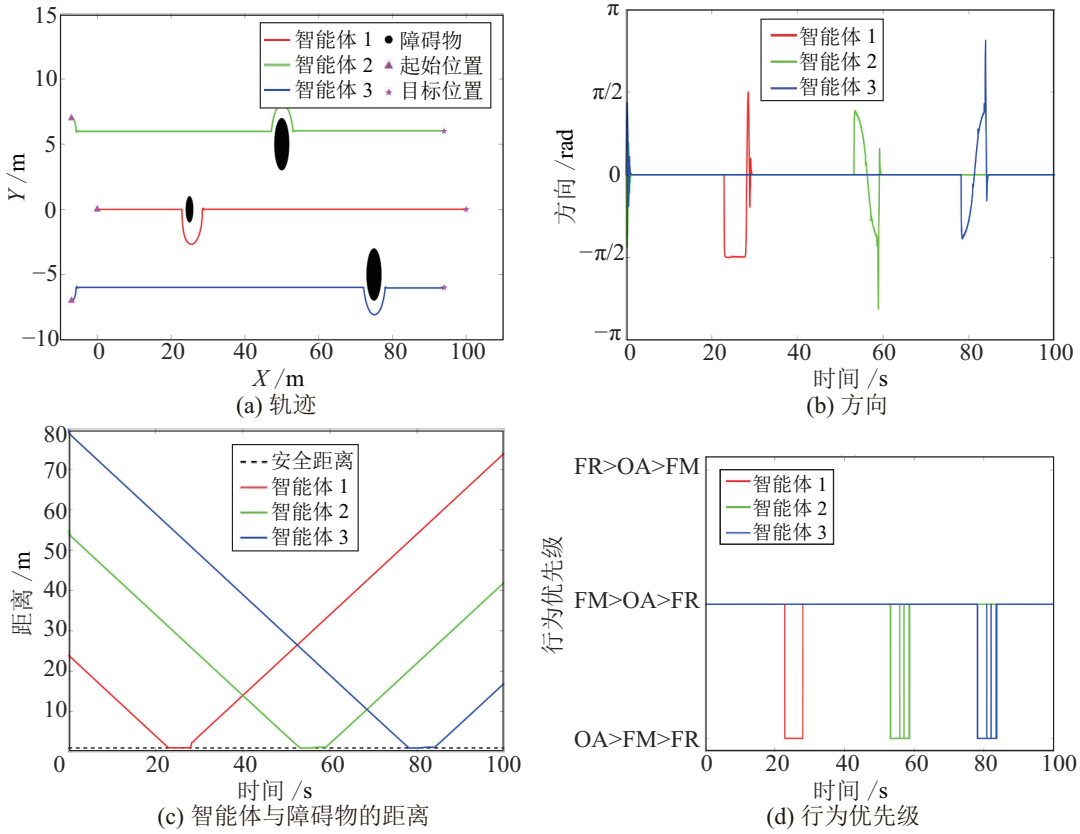


图 12　离线训练和在线执行环境不一致时 MARLMS 的仿真结果

Fig.12　Simulation results of MARLMS when offline training and online execution environments are inconsistent

4.4　未知障碍物场景的仿真测试

因为未知障碍物在强化学习的离线训练阶段是无法获取的，所以任务性能完全依赖于行为优先级策略的泛化性。针对未知障碍物的案例，需要在 MARLMS 的状态集合中补充第 i 个差速机器人与障碍物的最小距离，即避障行为中的 $\min\{d_i^{\mathrm{o}}\}$。由于原状态集合只包含多差速机器人和编队的状态，因此对环境的表征并不充分，而扩充状态集合并未改变算法 1 的整体框架，且若使用视觉或者图像等更高维度信息作为状态集合的元素，那么在实际多差速机器人应用时其任务性能会更佳。在该案例中，障碍物 1、2 和 4 设置为未知障碍物，因此在离线训练时，它们对于多差速机器人是未知的。此时，离线训练完成时的多差机器人轨迹图如图 10 所示。然后，将学习所得的最优行为优先级策略作用于原仿真场景，此时多差速机器人的轨迹如图 11 所示。

即使部分障碍物对于离线训练阶段是未知的，只要环境状态表征得足够充分，那么所学习的策略也能够凭借算法的泛化性保证多差速机器人完成预定的任务目标。

4.5 离线训练和在线执行环境不一致的仿真测试

在离线训练和在线执行环境不一致时的仿真测试中离线训练环境使用先前的仿真配置，而在线执行环境不包含障碍物 2～5，且增加了 2 个新的障碍物，分别为 $(x-50)^2+(y-5)^2=2^2$ 和 $(x-75)^2+(y+5)^2=2^2$。MARLMS 先在离线环境中学习至收敛，再将学习的策略应用于在线执行环境，其仿真结果如图 12 所示。仿真结果验证了多差速机器人系统仍能避开障碍物且实现任务目标。因为避开路径附近的障碍物是通过 OA 行为实现的，所以无论离线训练和在线执行环境是否一致，多差速机器人系统在遭遇障碍物时都会执行 OA 行为以避开障碍物。然而，从图 12(d) 中不难发现行为优先级存在不理想的切换。由于离线环境和在线执行环境是不一致的，因此 MARLMS 的联合行为优先级策略对于在线执行环境不是最优的，所以必然会存在不合理的行为优先级切换。为此，需要 MARLMS 在在线任务环境中利用所得到的经历进行学习，从而获取在线环境下的最优策略。

