(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112528552 B
(45)授权公告日 2024.09.06

(21)申请号 202011149361.4

(22)申请日 2020.10.23

(65)同一申请的已公布的文献号
申请公布号 CN 112528552 A

(43)申请公布日 2021.03.19

(73)专利权人 洛阳银杏科技有限公司
地址 471000 河南省洛阳市中国(河南)自由贸易试验区洛阳片区涧西区蓬莱路2号洛阳国家大学科技园2幢1-501

(72)发明人 徐巧玉 姬周珂 李坤鹏 方梦娟 王军委

(74)专利代理机构 洛阳九创知识产权代理事务所(普通合伙) 41156
专利代理师 张龙

(51)Int.Cl.
G06F 30/27(2020.01)
G06N 3/092(2023.01)
G06N 3/084(2023.01)
G06N 3/0985(2023.01)
G06N 3/045(2023.01)

(56)对比文件
李鹤宇等.基于深度强化学习的机械臂控制方法.《系统仿真学报》.2019,第31卷(第11期),第2452-2457页.

审查员 赵倩

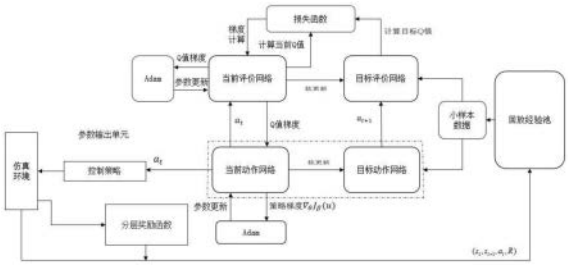
权利要求书2页 说明书8页 附图5页

(54)发明名称

一种基于深度强化学习的机械臂控制模型构建方法

(57)摘要

一种基于深度强化学习的机械臂控制模型构建方法,包括:基于真实机械臂构建仿真机械臂,真实机械臂和仿真机械臂均包括若干个关节;设定多个目标点,并且控制真实机械臂的末端向目标点移动,记录真实机械臂的真实结果参数,真实结果参数包括实际关节参数和末端位置参数;基于仿真机械臂构建选定深度强化学习算法;在深度强化学习算法中生成控制策略;基于目标点对深度强化学习算法进行训练;将训练过的深度强化学习算法中的控制策略输出为控制模型。本发明提供一种基于深度强化学习的机械臂控制模型构建方法,收敛速度快,并且生成的控制模型控制精度高。



CN 112528552 B

1.一种基于深度强化学习的机械臂控制模型构建方法,其特征在于:包括如下步骤:

S1、基于真实机械臂构建仿真机械臂,真实机械臂和仿真机械臂均包括若干个关节;

S2、设定多个目标点,并且控制真实机械臂的末端向目标点移动,记录真实机械臂的真实结果参数,真实结果参数包括实际关节参数和末端位置参数;

S3、基于仿真机械臂构建选定深度强化学习算法;S3中,深度强化学习算法包括DDPG(深度确定性策略梯度)智能体,DDPG智能体包括回放经验池、当前动作网络、目标动作网络、当前评价网络和目标评价网络,其中当前动作网络和目标动作网络用于生成控制策略,当前评价网络和目标评价网络用于生成评价值,当前动作网络与目标动作网络之间以及当前评价网络与目标评价网络之间均通过软更新方式传输参数;

S4、在深度强化学习算法中生成控制策略;S4中,生成控制策略的具体方法为:

S4.1、定义仿真机械臂的状态量$s=[x_1,y_1,z_1,x_2,y_2,z_2,a_1,a_2,\cdots,a_\lambda]$,其中$(x_1,y_1,z_1)$为仿真机械臂末端坐标,$(x_2,y_2,z_2)$为目标点坐标,$(a_1,a_2,\cdots,a_\lambda)$为仿真机械臂的关节参数,λ为真实机械臂和仿真机械臂的关节数,且有$1\leqslant\lambda\leqslant6$;

S4.2、当前动作网络生成控制策略$a_t=u(s_t|\theta^u)+N$,其中s_t为仿真机械臂的当前状态量,θ^u为当前动作网络的动作内参数,u(*)为当前动作网络的控制动作函数,N为随机噪声;

S4.3、当前动作网络将控制策略输出到仿真机械臂中对仿真机械臂进行控制;

S5、随机选取一个新的目标点;S6中,仿真结果参数包括仿真机械臂的结束状态量s_{t+1},结束状态量s_{t+1}为仿真机械臂按照控制策略动作之后的状态量;

S6、根据控制策略对仿真机械臂进行控制,获取仿真机械臂的仿真结果参数;

S7、将真实结果参数与仿真结果参数进行对比判断仿真结果参数是否符合精度要求,若符合则执行S8,若不符合则将根据真实结果参数对仿真结果参数进行修正生成奖励数据并且执行S9;S7中,奖励数据的计算方法为:

S7.1、根据当前目标点逆解出仿真机械臂的逆解关节参数$(b_1,b_2,\ldots,b_\lambda)$;

S7.2、计算仿真机械臂当前的关节参数相对于逆解关节参数的第一误差值$L_1=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-b_j)^2}$;

S7.3、计算仿真机械臂当前的关节参数相对于实际机械臂关节参数$(c_1,c_2,\ldots,c_\lambda)$的第二误差值$L_2=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-c_j)^2}$;

S7.4、计算仿真机械臂末端与目标点之间的距离值

$$d=\sqrt[2]{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2};$$

S7.5、计算奖励数据

$$R=\begin{cases}-L_1,d\geq0.1\\-d-0.1\lg(d)-L_2-0.1\lg(L_2),0.006\leq d<0.1\\R+3,0<d<0.006\end{cases};$$

S8、若存在未被选取过的目标点则返回S5,否则执行S11;

S9、将仿真结果参数和奖励数据输入到深度强化学习算法中;

S10、对深度强化学习算法进行训练,训练过程中深度强化学习算法对控制策略进行更

新,并且返回S6;

S11、将训练过的深度强化学习算法中的控制策略输出为控制模型。

2.如权利要求1所述的一种基于深度强化学习的机械臂控制模型构建方法,其特征在于:S9的具体方法为:

S9.1、构建小样本经验数据,小样本经验数据包括仿真结果参数和奖励数据;

