(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114879671 B
(45)授权公告日 2024.10.15

(21)申请号 202210477463.1

(22)申请日 2022.05.04

(65)同一申请的已公布的文献号
申请公布号 CN 114879671 A

(43)申请公布日 2022.08.09

(73)专利权人 哈尔滨工程大学
地址 150001 黑龙江省哈尔滨市南岗区南通大街145号哈尔滨工程大学科技处知识产权办公室

(72)发明人 王元慧 郝洋 张晓云 徐明 刘冲 谢可超 程基涛 鄂继洋 关一田 秦紫琦

(51)Int.Cl.
G05D 1/43(2024.01)
G05B 13/04(2006.01)

(56)对比文件
CN 106338919 A,2017.01.18
CN 108319138 A,2018.07.24

审查员 张慧

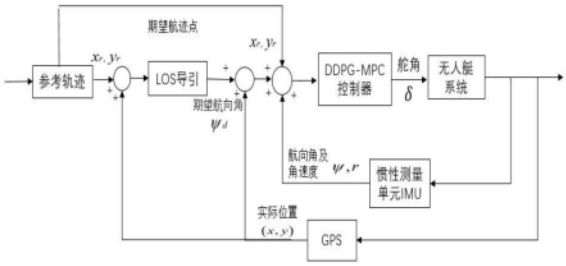
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种基于强化学习MPC的无人艇轨迹跟踪控制方法

(57)摘要

本发明属于水面无人艇轨迹跟踪控制技术领域,具体涉及一种基于强化学习MPC的无人艇轨迹跟踪控制方法。本发明在无人艇的MPC轨迹跟踪控制器设计过程中,选用无人艇的运动学模型和操纵响应模型作为预测模型,根据无人艇轨迹跟踪任务需求构造控制性能指标函数,在MPC滚动优化过程中利用强化学习的DDPG算法构建性能指标函数的求解器,通过最小化性能指标函数求解出轨迹跟踪的最优控制序列,最终将每时刻控制序列的第一个控制量作用于无人艇系统上。本发明提高了轨迹跟踪控制的鲁棒性和抗干扰,同时具备自学习能力,适应于复杂的海况环境,相较于传统的MPC控制算法其自主性和实时性更强,跟踪误差更小。



CN 114879671 B

1.一种基于强化学习MPC的无人艇轨迹跟踪控制方法,其特征在于,包括如下步骤:

步骤1:获取无人艇的实时状态信息(x,y,ψ,r),包括无人艇的位置坐标(x,y)、实时航向角ψ、艏摇角速度r;

步骤2:根据当前轨迹跟踪点$P_{k+1}(x_{k+1},y_{k+1})$和上一航迹跟踪点为$P_k(x_k,y_k)$,利用LOS导引算法计算无人艇实时的期望航向角ψ_d;

步骤3:建立无人艇轨迹跟踪的水平面数学模型,设计无人艇轨迹跟踪的DDPG-MPC智能控制器;

无人艇以恒定速度U航行,轨迹跟踪数学模型为:

$$\begin{cases}\dot{x}=U\cos\psi\\ \dot{y}=U\sin\psi\\ \dot{\psi}=r\\ \dot{r}=-\dfrac{r}{T}-\dfrac{\alpha}{T}r^3+\dfrac{K}{T}\delta\end{cases}$$

其中,T表示无人艇对舵的快速应答性和航向稳定性;K为增益系数;α为非线性系数;δ为操舵角;

在控制器设计中,状态变量χ=(x,y,ψ,r),输出量Y=(x,y,ψ),控制量u=δ;

离散状况下无人艇轨迹跟踪非线性系统的预测模型表示为:

χ(k+1)=f(χ(k),u(k),w(k))

其中,w(k)为系统扰动;f(·)为系统的非线性函数;

考虑k时刻对k+i时刻状态变量χ的预测值可表示为χ(k+i|k),其对应的系统输出值Y(k+i|k)=Cχ(k+i|k),k+i时刻输入系统参考轨迹为$Y_{ref}(k+i|k)$,作用于系统的控制量δ(k+i|k);

考虑k时刻开始由预测模型预测未来N个时刻的状态序列χ(k)、输出序列Y(k)、控制序列u(k)以及轨迹参考序列$Y_{ref}(k)$表示为:

$\chi(k)=(\chi(k+1|k),.....,\chi(k+N|k))^T$

$Y(k)=(Y(k+1|k),.....,Y(k+N|k))^T$

$u(k)=(u(k|k),.....,u(k+N-1|k))^T$

$Y_{ref}(k)=(Y_{ref}(k+1|k),.....,Y_{ref}(k+N|k))^T$

由此根据上述部分建立无人艇轨迹跟踪控制的性能指标:

$$J(k)=\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2+\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

其中,Q,R为性能指标函数的权值矩阵;

步骤4:利用DDPG算法求解MPC滚动优化过程中的最优控制序列,控制序列的第一个控制量作用于无人艇系统上。

2.根据权利要求1所述的一种基于强化学习MPC的无人艇轨迹跟踪控制方法,其特征在于:所述步骤4具体为:

步骤4.1:构建DDPG算法的Actor-Critic网络,包括4个网络结构:Actor策略网络$\mu(a|\theta^{\pi})$、Critic价值网络$Q(s,a|\theta^{Q})$、Actor目标策略网络$\mu(a|\theta^{\pi'})$、Critic目标价值网络Q(s,a|

$\theta^{Q'}$),当前网络和目标网络的网络结构一致;

步骤4.2:初始化网络网络模型参数θ^{π}、θ^{Q},当前网络的参数复制到目标网络;

$\theta^{\pi'}\leftarrow\theta^{\pi}$,$\theta^{Q'}\leftarrow\theta^{Q}$

步骤4.3:选择无人艇的状态序列$\chi(k)$作为DDPG算法的状态s_t,控制序列$u(k)$作为执行的动作a_t;选择无人艇轨迹跟踪控制的性能指标函数的负数作为DDPG算法的奖励回报;

$$r_t=-\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2-\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

步骤4.4:初始化无人艇的状态,根据当前无人艇的状态s_t,由策略网络根据当前的策略$\mu(a|\theta^{\pi})$给出状态s_t下的执行动作a_t,同时价值网络给出在状态s_t下执行动作a_t的价值$Q(s,a|\theta^{Q})$,将动作序列的第一个动作即控制序列的第一个控制量作用于无人艇系统上使其与环境交互,得到下一时刻的状态s_{t+1}并得到及时的奖励r_t,将每个过程产生的数据样本(s_t,a_t,r_t,s_{t+1})储存在经验池中用于训练策略网络和价值网络;

步骤4.5:从经验池中随机选取M个数据样本(s_i,a_i,r_i,s_{i+1}),对于价值网络的训练,构造价值网络训练的损失函数:

$$L=\frac{1}{M}\sum_i\left(r_i+\gamma Q\left(s_{i+1},\mu\left(s_{i+1}|\theta^{\pi'}\right)|\theta^{Q'}\right)-Q\left(s_i,a_i|\theta^{Q}\right)\right)^2$$

对于策略网络的训练采用策略梯度算法更新训练策略网络的参数,计算策略网络的策略梯度:

$$\nabla_{\theta^{\pi}}\mu\approx\frac{1}{M}\sum_i\left(\nabla_a Q\left(s,a|\theta^{Q}\right)|_{s=s_i,a=\mu(s_i)}\cdot\nabla_{\theta^{\pi}}\mu\left(s|\theta^{\pi}\right)|_{s=s_i}\right)$$

对于目标网络采用软更新的方式对其参数进行更新训练:

$\theta^{Q'}=\tau\theta^{Q}+(1-\tau)\theta^{Q'}$

$\theta^{\pi'}=\tau\theta^{\pi}+(1-\tau)\theta^{\pi'}$

步骤4.6:经过多次迭代训练,策略网络产生最优策略$\mu\left(s|\theta^{\pi^*}\right)$,将训练好的策略网络作为MPC滚动优化的求解器。