4.6 5 个差速机器人系统案例的仿真测试

5 个差速机器人系统的编队质心期望轨迹为 $[-6+t,0]$ m。编队相对位置分别为 $\boldsymbol{p}_1^{\mathrm{c}}=[6,0]$ m、$\boldsymbol{p}_2^{\mathrm{c}}=[0,6]$ m、$\boldsymbol{p}_3^{\mathrm{c}}=[0,-6]$ m、$\boldsymbol{p}_4^{\mathrm{c}}=[-3,9]$ m 和 $\boldsymbol{p}_5^{\mathrm{c}}=[-3,-9]$ m。编队重构矩阵分别为 $\boldsymbol{\Gamma}_{\mathrm{FR},1}=\boldsymbol{0}_2$、$\boldsymbol{\Gamma}_{\mathrm{FR},2}=\begin{bmatrix}0 & 1/6\\ 0 & 0\end{bmatrix}$、$\boldsymbol{\Gamma}_{\mathrm{FR},3}=\begin{bmatrix}0 & 1/3\\ 0 & 0\end{bmatrix}$、$\boldsymbol{\Gamma}_{\mathrm{FR},4}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$ 和 $\boldsymbol{\Gamma}_{\mathrm{FR},5}=\begin{bmatrix}4/3 & 0\\ 0 & 0\end{bmatrix}$。MARLMS 完成离线训练后的仿真结果如图 13 所示。图 13(a) 显示了 5 个差速机器人系统可以通过执行 FM、FR 和 OA 行为，形成期望队形且避开路径附近的障碍物。图 13(c) 显示了 5 个差速机器人系统均不会违反安全约束，在任务过程中始终与障碍物保持安全距离。图 13(d) 显示了行为优先级切换是理想的，不存在不合理的行为优先级切换。上述仿真结果验证了所提出的 MARLMS 具有一定的可扩展性。此外，MARLMS 可以通过云平台或并行训练加快学习。由于状态空间和行为空间随智能体数量呈指数增长，因此 MARLMS 不适用于大规模系统。

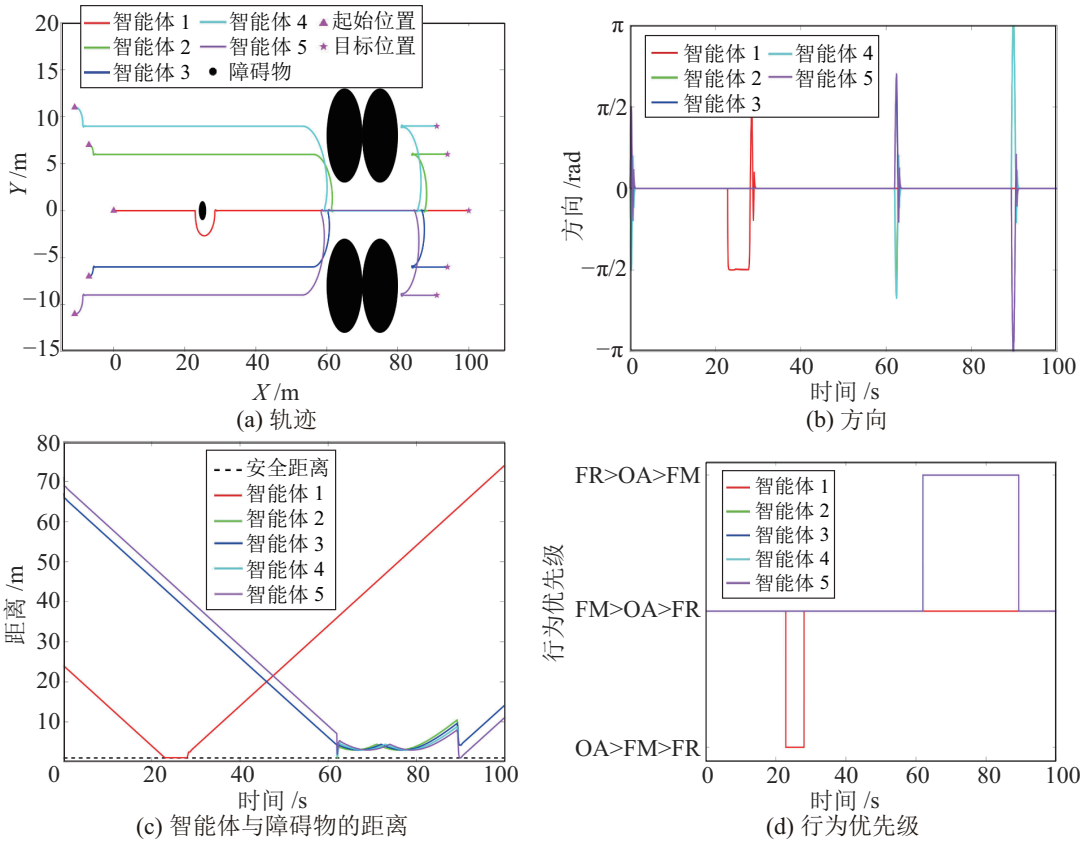
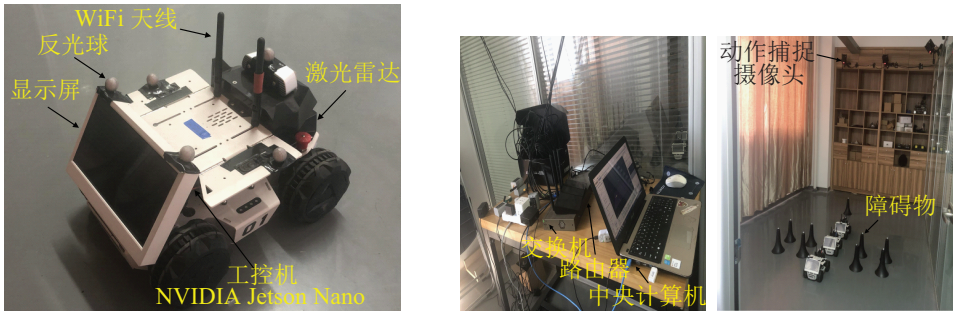