S9.2、将小样本经验数据随机存入到回放经验池中;

S9.3、当回放经验池中的小样本经验数据数量超过设定阈值后,将回放经验池划分为多个区域,每个区域对应于一个目标点;

S9.4、根据目标点将小样本经验数据存入到对应的区域中。

3.如权利要求1所述的一种基于深度强化学习的机械臂控制模型构建方法,其特征在于:S10的具体方法为:

S10.1、目标动作网络和目标评价网络均从回放经验池提取一组小样本经验数据;

S10.2、目标评价网络生成a_{t+1}并且计算目标评价值$y_j=R_j+\gamma Q'(s_{j+1},u'(s_{j+1}|\theta^{u'})|\theta^{Q'})$,$s_{j+1}$为仿真机械臂的结束状态量参数,并且1≤j≤n,γ为衰减因子,$\theta^{u'}$为目标动作网络的动作内参数,u'(*)为目标动作网络的控制动作函数并且用于生成a_{t+1},$\theta^{Q'}$为目标评价网络的评价内参数,Q'(*)为目标评价网络的目标评价算子;S10.3、当前评价网络根据控制策略a_t计算当前评价值Q;

S10.4、将y_j和Q输入到损失函数中进行计算得到评价损失值$M=\frac{1}{n}\sum_j(y_j-Q(s_j,a_j|\theta^Q))^2$,其中n为迭代训练次数,1≤j≤n,Q(*)为当前评价网络的当前评价算子并且用于生成当前评价值Q,θ^Q为当前评价网络的评价内参数,s_j为当前状态量参数,a_j为控制策略参数;

S10.5、当前评价网络利用Adam算法对评价值梯度进行更新;

S10.6、当前评价网络将更新后的评价值梯度送入到当前动作网络中;

S10.7、当前动作网络根据评价值梯度计算策略梯度

$\nabla_{\theta^u}J_\beta(u)\approx\frac{1}{n}\sum_j(\nabla_a Q(s,a|\theta^u)|_{s=s_j,a=u(s_j)}\cdot\nabla_{\theta^u}u(s|\theta^u)|_{s=s_j})$,其中n为迭代训练次数,1≤j≤n,$\nabla_a Q(*)$为评价值梯度,$\nabla_{\theta^u}u(*)$为动作梯度,s=$s_j$为当前状态量参数,u(*)为当前动作网络的控制动作函数,a为当前状态下的策略动作参数,θ^u为当前动作网络的动作内参数;

S10.8、当前动作网络通过Adam算法对策略梯度进行更新。

一种基于深度强化学习的机械臂控制模型构建方法

技术领域

[0001] 本发明涉及自动控制技术领域,具体的说是一种基于深度强化学习的机械臂控制模型构建方法。

背景技术

[0002] 目前,随着工业领域大型装备使用需求增多,液压机械臂广泛应用于重型工件及设备的运输装卸等任务中,但由于液压机械臂内部结构复杂、质量重、体积大,其控制易受到惯性、摩擦等各方面因素的影响,因此液压机械臂的精确控制问题亟需解决。

[0003] 深度强化学习由于具备自适应学习的特点,目前许多研究者基于深度强化学习进行机械臂控制研究。郭宪等在“郭宪.基于DQN的机械臂控制策略的研究[D].北京交通大学,2018.”一文中提出一种基于深度Q学习(Deep-Q Learning,DQN)算法的机械臂控制策略,采用了引导式DQN算法的控制策略,为提高算法训练效率,在低精度要求的机械臂抓取任务中训练;卜令正等在“卜令正.基于深度强化学习的机械臂控制研究[D].中国矿业大学,2019.”一文中提出基于DDPG算法设计复合奖励函数来促进算法收敛,提高了机械臂在固定目标点抓取的精度,不足的是每个目标点需要分别进行训练;Gu等在“Gu S,Holly E,Lillicrap T,et al.Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates[C]//2017IEEE international conference on robotics and automation(ICRA).IEEE,2017:3389-3396.”一文中提出基于归一化优势函数(Normalized Advantage Function,NAF)算法在机械臂开门任务中训练,提高机械臂到达门把手的准确度,但该算法需要多个机械臂协同工作来促进算法收敛;Mahmood等在“Mahmood AR,Korenkevych D,Komer B J,et al.Setting up a reinforcement learning task with a real-world robot[C]//2018IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).IEEE,2018:4635-4640.”一文中提出采用信赖域策略优化(Trust Region Policy Optimization,TRPO)算法对UR5机械臂进行到达目标点训练,训练到了一定的末端控制精度,但没有往更高精度进行探索。在复杂环境下,目前深度强化学习机械臂控制方法对多个目标的精确控制能力较弱,并且无法做到收敛速度和控制精度的兼顾。

发明内容

[0004] 为了解决现有技术中的不足,本发明提供一种基于深度强化学习的机械臂控制模型构建方法,收敛速度快,并且生成的控制模型控制精度高。

[0005] 为了实现上述目的,本发明采用的具体方案为:一种基于深度强化学习的机械臂控制模型构建方法,包括如下步骤:

[0006] S1、基于真实机械臂构建仿真机械臂,真实机械臂和仿真机械臂均包括若干个关节;

[0007] S2、设定多个目标点,并且控制真实机械臂的末端向目标点移动,记录真实机械臂

的真实结果参数,真实结果参数包括实际关节参数和末端位置参数;

[0008] S3、基于仿真机械臂构建选定深度强化学习算法;

[0009] S4、在深度强化学习算法中生成控制策略;

[0010] S5、随机选取一个新的目标点;

[0011] S6、根据控制策略对仿真机械臂进行控制,获取仿真机械臂的仿真结果参数;

[0012] S7、将真实结果参数与仿真结果参数进行对比判断仿真结果参数是否符合精度要求,若符合则执行S8,若不符合则将根据真实结果参数对仿真结果参数进行修正生成奖励数据并且执行S9;

[0013] S8、若存在未被选取过的目标点则返回S5,否则执行S11;

[0014] S9、将仿真结果参数和奖励数据输入到深度强化学习算法中;

[0015] S10、对深度强化学习算法进行训练,训练过程中深度强化学习算法对控制策略进行更新,并且返回S6;