一种基于强化学习MPC的无人艇轨迹跟踪控制方法

技术领域

[0001] 本发明属于水面无人艇轨迹跟踪控制技术领域,具体涉及一种基于强化学习MPC的无人艇轨迹跟踪控制方法。

背景技术

[0002] 无人艇(USV)是一种无人操作的水面舰艇,具有自主性强,智能化高等特点,在军事和民用领域有着广泛的用途。如何保证无人艇在海面上安全地自主航行,是无人艇研究领域的重要课题。在实际的应用中,无人艇多工作于复杂海况,除受风浪流等海洋环境的干扰外,无人艇在航行过程中多遇到岛屿、暗礁、船只,浮标等障碍物,这些障碍物又分为静止的障碍物和运动的障碍物,都会产生无人艇在航行过程发生碰撞的风险。因此,在无人艇航行之前一般会进行路径规划,为无人艇规划出一条安全的最短航迹,使无人艇沿着预设的航迹自主地航行。

[0003] 无人艇多航行于复杂的海洋环境,易受风、浪、流的影响,加上无人艇多为欠驱动系统,非线性度较高,导致无人艇的轨迹跟踪控制变得异常复杂。从国内外的发展现状来看,无人艇轨迹跟踪控制一般采用滑模控制、反步法、神经网络PID、模糊PID,自抗扰控制等方法。随着人工智能的发展和进步,研究者们越来越重视将深度学习、强化学习、神经网络、群智能算法等智能算法融入到轨迹跟踪控制器的设计中来弥补当前控制算法的缺陷。

发明内容

[0004] 本发明的目的在于提供一种基于强化学习MPC的无人艇轨迹跟踪控制方法。

[0005] 一种基于强化学习MPC的无人艇轨迹跟踪控制方法,包括如下步骤:

[0006] 步骤1:获取无人艇的实时状态信息(x,y,ψ,r),包括无人艇的位置坐标(x,y)、实时航向角ψ、艏摇角速度r;

[0007] 步骤2:根据当前轨迹跟踪点$P_{k+1}(x_{k+1},y_{k+1})$和上一航迹跟踪点为$P_k(x_k,y_k)$,利用LOS导引算法计算无人艇实时的期望航向角ψ_d;

[0008] 步骤3:建立无人艇轨迹跟踪的水平面数学模型,设计无人艇轨迹跟踪的DDPG-MPC智能控制器;

[0009] 无人艇以恒定速度U航行,轨迹跟踪数学模型为:

[0010] $$\begin{cases}\dot{x}=U\cos\psi\\ \dot{y}=U\sin\psi\\ \dot{\psi}=r\\ \dot{r}=-\frac{r}{T}-\frac{\alpha}{T}r^3+\frac{K}{T}\delta\end{cases}$$

[0011] 其中,T表示无人艇对舵的快速应答性和航向稳定性;K为增益系数;α为非线性系数;δ为操舵角;

[0012] 在控制器设计中,状态变量χ=(x,y,ψ,r),输出量Y=(x,y,ψ),控制量u=δ;

[0013] 离散状况下无人艇轨迹跟踪非线性系统的预测模型表示为：

[0014] $\chi(k+1)=f(\chi(k),u(k),w(k))$

[0015] 其中,w(k)为系统扰动;f(·)为系统的非线性函数;

[0016] 考虑k时刻对k+i时刻状态变量χ的预测值可表示为$\chi(k+i|k)$,其对应的系统输出值$Y(k+i|k)=C\chi(k+i|k)$,k+i时刻输入系统参考轨迹为$Y_{ref}(k+i|k)$,作用于系统的控制量$\delta(k+i|k)$;

[0017] 考虑k时刻开始由预测模型预测未来N个时刻的状态序列χ(k)、输出序列Y(k)、控制序列u(k)以及轨迹参考序列$Y_{ref}(k)$表示为：

[0018] $\chi(k)=(\chi(k+1|k),.....,\chi(k+N|k))^T$

[0019] $Y(k)=(Y(k+1|k),.....,Y(k+N|k))^T$

[0020] $u(k)=(u(k|k),.....,u(k+N-1|k))^T$

[0021] $Y_{ref}(k)=(Y_{ref}(k+1|k),.....,Y_{ref}(k+N|k))^T$

[0022] 由此根据上述部分建立无人艇轨迹跟踪控制的性能指标：

[0023] $$J(k)=\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2+\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

[0024] 其中,Q,R为性能指标函数的权值矩阵;