图 13 5 个差速机器人系统的 MARLMS 仿真结果

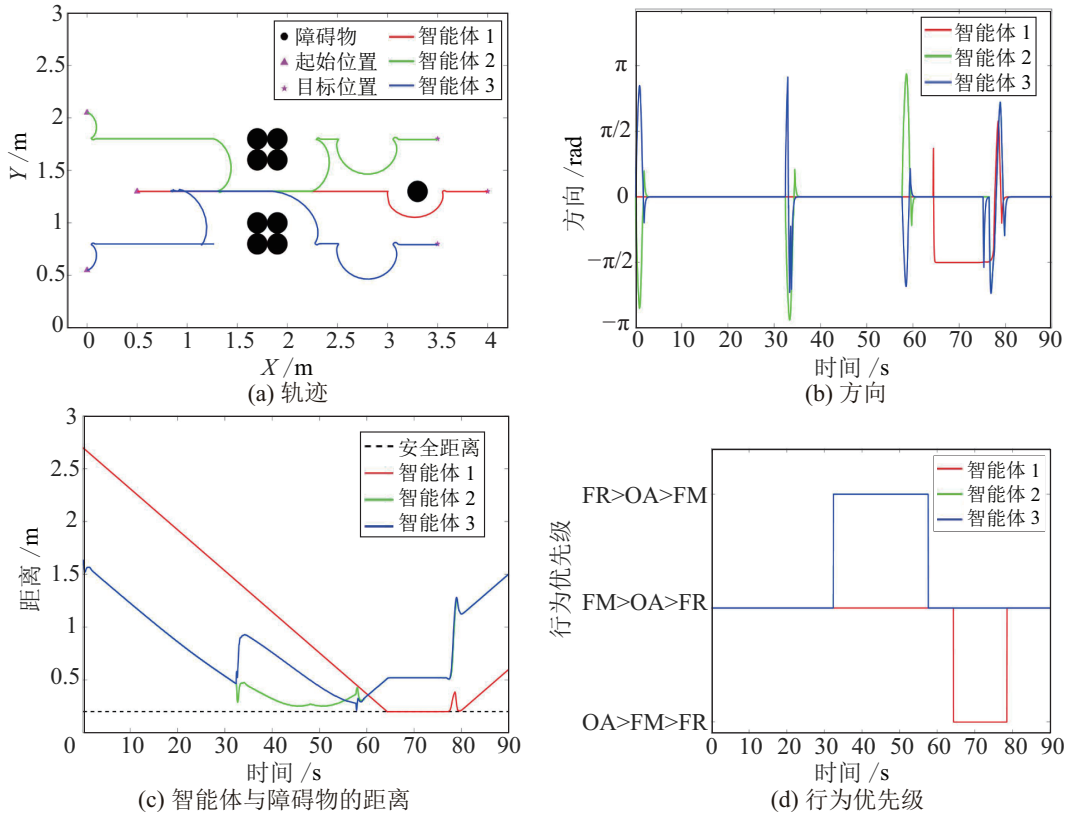
Fig.13 Simulation results of MARLMS of five differential drive robots