[0016] S11、将训练过的深度强化学习算法中的控制策略输出为控制模型。

[0017] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S3中,深度强化学习算法包括DDPG(深度确定性策略梯度)智能体,DDPG智能体包括回放经验池、当前动作网络、目标动作网络、当前评价网络和目标评价网络,其中当前动作网络和目标动作网络用于生成控制策略,当前评价网络和目标评价网络用于生成评价值,当前动作网络与目标动作网络之间以及当前评价网络与目标评价网络之间均通过软更新方式传输参数。

[0018] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S4中,生成控制策略的具体方法为:

[0019] S4.1、定义仿真机械臂的状态量$s=[x_1,y_1,z_1,x_2,y_2,z_2,a_1,a_2,\cdots,a_\lambda]$,其中$(x_1,y_1,z_1)$为仿真机械臂末端坐标,$(x_2,y_2,z_2)$为目标点坐标,$(a_1,a_2,\cdots,a_\lambda)$为仿真机械臂的关节参数,λ为真实机械臂和仿真机械臂的关节数,且有$1\leqslant\lambda\leqslant6$;

[0020] S4.2、当前动作网络生成控制策略$a_t=u(s_t|\theta^u)+N$,其中s_t为仿真机械臂的当前状态量,θ^u为当前动作网络的动作内参数,u(*)为当前动作网络的控制动作函数,N为随机噪声;

[0021] S4.3、当前动作网络将控制策略输出到仿真机械臂中对仿真机械臂进行控制。

[0022] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S6中,仿真结果参数包括仿真机械臂的结束状态量s_{t+1},结束状态量s_{t+1}为仿真机械臂按照控制策略动作之后的状态量。

[0023] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S7中,奖励数据的计算方法为:

[0024] S7.1、根据当前目标点逆解出仿真机械臂的逆解关节参数$(b_1,b_2,\cdots,b_\lambda)$;

[0025] S7.2、计算仿真机械臂当前的关节参数相对于逆解关节参数的第一误差值$L_1=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-b_j)^2}$;

[0026] S7.3、计算仿真机械臂当前的关节参数相对于实际机械臂关节参数$(c_1,c_2,\cdots,c_\lambda)$的第二误差值$L_2=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-c_j)^2}$;

[0027] S7.4、计算仿真机械臂末端与目标点之间的距离值

$$d=\sqrt[2]{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$；

[0028] S7.5、计算奖励数据

[0029] $$R=\begin{cases}-L_1, d\geq 0.1\\ -d-0.1\lg(d)-L_2-0.1\lg(L_2), 0.006\leq d<0.1\\ R+3, 0<d<0.006\end{cases}$$。

[0030] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S9的具体方法为:

[0031] S9.1、构建小样本经验数据,小样本经验数据包括仿真结果参数和奖励数据;

[0032] S9.2、将小样本经验数据随机存入到回放经验池中;

[0033] S9.3、当回放经验池中的小样本经验数据数量超过设定阈值后,将回放经验池划分为多个区域,每个区域对应于一个目标点;

[0034] S9.4、根据目标点将小样本经验数据存入到对应的区域中。

[0035] 作为上述基于深度强化学习的机械臂控制模型构建方法的进一步优化:S10的具体方法为:

[0036] S10.1、目标动作网络和目标评价网络均从回放经验池提取一组小样本经验数据;

[0037] S10.2、目标评价网络生成a_{t+1}并且计算目标评价值$y_j=R_j+\gamma Q'(s_{j+1},u'(s_{j+1}|\theta^{u'})|\theta^{Q'})$,$s_{j+1}$为仿真机械臂的结束状态量参数,并且1≤j≤n,γ为衰减因子,$\theta^{u'}$为目标动作网络的动作内参数,u'(*)为目标动作网络的控制动作函数并且用于生成a_{t+1},$\theta^{Q'}$为目标评价网络的评价内参数,Q'(*)为目标评价网络的目标评价算子;

[0038] S10.3、当前评价网络根据控制策略a_t计算当前评价值Q;

[0039] S10.4、将y_j和Q输入到损失函数中进行计算得到评价损失值$M=\frac{1}{n}\sum_j(y_j-Q(s_j,a_j|\theta^Q))^2$,其中n为迭代训练次数,1≤j≤n,Q(*)为当前评价网络的当前评价算子并且用于生成当前评价值Q,θ^Q为当前评价网络的评价内参数,s_j为当前状态量参数,a_j为控制策略参数;

[0040] S10.5、当前评价网络利用Adam算法对评价值梯度进行更新;

[0041] S10.6、当前评价网络将更新后的评价值梯度送入到当前动作网络中;

[0042] S10.7、当前动作网络根据评价值梯度计算策略梯度$\nabla_{\theta^u}J_\beta(u)\approx\frac{1}{n}\sum_j(\nabla_a Q(s,a|\theta^u)|_{s=s_j,a=u(s_j)}\cdot\nabla_{\theta^u}u(s|\theta^u)|_{s=s_j})$,其中n为迭代训练次数,1≤j≤n,$\nabla_a Q(*)$为评价值梯度,$\nabla_{\theta^u}u(*)$为动作梯度,s=$s_j$为当前状态量参数,u(*)为当前动作网络的控制动作函数,a为当前状态下的策略动作参数,θ^u为当前动作网络的动作内参数;

[0043] S10.8、当前动作网络通过Adam算法对策略梯度进行更新。

[0044] 有益效果:本发明收敛速度快,具有较强的抗干扰能力和自适应能力,并且生成的控制模型控制精度高。

附图说明

[0045] 图1是本发明中深度强化学习算法的原理图；

[0046] 图2是具体实施方式中所采用的中信机械臂的实体结构示意图；

[0047] 图3是具体实施方式中基于中信机械臂构建的仿真机械臂结构示意图；

[0048] 图4是实例中的训练曲线图；

[0049] 图5是平滑处理后的训练曲线图；

[0050] 图6是单点重复性测试结果曲线图；

[0051] 图7是多点误差测试结果曲线图。

具体实施方式

[0052] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0053] 请参阅图1,一种基于深度强化学习的机械臂控制模型构建方法,包括S1至S8。