[0025] 步骤4:利用DDPG算法求解MPC滚动优化过程中的最优控制序列,控制序列的第一个控制量作用于无人艇系统上。

[0026] 进一步地,所述步骤4具体为：

[0027] 步骤4.1:构建DDPG算法的Actor-Critic网络,包括4个网络结构:Actor策略网络$\mu(a|\theta^{\pi})$、Critic价值网络$Q(s,a|\theta^{Q})$、Actor目标策略网络$\mu(a|\theta^{\pi'})$、Critic目标价值网络$Q(s,a|\theta^{Q'})$,当前网络和目标网络的网络结构一致;

[0028] 步骤4.2:初始化网络网络模型参数θ^{π}、θ^{Q},当前网络的参数复制到目标网络;

[0029] $\theta^{\pi'}\leftarrow\theta^{\pi},\theta^{Q'}\leftarrow\theta^{Q}$

[0030] 步骤4.3:选择无人艇的状态序列χ(k)作为DDPG算法的状态s_t,控制序列u(k)作为执行的动作a_t;选择无人艇轨迹跟踪控制的性能指标函数的负数作为DDPG算法的奖励回报;

[0031] $$r_t=-\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2-\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

[0032] 步骤4.4:初始化无人艇的状态,根据当前无人艇的状态s_t,由策略网络根据当前的策略$\mu(a|\theta^{\pi})$给出状态s_t下的执行动作a_t,同时价值网络给出在状态s_t下执行动作a_t的价值$Q(s,a|\theta^{Q})$,将动作序列的第一个动作即控制序列的第一个控制量作用于无人艇系统上使其与环境交互,得到下一时刻的状态s_{t+1}并得到及时的奖励r_t,将每个过程产生的数据样本(s_t,a_t,r_t,s_{t+1})储存在经验池中用于训练策略网络和价值网络;

[0033] 步骤4.5:从经验池中随机选取M个数据样本(s_i,a_i,r_i,s_{i+1}),对于价值网络的训练,构造价值网络训练的损失函数：

[0034] $$L=\frac{1}{M}\sum_i\left(r_i+\gamma Q\left(s_{i+1},\mu\left(s_{i+1}|\theta^{\pi'}\right)|\theta^{Q'}\right)-Q\left(s_i,a_i|\theta^{Q}\right)\right)^2$$

[0035] 对于策略网络的训练采用策略梯度算法更新训练策略网络的参数,计算策略网络

的策略梯度：

[0036] $$\nabla_{\theta^{\pi}}\mu \approx \frac{1}{M}\sum_{i}\left(\nabla_{a}Q\left(s,a\mid\theta^{Q}\right)\mid_{s=si,a=\mu(si)}\bullet\nabla_{\theta^{\pi}}\mu\left(s\mid\theta^{\pi}\right)\mid_{s=si}\right)$$

[0037] 对于目标网络采用软更新的方式对其参数进行更新训练：

[0038] $\theta^{Q'}=\tau\theta^{Q}+(1-\tau)\theta^{Q'}$

[0039] $\theta^{\pi'}=\tau\theta^{\pi}+(1-\tau)\theta^{\pi'}$

[0040] 步骤4.6:经过多次迭代训练,策略网络产生最优策略$\mu\left(s\mid\theta^{\pi^{*}}\right)$,将训练好的策略网络作为MPC滚动优化的求解器。

[0041] 本发明的有益效果在于：

[0042] 本发明在无人艇的MPC轨迹跟踪控制器设计过程中,选用无人艇的运动学模型和操纵响应模型作为预测模型,根据无人艇轨迹跟踪任务需求构造控制性能指标函数,在MPC滚动优化过程中利用强化学习的DDPG算法构建性能指标函数的求解器,通过最小化性能指标函数求解出轨迹跟踪的最优控制序列,最终将每时刻控制序列的第一个控制量作用于无人艇系统上。本发明提高了轨迹跟踪控制的鲁棒性和抗干扰,同时具备自学习能力,适应于复杂的海况环境,相较于传统的MPC控制算法其自主性和实时性更强,跟踪误差更小。