(a) AgileX Limo 机器人　　(b) 实验装置

图 14　实验配置示意

Fig.14　Experimental configuration schematics

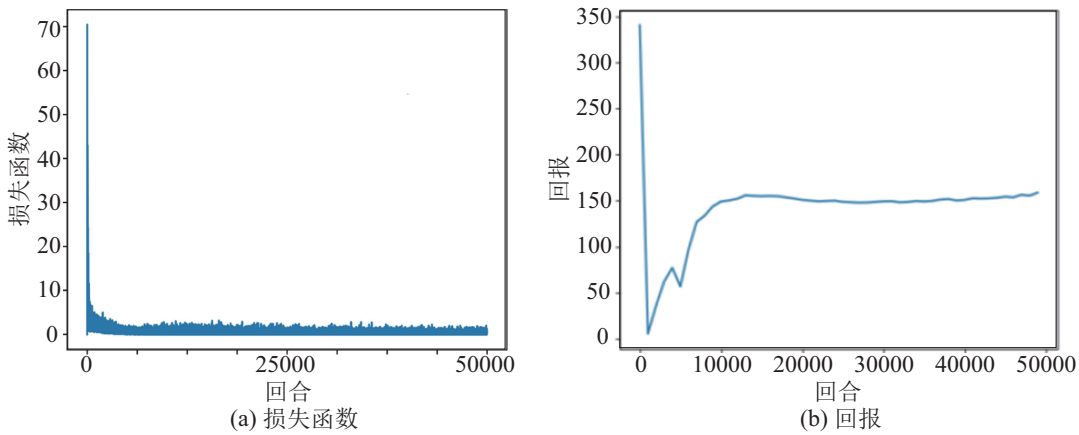


(a) 轨迹　(b) 方向

(c) 智能体与障碍物的距离　(d) 行为优先级

图 15　实验环节中 MARLMS 学习后的训练结果

Fig.15　Training results of the MARLMS after learning in the experiment



(a) 损失函数　(b) 回报

图 16　实验环节中 MARLMS 的训练曲线

Fig.16　Training curves of the MARLMS in the experiment

5 实验验证（Experiment validation）

5.1 实验配置

实验配置示意图如图 14 所示，其中图 14(a) 显示的每个 AgileX Limo 机器人均设置为差速模式。在任务执行过程中，多 AgileX Limo 机器人系统使用激光雷达探测障碍物。具体来说，中央计算机分别使用 WiFi 和动作捕捉相机通过运行一个 Python 编码的 MARLMS 程序来获取传感器信息和 AgileX Limo 机器人位置信息。然后，中央计算机基于接收到的传感器和位置信息计算多 AgileX Limo 机器人系统的联合行为优先级。之后，中央计算机将位置信息和联合行为优先级发送给每个 AgileX Limo 机器人。根据接收到的位置信息和联合行为优先级，每个 AgileX Limo 机器人使用工控机（NVIDIA Jetson Nano）计算参考速度命令。最后，多 AgileX Limo 机器人系统执行参考速度命令，且中央计算机接收新的传感器和位置信息。整个过程一直持续至多 AgileX Limo 机器人系统移动到预定的目标位置。在整个实验环节中，MARLMS 需要先在离线环境中训练，直至网络收敛且学习到一个联合行为优先级策略。在训练完成后，再将 MARLMS 导入实际的多差速机器人中，以在线指导智能体智能地切换行为优先级。MARLMS 的实验参数值如表 4 所示。在实验环节的离线训练阶段，MARLMS 学习后的训练结果分别如图 15 所示。在整个实验环节的离线训练过程中，MARLMS 的损失函数和回报如图 16 所示。

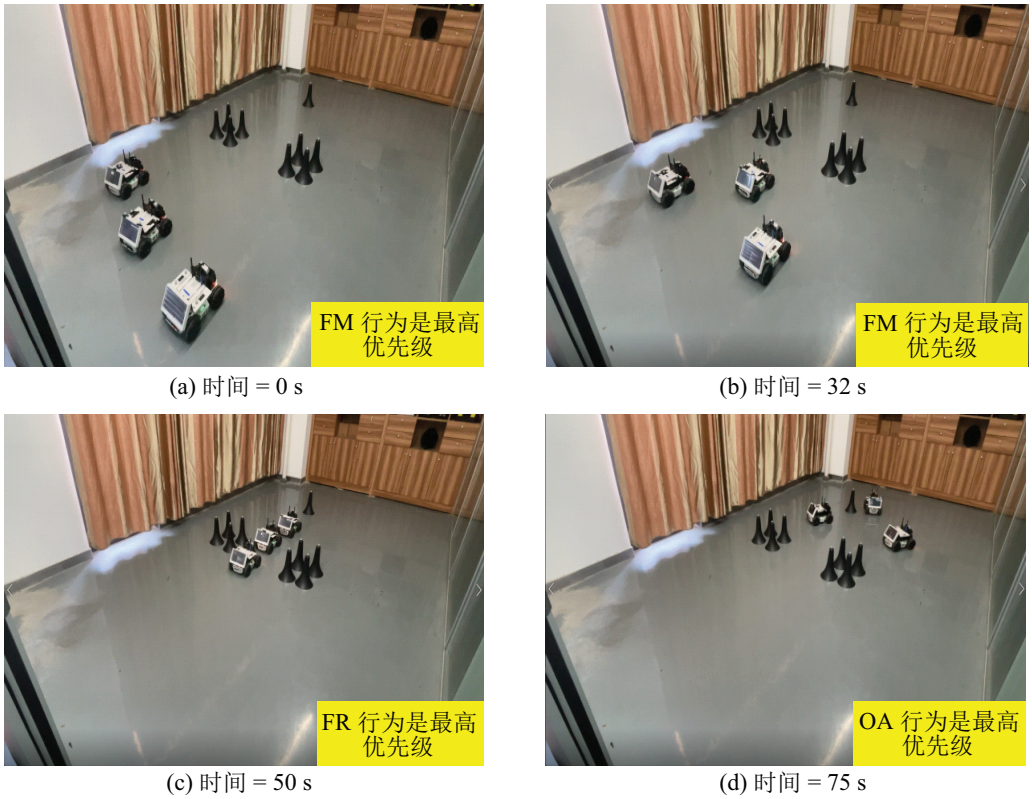
表 4 MARLMS 的实验参数值

Tab.4 Experimental parameter values of the MARLMS

参数名称	参数符号	参数数值
训练的总回合数	T_{e}	50 000
一个回合的总时间步长	T_{s}	600
宽松值的适度因子	$\kappa_{\mathcal{C}}$	2
温度指数	ρ_{γ}	−0.01
衰减率	d_{γ}	0.9
学习率	α_t	0.000 1
探索指数	ζ	0.995
经验池数量	$\mathcal{D}$	20 000
采样时间	Δt	0.15 s

5.2 实验结果

本文 MARLMS 在多 AgileX Limo 机器人系统上的实验验证结果如图 17～图 20 所示，其中图 17 是整个实验过程的快照，图 18 是 MARLMS 的实验结果图。图 17(a) 显示了多 AgileX Limo 机器人系统在起始阶段未形成编队，因此 FM 行为是最高优