[0054] S1、基于真实机械臂构建仿真机械臂,真实机械臂和仿真机械臂均包括若干个关节。在本实施例中,真实机械臂选择六关节机械臂,六个关节分别为大臂摆动关节、大臂俯仰关节、大臂伸缩关节、小臂摆动关节、小臂翻转关节和末端俯仰关节,属于常见的机械臂,不再赘述其结构。仿真环境可以采用ROS(Robot Operating System)。

[0055] S2、设定多个目标点,并且控制真实机械臂的末端向目标点移动,记录真实机械臂的真实结果参数,真实结果参数包括实际关节参数和末端位置参数。S2、设定多个目标点,并且控制真实机械臂的末端向目标点移动,记录真实机械臂的真实结果参数,真实结果参数包括实际关节参数和末端位置参数。目标点的坐标采用三维坐标表示。目标点的数量可以根据控制泛化性要求确定,控制泛化性要求越高,则目标点的数量越多,但是相应地会造成复杂度提升、耗时增多,因此需要根据实际情况灵活选择,例如在本实施例中目标点的数量设置为5000个。

[0056] S3、基于仿真机械臂构建选定强化学习算法。S3中,深度强化学习算法包括DDPG(深度确定性策略梯度)智能体,DDPG智能体包括回放经验池、当前动作网络、目标动作网络、当前评价网络和目标评价网络,其中当前动作网络和目标动作网络用于生成控制策略,当前评价网络和目标评价网络用于生成评价值,当前动作网络与目标动作网络之间以及当前评价网络与目标评价网络之间均通过软更新方式传输参数。

[0057] S4、在深度强化学习算法中生成控制策略。控制策略包括各个关节的控制动作增量,控制动作增量即关节的动作量,机械臂的各个关节按照控制动作增量动作即可使机械臂整体运动,实现对机械臂的控制。S4中,生成控制策略的具体方法为S4.1至S4.3。

[0058] S4.1、定义仿真机械臂的状态量$s=[x_1,y_1,z_1,x_2,y_2,z_2,a_1,a_2,\cdots,a_\lambda]$,其中$(x_1,y_1,z_1)$为仿真机械臂末端坐标,$(x_2,y_2,z_2)$为目标点坐标,$(a_1,a_2,\cdots,a_\lambda)$为仿真机械臂的关节参数,$\lambda$为真实机械臂和仿真机械臂的关节数,且有$1\leqslant\lambda\leqslant6$。

[0059] S4.2、当前动作网络生成控制策略$a_t=u(s_t|\theta^u)+N$,其中s_t为仿真机械臂的当前状态量,θ^u为当前动作网络的动作内参数,$u(*)$为当前动作网络的控制动作函数,N为随机噪

声。

[0060] S4.3、当前动作网络将控制策略输出到仿真机械臂中对仿真机械臂进行控制。

[0061] S5、随机选取一个新的目标点。需要说明的是,本发明中设定多个目标点用于提升训练模型的精确度,需要基于所有目标点对深度强化学习算法进行训练优化,因此每次选取的目标点都需要是不同的,这里新的目标点指未被选取过的目标点。

[0062] S6、根据控制策略对仿真机械臂进行控制,获取仿真机械臂的仿真结果参数。仿真结果参数包括仿真机械臂的结束状态量s_{t+1},结束状态量s_{t+1}为仿真机械臂按照控制策略动作之后的状态量。

[0063] S7、将真实结果参数与仿真结果参数进行对比判断仿真结果参数是否符合精度要求,若符合则执行S8,若不符合则将根据真实结果参数对仿真结果参数进行修正生成奖励数据并且执行S9。需要说明的是,在基于一部分目标点进行训练时,可能经过长时间的训练仿真结果参数仍然无法符合精度要求,如果继续进行训练会导致算法收敛速度缓慢,为了避免这一情况,可以设定一个训练次数的阈值,当基于某一个目标点的训练次数达到阈值时仿真结果参数仍然无法符合精度要求时忽略该目标点,重新返回S5。

[0064] S7中,奖励数据的计算方法为S7.1至S7.5。

[0065] S7.1、根据当前目标点逆解出仿真机械臂的逆解关节参数$(b_1,b_2,\cdots,b_\lambda)$。

[0066] S7.2、计算仿真机械臂当前的关节参数相对于逆解关节参数的第一误差值$L_1=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-b_j)^2}$。

[0067] S7.3、计算仿真机械臂当前的关节参数相对于实际机械臂关节参数$(c_1,c_2,\cdots,c_\lambda)$的第二误差值$L_2=\sqrt[2]{\sum_{i,j=1}^{\lambda}(a_i-c_j)^2}$。

[0068] S7.4、计算仿真机械臂末端与目标点之间的距离值

$$d=\sqrt[2]{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$$。

[0069] S7.5、计算奖励数据

[0070] $$R=\begin{cases}-L_1, d\geq 0.1\\ -d-0.1\lg(d)-L_2-0.1\lg(L_2), 0.006\leq d<0.1\\ R+3, 0<d<0.006\end{cases}$$。

[0071] 本发明生成奖励数据时采用三层奖励的方式,第一层奖励用于引导机械臂在目标点范围运动,当距离值d大于等于0.1m时,以当前关节值与逆解关节值的绝对值差的相反数作为奖励;第二层奖励以机械臂末端与目标点距离d的相反数作为距离奖励,以当前关节值与实际关节值的绝对值差的相反数作为关节奖励,当距离d在0.006m～0.1m之间时,进行距离奖励和关节奖励的组合,并引入log函数表达式作为非线性奖励;第三层奖励为精度奖励,当距离值d小于0.006m时给当前奖励数据加3,提高精度。通过三层奖励数据,能够有效地提升深度强化学习算法的收敛速度和对机械臂控制的精度。

[0072] S8、若存在未被选取过的目标点则返回S5,否则执行S11。为了进一步保证控制策略的精确度,本发明中将针对一个目标点的训练过程称为一个训练回合,当一个回合的控制策略训练完后,选择新的目标点进行进一步训练,从而不断地对深度强化学习算法进行

训练,同时,为了避免因为目标点选取的过于集中造成深度强化学习算法产生片面性,因此本发明中,一个训练回合结束后将当前目标点略过,并且从其余的目标点中随机选取一个进行下一个训练回合。当所有目标点都训练结束之后,即可结束训练过程,即跳转至S10。