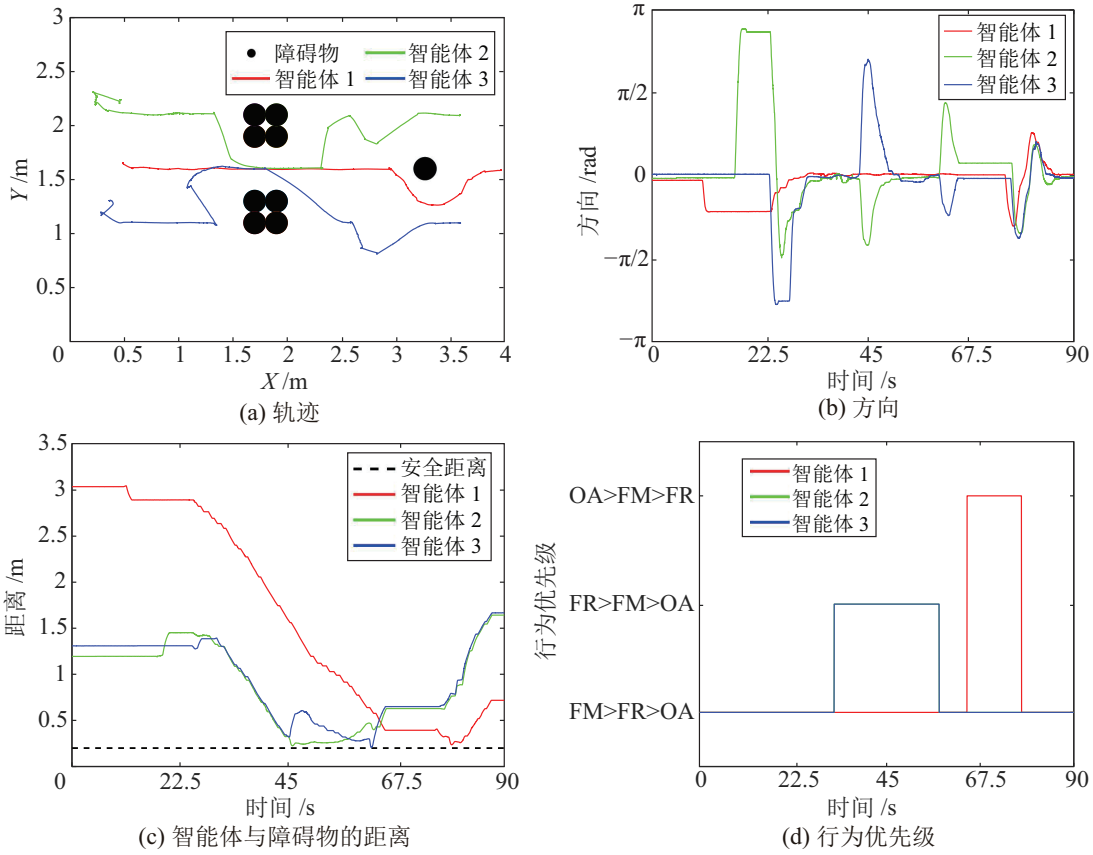


(a) 时间 = 0 s (b) 时间 = 32 s

(c) 时间 = 50 s (d) 时间 = 75 s

图 17 多 AgileX Limo 机器人的实验快照

Fig.17 The snapshots of AgileX Limo robots

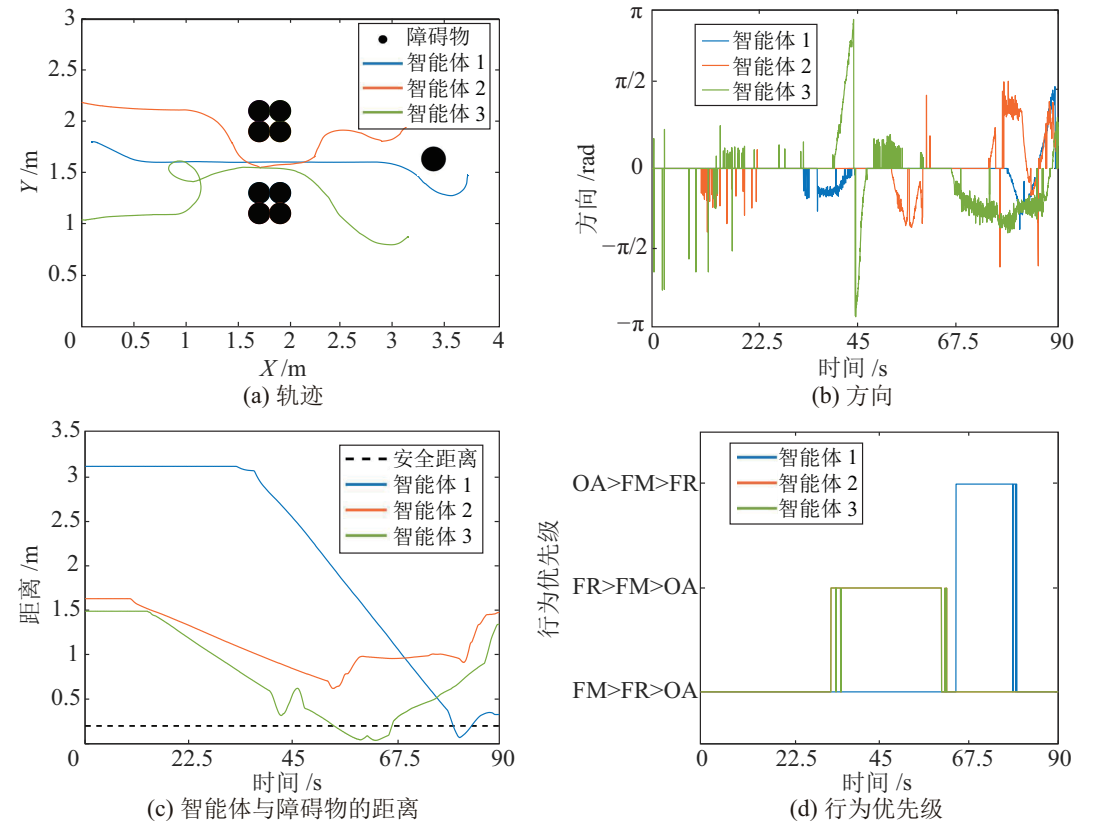


(a) 轨迹　(b) 方向

(c) 智能体与障碍物的距离　(d) 行为优先级

图 18　MARLMS 的实验结果

Fig.18　Experimental results of the MARLMS

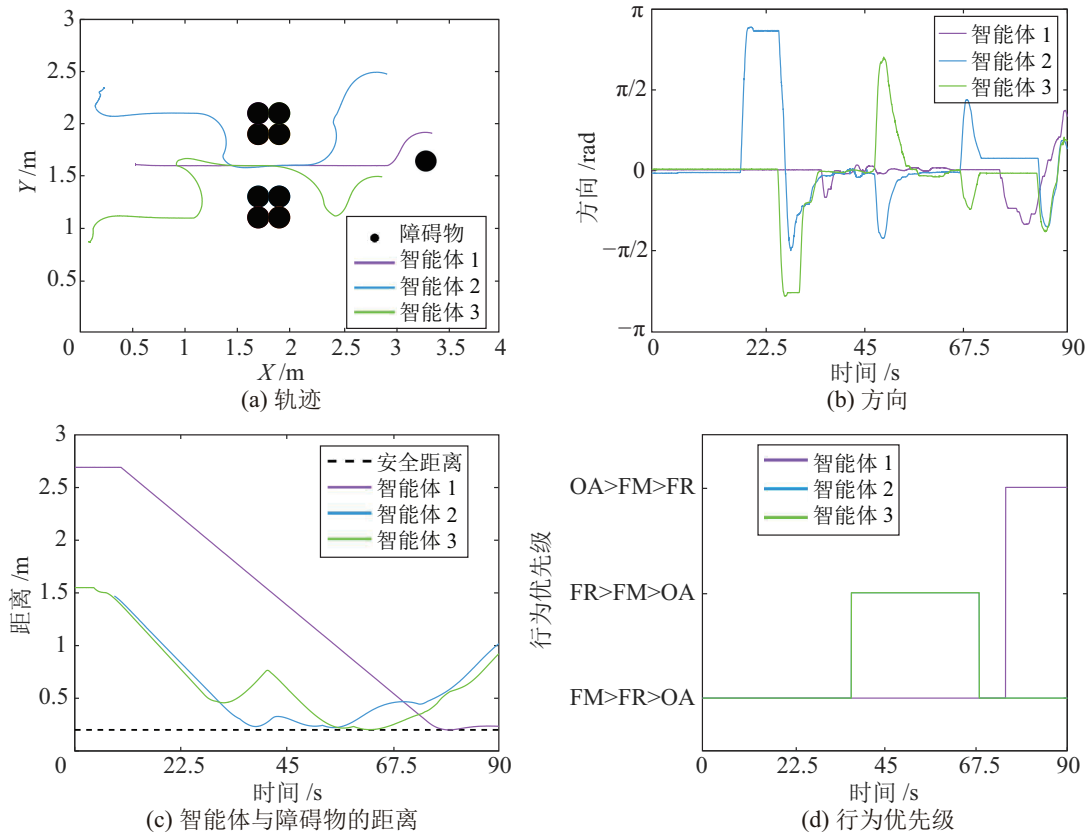


(a) 轨迹　(b) 方向

(c) 智能体与障碍物的距离　(d) 行为优先级

图 19　FSAMS 的实验结果

Fig.19　Experimental results of the FSAMS



(a) 轨迹 (b) 方向

(c) 智能体与障碍物的距离 (d) 行为优先级

图 20 MPCMS 的实验结果

Fig.20 Experimental results of the MPCMS

先级，旨在驱使系统形成期望的队形。图 17(b) 显示了多 AgileX Limo 机器人系统在执行任务 32 s 时，已经形成了期望的队形，且以编队形式向预定的目标位置移动。图 17(c) 显示了当多 AgileX Limo 机器人系统遭遇障碍物群时，重构了新的队形，以协同的方式避开路径上的障碍物。图 17(d) 显示了当多 AgileX Limo 机器人系统遭遇单个障碍物时，1 号 AgileX Limo 机器人会切换 OA 行为作为最高优先级以避障，而其他 AgileX Limo 机器人会维持编队。FSAMS 和 MPCMS 的实验结果分别如图 19 和图 20 所示。FSAMS 的行为优先级切换最为频繁，导致多差速机器人在队形切换时轨迹不光滑，以及在避障时违反安全规则。MPCMS 的实时性最差，导致多差速机器人无法在 90 s 的任务时间内移动至目标位置。相较于 FSAMS 和 MPCMS，本文 MARLMS 兼顾了行为优先级的切换性能和算法的实时性。实验结果验证了本文 MARLMS 的有效性、优越性和实用性。