[0073] S9、将仿真结果参数和奖励数据输入到深度强化学习算法中。S9中,将小样本经验数据存入到回放经验池的具体方法为S9.1至S9.3。

[0074] S9.1、构建小样本经验数据,小样本经验数据包括仿真结果参数和奖励数据。小样本经验数据还包括控制策略。

[0075] S9.2、将小样本经验数据随机存入到回放经验池中。

[0076] S9.3、当回放经验池中的小样本经验数据数量超过设定阈值后,将回放经验池划分为多个区域,每个区域对应于一个目标点。

[0077] S9.4、根据目标点将小样本经验数据存入到对应的区域中。

[0078] 在对深度强化学习算法进行训练的前期,为增加数据的多样性,将小样本经验数据随机存入到回放经验池中,相应地深度强化学习算法在从回放经验池中抽取小样本经验数据时采用随机抽样的方式;当回放经验池中的数据到达一定数量之后,将实时采集的数据以当前回合训练的目标点为中心进行区域存储,本回合训练时提高区域内数据的采样概率;下一回合更换目标点训练,重新将数据按照下一回合训练目标点为中心进行存储。以该方式进行数据存储和采样,减少了机械臂的无效采样行为,提高了采样效率。

[0079] S10、对深度强化学习算法进行训练,训练过程中深度强化学习算法对控制策略进行更新,并且返回S6。S10的具体方法为S10.1至S10.8。

[0080] S10.1、目标动作网络和目标评价网络均从回放经验池提取一组小样本经验数据。

[0081] S10.2、目标评价网络生成a_{t+1}并且计算目标评价值$y_j=R_j+\gamma Q'(s_{j+1},u'(s_{j+1}|\theta^{u'})|\theta^{Q'})$,$s_{j+1}$为仿真机械臂的结束状态量参数,并且1≤j≤n,γ为衰减因子,$\theta^{u'}$为目标动作网络的动作内参数,u'(*)为目标动作网络的控制动作函数并且用于生成a_{t+1},$\theta^{Q'}$为目标评价网络的评价内参数,Q'(*)为目标评价网络的目标评价算子。需要说明的是,控制动作函数基于小样本经验数据生成a_{t+1}。

[0082] S10.3、当前评价网络根据控制策略a_t计算当前评价值Q。

[0083] S10.4、将y_j和Q输入到损失函数中进行计算得到评价损失值$M=\frac{1}{n}\sum_j(y_j-Q(s_j,a_j|\theta^Q))^2$,其中n为迭代训练次数,1≤j≤n,Q(*)为当前评价网络的当前评价算子并且用于生成当前评价值Q,θ^Q为当前评价网络的评价内参数,s_j为当前状态量参数,a_j为控制策略参数。需要说明的是由评价损失值M可以得到评价值梯度。

[0084] S10.5、当前评价网络利用Adam算法对评价值梯度进行更新。

[0085] S10.6、当前评价网络将更新后的评价值梯度送入到当前动作网络中。

[0086] S10.7、当前动作网络根据评价值梯度计算策略梯度$\nabla_{\theta^u}J_\beta(u)\approx\frac{1}{n}\sum_j(\nabla_a Q(s,a|\theta^u)|_{s=s_j,a=u(s_j)}\cdot\nabla_{\theta^u}u(s|\theta^u)|_{s=s_j})$,其中n为迭代训练次数,1≤j≤n,$\nabla_a Q(*)$为评价值梯度,$\nabla_{\theta^u}u(*)$为动作梯度,s=$s_j$为当前状态量参数,u(*)为当前动作网络的控制动作函数,a为当前状态下的策略动作参数,θ^u为当前动作网络的动作内

参数。

[0087] S10.8、当前动作网络通过Adam算法对策略梯度进行更新。

[0088] S11、将训练过的深度强化学习算法中的控制策略输出为控制模型。在对所有目标点都训练完成之后生成深度强化学习算法控制模型,深度强化学习算法控制模型已经可以在外部干扰下对仿真机械臂进行精确控制,在工程实践中可以利用控制模型对真实机械臂进行控制。

[0089] 以下通过一个实例对本发明进行验证。

[0090] 在该实例中,选用中信重工机械臂作为真实机械臂,其实体结构如图2所示,相应地,真实结果参数中,实际关节参数由绝对值编码器获得,机械臂末端位置参数由全站仪获得。根据真实机械臂构建的仿真机械臂如图3所示,其机械臂数据可由仿真环境直接获得,仿真环境选择为ROS-Kinetic Gazebo-7.16。

[0091] 以传统的DDPG(深度确定性策略梯度)算法作为对比,在相同训练回合的条件下,以机械臂末端控制精度为指标进行训练对比,训练实验曲线如图4所示,其中虚线为改传统DDPG算法,实线为本发明,横坐标代表训练的回合数,纵坐标代表末端控制精度,纵坐标数值越大,精度越高。

[0092] 为将图4的曲线进行直观清晰的表示,对训练数据进行平滑处理,平滑后的训练曲线如图5所示。从图5可以看出,传统DDPG算法在3920回合左右收敛,本发明在3350回合左右收敛,收敛速度提升了16%。

[0093] 为验证本发明对机械臂的精确控制能力,设定末端位置控制精度要求为±6mm,对机械臂添加扰动噪声,扰动噪声为在机械臂末端添加的随机噪声,噪声大小范围为-0.005m～0.005m,分别进行了单点重复性误差的测试和多点误差的测试。

[0094] 为验证控制模型的单点重复控制性能,并模拟机械臂的实际工况,分别在仿真机械臂有扰动和无扰动的条件下进行25次的单点重复性误差测试实验。测试结果如图6所示,横坐标为重复测试的次数,纵坐标为末端位置控制误差大小,图中实线为不加扰动时的单点测试效果,虚线为加扰动以后的单点测试效果,相对应的单点测试结果如表1所示。

[0095] 表1单点重复性测试结果

[0096]

	重复次数	最大控制误差	最小控制误差	平均控制误差
未扰动	25	5.335mm	4.616mm	4.924mm
有扰动	25	6.146mm	4.780mm	5.411mm

[0097] 实验表明,本发明生成的控制模型具有很好的抗干扰能力,在不加扰动时末端的最大控制误差5.335mm,最小误差4.616mm,平均误差4.924mm;加扰动后末端最大控制误差6.146mm,最小误差4.780mm,平均误差5.411mm。