图 15 是离线训练的结果图，是轨迹完美跟踪时的理想结果图。图 18 是实际 AgileX Limo 机器人系统的结果图，其中为了实验验证方便，本文使用了不依赖模型的 PID 控制器作为行为控制器以跟踪图 15 的结果。由于地面摩擦力和 AgileX Limo 机器人系统内部扰动的存在，跟踪性能不可能是完美的，因此图 18 的结果会比图 15 差。然而，图 18(d) 和图 15(d) 的行为优先级切换性能是一致的，因此 MARLMS 的最优行为优先级策略是有效的。

6 结论（Conclusion）

针对基于行为的多差速机器人系统提出了一个新颖的 MARLMS，通过学习一个联合的行为优先级策略，在任务执行过程中智能且动态地决策行为优先级。通过引入差速模型代替质点模型，提升了 NSBC 方法对于最小极值状态的鲁棒性，且更适用于多差速机器人系统。通过学习一个最优的联合行为优先级策略，不仅打破了单机器人系统只能进行独立学习的限制，允许更多学习者加入并通过合作完成全局行为，而且降低了人工设计行为优先级切换规则的负担和对高性能硬件平台在线计算和存储能力的依赖。未来的工作是将 NSBC 方法的任务层改进为分布式，且在拓扑结构限制下学习一组分布式的最优行为优先级策略，弱化 NSBC 方法的集中式，且提升可扩展性。

本文沿用经典 NSBC 法的框架，因此 MARLM-

S 的任务层采用集中模式、控制层采用分散模式，存在隐含集中模式的致命缺陷，极大地限制了方法的可扩展性。为了解决可扩展问题，分布式系统是最为常用的手段之一。然而，行为控制系统的分布式化，不是简单地使用分布式多智能体强化学习算法就可以解决的，因为其在任务层和控制层均是分布式的。分布式行为控制框架设计的难点包括协作式任务的分布式化、分布式行为优先级策略学习的强化学习问题建模、拓扑结构的切换问题和奖励函数的设计问题等。分布式任务监管器的设计问题存在诸多难点，这是未来的重点研究内容。

参考文献（References）

[1] HU J Q, ZHANG Y M, RAKHEJA S. Adaptive trajectory tracking for car-like vehicles with input constraints[J]. IEEE Transactions on Industrial Electronics, 2022, 69(3): 2801-2810.

[2] QIN B, YAN H C, ZHANG H, et al. Enhanced reduced-order extended state observer for motion control of differential driven mobile robot[J]. IEEE Transactions on Cybernetics, 2023, 53(2): 1299-1310.

[3] YU X, SU R. Decentralized circular formation control of nonholonomic mobile robots under a directed sensor graph[J]. IEEE Transactions on Automatic Control, 2023, 68(6): 3656-3663.

[4] 王伟嘉，郑雅婷，林国政，等. 集群机器人研究综述[J]. 机器人，2020，42(2)：232-256.
WANG W J, ZHENG Y T, LIN G Z, et al. Swarm robotics: A review[J]. Robot, 2020, 42(2): 232-256.

[5] GARATTONI L, BIRATTARI M. Autonomous task sequencing in a robot swarm[J]. Science Robotics, 2018, 3(20). DOI: 10.1126/scirobotics.aat0430.

[6] 李勇，李坤成，孙柏青，等. 智能体 Petri 网融合的多机器人－多任务协调框架[J]. 自动化学报，2021，47(8)：2029-2049.
LI Y, LI K C, SUN B Q, et al. Multi-robot-multi-task coordination framework based on the integration of intelligent agent and Petri net[J]. Acta Automatica Sinica, 2021, 47(8): 2029-2049.

[7] MUSIĆ S, HIRCHE S. Control sharing in human-robot team interaction[J]. Annual Reviews in Control, 2017, 44: 342-354.

[8] XU L, XU Q M, CHEN C L, et al. Efficient task-network scheduling with task conflict metric in time-sensitive networking[J]. IEEE Transactions on Industrial Informatics, 2024, 20(2): 1528-1538.

[9] 王峰，张衡，韩孟臣，等. 基于协同进化的混合变量多目标粒子群优化算法求解无人机协同多任务分配问题[J]. 计算机学报，2021，44(10)：1967-1983.
WANG F, ZHANG H, HAN M C, et al. Co-evolution based mixed-variable multi-objective particle swarm optimization for UAV cooperative multi-task allocation problem[J]. Chinese Journal of Computers, 2021, 44(10): 1967-1983.

[10] BROOKS R A. New approaches to robotics[J]. Science, 1991, 253(5025): 1227-1232.

[11] 王义萍，陈庆伟，胡维礼. 机器人行为选择机制综述[J]. 机器人，2009，31(5)：472-480.
WANG Y P, CHEN Q W, HU W L. A survey on robot behavior selection mechanism[J]. Robot, 2009, 31(5): 472-480.

[12] 居鹤华，崔平远，刘红云. 基于自主行为智能体的月球车运动规划与控制[J]. 自动化学报，2006，32(5)：704-712.
JU H H, CUI P Y, LIU H Y. Autonomous behavior agent-based lunar rover motion planning and control[J]. Acta Automatica Sinica, 2006, 32(5): 704-712.

[13] REZAEE H, ABDOLLAHI F. A decentralized cooperative control scheme with obstacle avoidance for a team of mobile robots [J]. IEEE Transactions on Industrial Electronics, 2014, 61(1): 347-354.

[14] MAC T T, COPOT C, DE KEYSER R, et al. MIMO fuzzy control for autonomous mobile robot[J]. Journal of Automation and Control Engineering, 2016, 4(1): 65-70.