[0098] 为验证控制模型的多点控制性能,选取50组目标点,分别在无扰动和有扰动的环境下进行测试,测试现象如图7所示,图中实线为不加扰动时的多点测试效果,虚线为加扰动以后的多点测试效果,横坐标为测试的目标点数量,纵坐标为末端控制误差大小。

[0099] 多点误差测试结果表明,本发明生成的控制模型在不加扰动时对机械臂末端的最大控制误差为7.55mm,最小误差为4.78mm,平均误差为5.517mm;加扰动后对机械臂末端的最大控制误差为8.52mm,最小误差为5.08mm,平均误差为6.103mm。

[0100] 为验证模型的稳定性,以每50个随机目标点数据为一组进行测试,共测试6组,测

试结果如表2所示,控制误差在6mm以内为成功,完成率为成功次数与测试数量的比值。

[0101] 表2任务完成率测试

[0102]

组名	Test1	Test2	Test3	Test4	Test5	Test6
测试数量	50	50	50	50	50	50
成功次数	45	43	45	42	44	44
完成率/%	90	86	90	84	88	88

[0103] 实验表明,在多点稳定性测试中,本发明的任务完成率保持在80%以上,最高达到90%。

[0104] 综上所述,本发明收敛速度快,具有较强的抗干扰能力和自适应能力,并且生成的控制模型控制精度高。

[0105] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

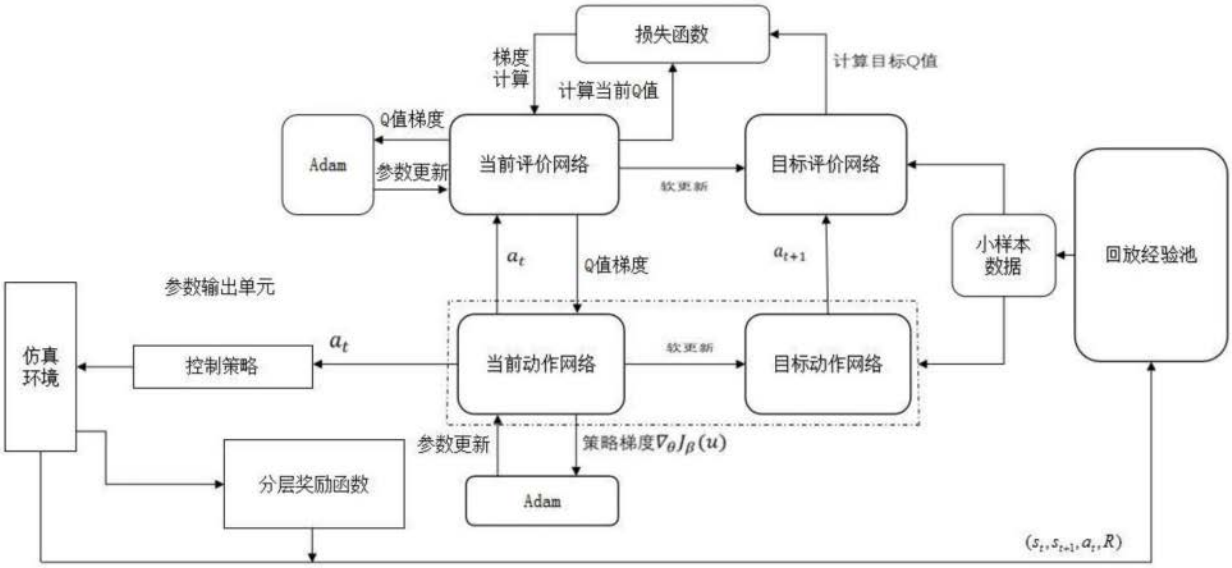

损失函数
梯度计算
计算当前Q值
计算目标Q值
Q值梯度
Adam
参数更新
当前评价网络
软更新
目标评价网络
小样本数据
回放经验池
a_t
Q值梯度
a_{t+1}
参数输出单元
仿真环境
控制策略
a_t
当前动作网络
软更新
目标动作网络
参数更新
策略梯度$\nabla_\theta J_\beta(u)$
分层奖励函数
Adam
(s_t, s_{t+1}, a_t, R)

图1

图2

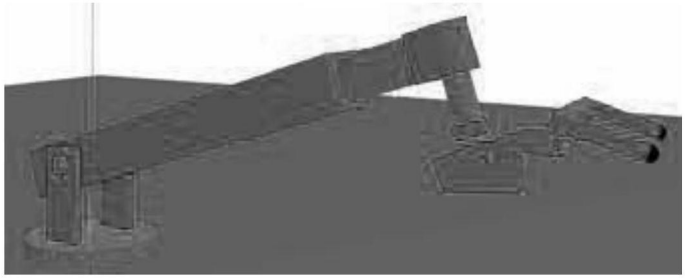


图3

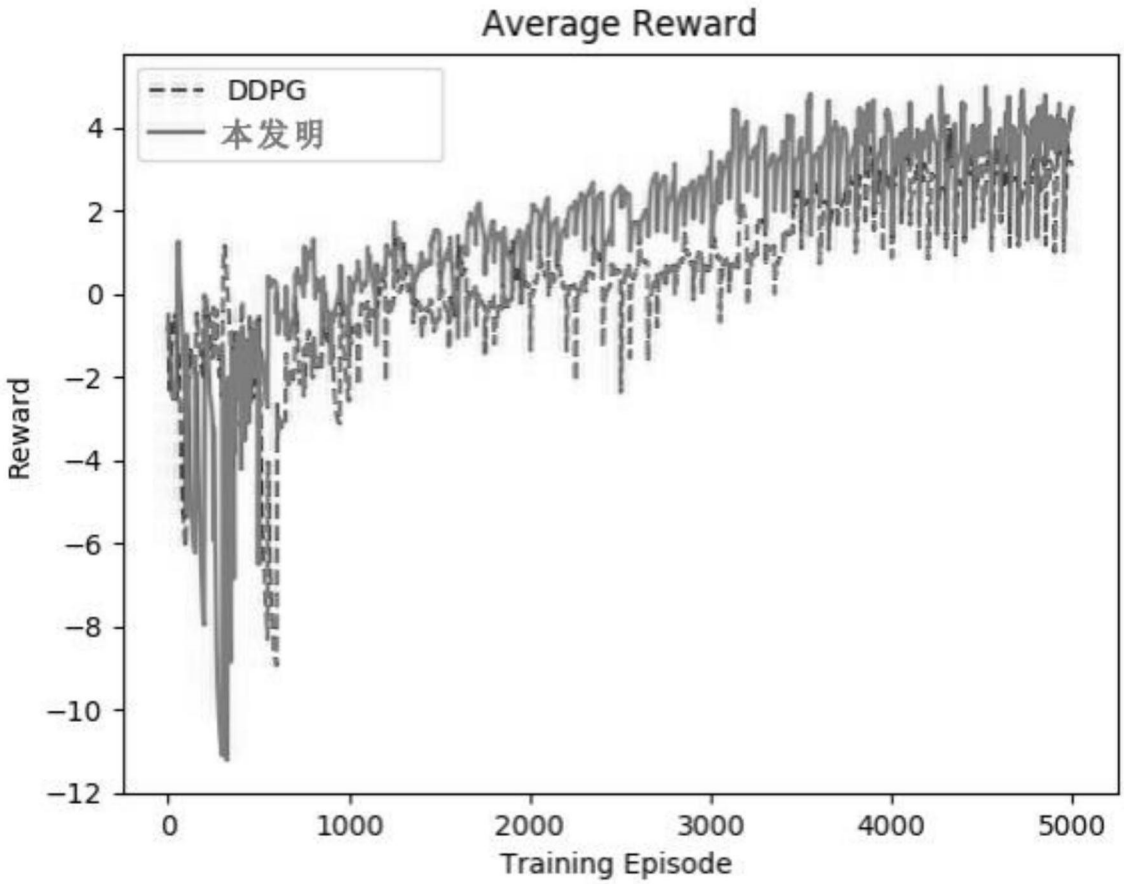

Average Reward
DDPG
本发明
Reward
Training Episode
4
2
0
−2
−4
−6
−8
−10
−12
0
1000
2000
3000
4000
5000

图4

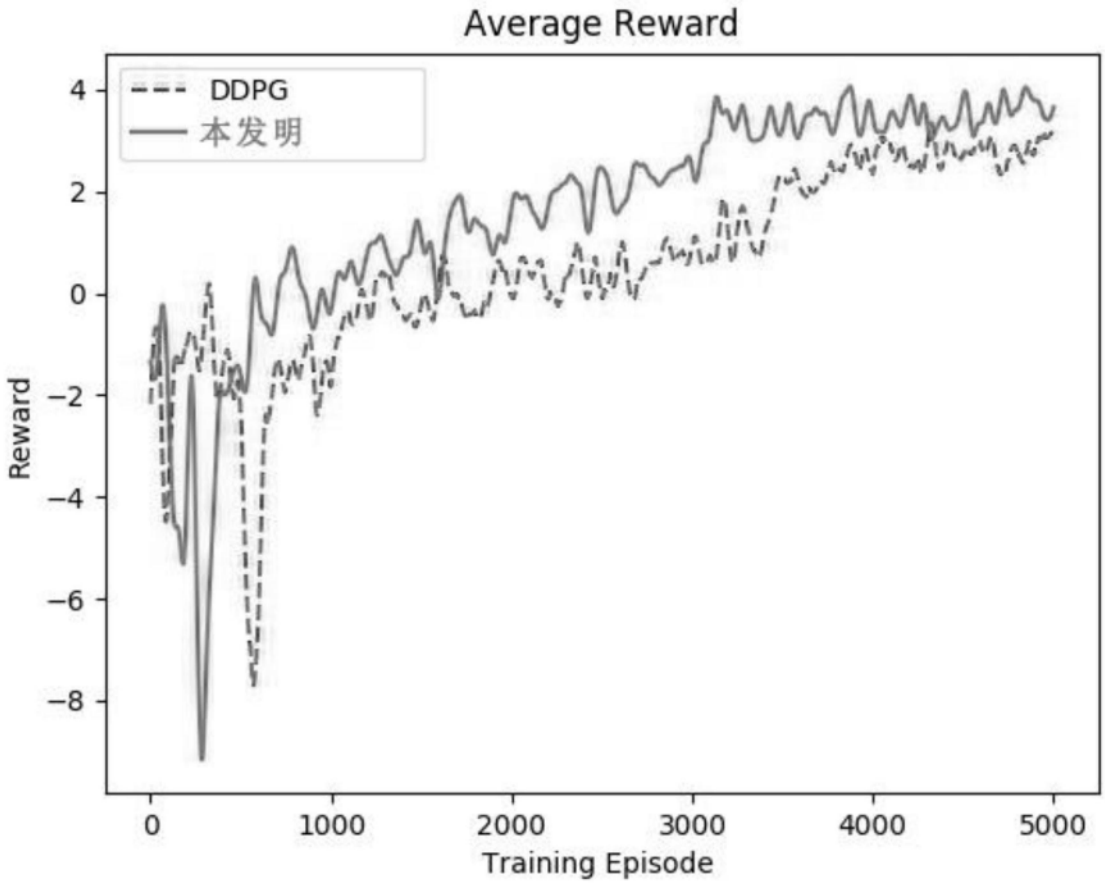

Average Reward
DDPG
本发明
Reward
Training Episode
4
2
0
−2
−4
−6
−8
0
1000
2000
3000
4000
5000