[15] ANTONELLI G, CHIAVERINI S. Kinematic control of platoons of autonomous vehicles[J]. IEEE Transactions on Robotics, 2006, 22(6): 1285-1292.

[16] MUSCIO G, PIERRI F, TRUJILLO M A, et al. Coordinated control of aerial robotic manipulators: Theory and experiments [J]. IEEE Transactions on Control Systems Technology, 2018, 26(4): 1406-1413.

[17] HUANG J, ZHOU N, CAO M. Adaptive fuzzy behavioral control of second-order autonomous agents with prioritized missions: Theory and experiments[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9612-9622.

[18] WANG W J, LI C J, GUO Y N. Relative position coordinated control for spacecraft formation flying with obstacle/collision avoidance[J]. Nonlinear Dynamics, 2021, 104: 1329-1342.

[19] ZHOU N, CHENG X D, SUN Z Q, et al. Fixed-time cooperative behavioral control for networked autonomous agents with second-order nonlinear dynamics[J]. IEEE Transactions on Cybernetics, 2022, 52(9): 9504-9518.

[20] YAO P, WEI Y X, ZHAO Z Y. Null-space-based modulated reference trajectory generator for multi-robots formation in obstacle environment[J]. ISA Transactions, 2022, 123: 168-178.

[21] ZHENG C B, PANG Z H, WANG J X, et al. Null-space-based time-varying formation control of uncertain nonlinear second-order multi-agent systems with collision avoidance[J]. IEEE Transactions on Industrial Electronics, 2023, 70(10): 10476-10485.

[22] MARINO A, CACCAVALE F, PARKER L E, et al. Fuzzy behavioral control for multi-robot border patrol[C]//17th Mediterranean Conference on Control and Automation. Piscataway, USA: IEEE, 2009: 246-251.

[23] CHEN Y T, ZHANG Z Y, HUANG J. Dynamic task priority planning for null-space behavioral control of multi-agent systems[J]. IEEE Access, 2020, 8: 149643-149651.

[24] WANG W, GUO J Y, TIAN G Q, et al. Event-triggered intervention framework for UAV-UGV coordination systems[J]. Machines, 2021, 9(12). DOI: 10.3390/machines9120371.

[25] ZHANG Z Y, MO Z B, CHEN Y T, et al. Reinforcement learning behavioral control for nonlinear autonomous system [J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(9): 1561-1573.

[26] HUANG J, MO Z B, ZHANG Z Y, et al. Behavioral control task supervisor with memory based on reinforcement learning for human-multi-robot coordination systems[J]. Frontiers of Information Technology & Electronic Engineering, 2022, 23: 1174-1188.

（下转第 424 页）

[21] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2012: 3354-3361.

[22] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2015: 3061-3070.

[23] CHENG X L, ZHONG Y R, HARANDI M, et al. Hierarchical neural architecture search for deep stereo matching[C] //Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2020: 22158-22169.

[24] ZHANG Y M, CHEN Y M, BAI X, et al. Adaptive unimodal cost volume filtering for deep stereo matching[C]//AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI Press, 2020: 12926-12934.

[25] ZENG K, WANG Y N, ZHU Q, et al. Deep progressive fusion stereo network[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 25437-25447.

[26] LIU B Y, YU H M, LONG Y Q. Local similarity pattern and cost self-reassembling for deep stereo matching networks[C]// AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI Press, 2022: 1647-1655.

[27] SHAMSAFAR F, WOERZ S, RAHIM R, et al. MobileStereoNet: Towards lightweight deep networks for stereo matching [C]//IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, USA: IEEE, 2022: 677-686.

[28] XU H F, ZHANG J Y. AANet: Adaptive aggregation network for efficient stereo matching[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 1956-1965.

[29] DENG Y, XIAO J M, ZHOU S Z, et al. Detail preserving coarse-to-fine matching for stereo matching and optical flow[J]. IEEE Transactions on Image Processing, 2021, 30: 5835-5847.

[30] CHEN S Y, XIANG Z Y, QIAO C Y, et al. PGNet: Panoptic parsing guided deep stereo matching[J]. Neurocomputing, 2021, 463: 609-622.

[31] TANKOVICH V, HANE C, ZHANG Y D, et al. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2021: 14357-14367.

[32] YE X Q, SANG X Z, CHEN D, et al. Superpixel guided network for three-dimensional stereo matching[J]. IEEE Transactions on Computational Imaging, 2022, 8: 54-68.

作者简介：

范诗萌（1994 –），女，博士生。研究领域：机器视觉，3 维成像等。

孙 炜（1975 –），男，博士，教授。研究领域：机器视觉，机器人控制，3 维成像等。

（上接第 413 页）

[27] VAN DANG C, AHN H, KIM J W, et al. Collision-free navigation in human-following task using a cognitive robotic system on differential drive vehicles[J]. IEEE Transactions on Cognitive and Developmental Systems, 2023, 15(1): 78-87.

[28] CHEN Y, LI Z J, KONG H Y, et al. Model predictive tracking control of nonholonomic mobile robots with coupled input constraints and unknown dynamics[J]. IEEE Transactions on Industrial Informatics, 2019, 15(6): 3196-3205.

[29] WEI E M, LUKE S. Lenient learning in independent-learner stochastic cooperative games[J]. The Journal of Machine Learning Research, 2016, 17(1): 2914-2955.

作者简介：

张祯毅（1994 –），男，博士生。研究领域：强化学习，机器行为学，多机器人系统。

黄 捷（1983 –），男，博士，教授。研究领域：复杂系统控制与决策，集群机器人系统，5G+ 工业互联网理论与技术。