图5

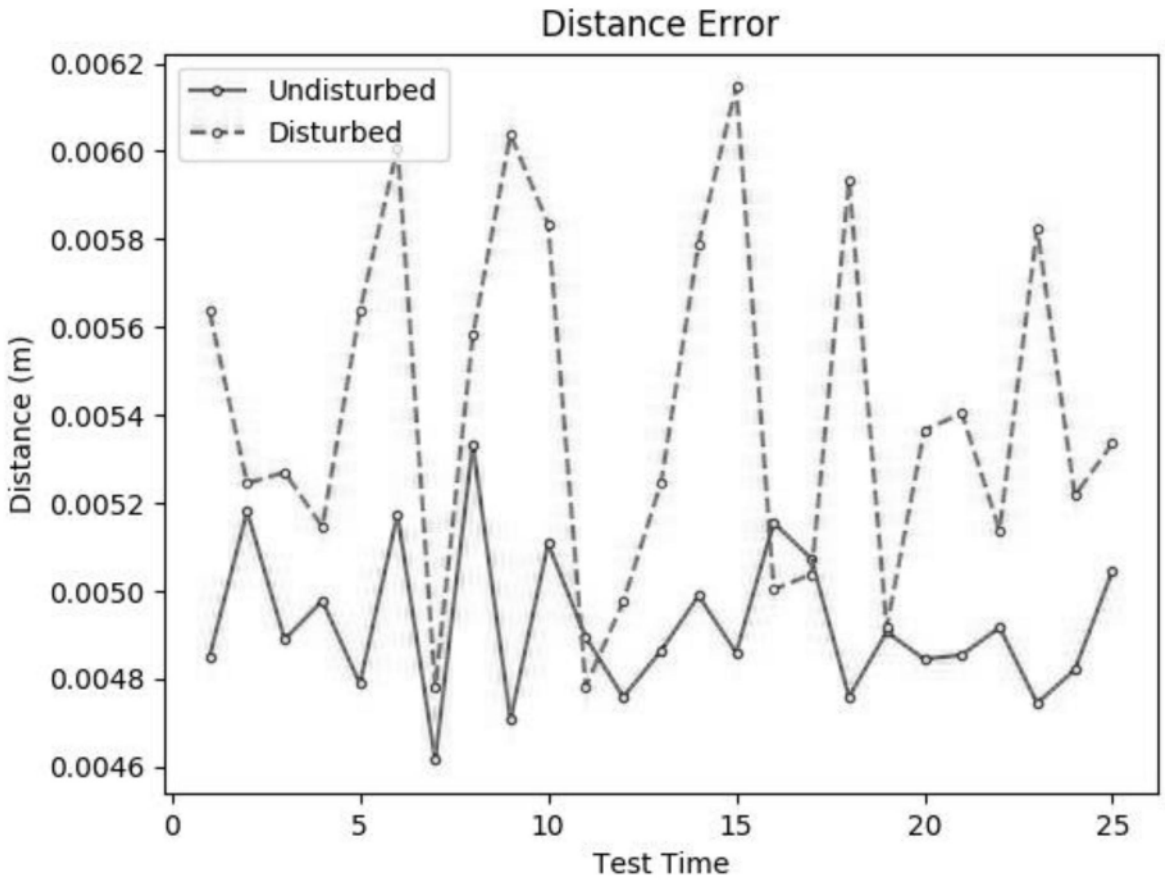

Distance Error
Undisturbed
Disturbed
Distance (m)
0.0062
0.0060
0.0058
0.0056
0.0054
0.0052
0.0050
0.0048
0.0046
0
5
10
15
20
25
Test Time

图6

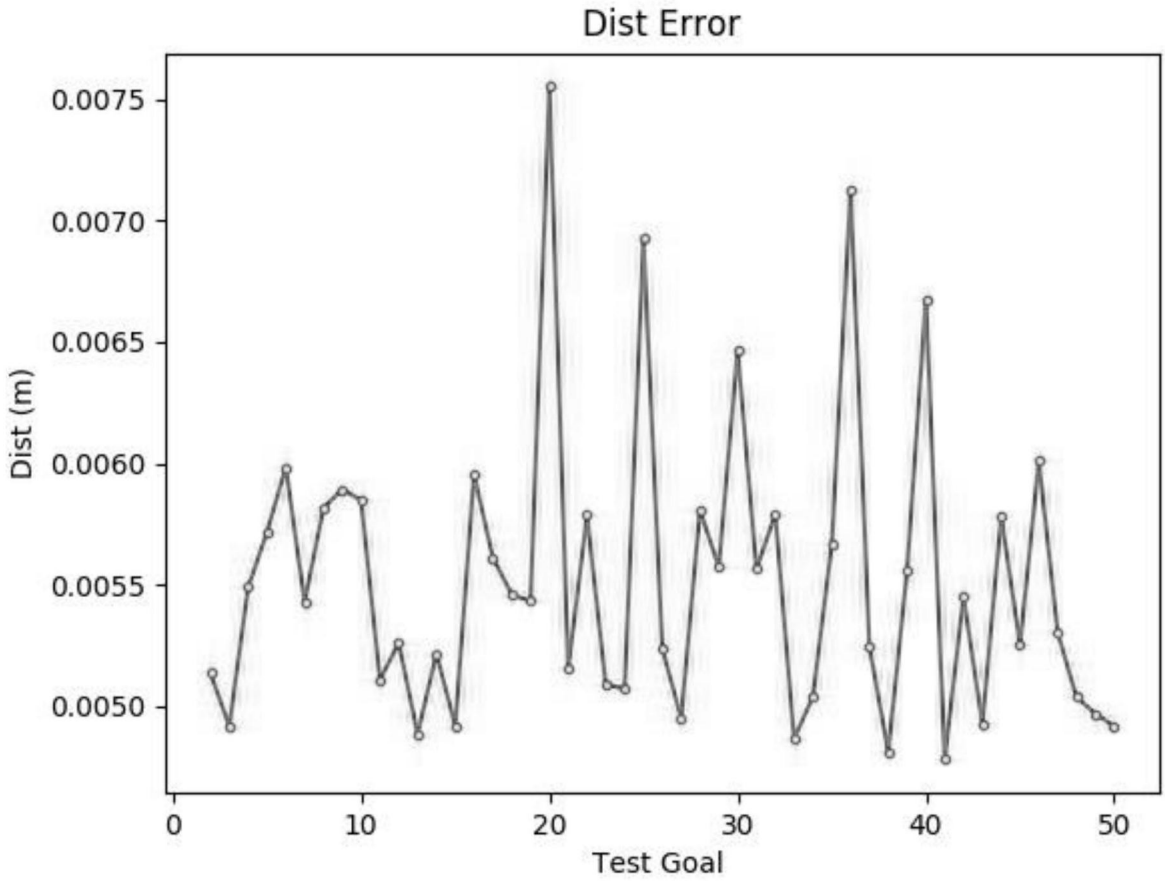

Dist Error
Dist (m)
0.0075
0.0070
0.0065
0.0060
0.0055
0.0050
0
10
20
30
40
50
Test Goal

图7