附图说明

[0043] 图1为本发明的总体流程图。

[0044] 图2为LOS导引算法示意图。

[0045] 图3为强化学习DDPG算法实现步骤示意图。

具体实施方式

[0046] 下面结合附图对本发明做进一步描述。

[0047] 本发明提出一种基于强化学习MPC的无人艇轨迹跟踪控制方法,实现无人艇轨迹跟踪的自主控制。在无人艇的MPC轨迹跟踪控制器设计过程中,选用无人艇的运动学模型和操纵响应模型作为预测模型,根据无人艇轨迹跟踪任务需求构造控制性能指标函数,在MPC滚动优化过程中利用强化学习的DDPG算法构建性能指标函数的求解器,通过最小化性能指标函数求解出轨迹跟踪的最优控制序列,最终将每时刻控制序列的第一个控制量作用于无人艇系统上。本发明所提出的方法提高了轨迹跟踪控制的鲁棒性和抗干扰,同时具备自学习能力,适应于复杂的海况环境。

[0048] 一种基于强化学习MPC的无人艇轨迹跟踪控制方法,流程图如图1所示,主要包括以下步骤：

[0049] 步骤1.实时监测无人艇状态信息,通过GPS导航定位系统获取无人艇的位置坐标(x,y)、利用罗经检测无人艇实时航向角ψ、利用陀螺仪检测无人艇艏摇角速度r。

[0050] 步骤2.实时获取无人艇的状态信息(x,y,ψ,r),由当前轨迹跟踪点$P_{k+1}(x_{k+1},y_{k+1})$和上一航迹跟踪点为$P_{k}(x_{k},y_{k})$利用LOS导引算法计算无人艇实时的期望航向角ψ_{d}。详细步骤如图2所示：

[0051] 将无人艇当前位置(x,y)投影至期望轨迹记为$(x_{d}(\omega),y_{d}(\omega))$,其中ω为轨迹参

数,以该点为原点建立Serret-Frenet坐标系,沿期望轨迹的切线方向记为x_p轴,与惯性系坐标轴的纵轴方向的夹角记为轨迹方位角ψ_p,则$\psi_p=\arctan(y_d'(\omega)/x_d'(\omega))$。其中,$\psi_p\in[-\pi,\pi]$,$y_d'(\omega)=dy_d(\omega)/d\omega$,$x_d'(\omega)=dx_d(\omega)/d\omega$。

[0052] 轨迹参数更新率为:$\dot{\omega}=U\sqrt{\left(x_d'(\omega)\right)^2+\left(y_d'(\omega)\right)^2}>0$, $U=\sqrt{u^2+v^2}$为无人艇航行速度。

[0053] 期望航向点(x_{LOS},y_{LOS})与无人艇当前位置在期望轨迹上投影点间的距离记为Δ,称为前视距离,Δ=nL,n=2~10,L为无人艇的长度。目标点(x_{LOS},y_{LOS})相对于无人艇当前位置的方位角即为无人艇的实时期望航向角ψ_d。

[0054] $\psi_d=\psi_p+\arctan(-y_e/\Delta)$

[0055] 步骤3:建立无人艇轨迹跟踪的水平面数学模型,利用模型预测控制方法(MPC)结合强化学习的深度确定性策略梯度(DDPG)算法设计无人艇轨迹跟踪的DDPG-MPC智能控制器。

[0056] 无人艇一阶非线性操纵响应模型为:

[0057] $$T\dot{r}+r+ar^3=K\delta$$

[0058] T表示无人艇对舵的快速应答性和航向稳定性;K为增益系数;α为非线性系数;r为转艏角速度;δ为操舵角。

[0059] 由此得到无人艇轨迹跟踪模型:

[0060] $$\begin{cases}\dot{x}=u\cos\psi-v\sin\psi\\ \dot{y}=u\sin\psi+v\cos\psi\\ \dot{\psi}=r\\ \dot{r}=-\dfrac{r}{T}-\dfrac{\alpha}{T}r^3+\dfrac{K}{T}\delta\end{cases}$$

[0061] 考虑在实际航行中无人艇的纵向速度远远大于横向速度u>>v,横向速度v≈0,无人艇以恒定速度U航行,上述数学模型可简化为:

[0062] $$\begin{cases}\dot{x}=U\cos\psi\\ \dot{y}=U\sin\psi\\ \dot{\psi}=r\\ \dot{r}=-\dfrac{r}{T}-\dfrac{\alpha}{T}r^3+\dfrac{K}{T}\delta\end{cases}$$

[0063] 设置采样时间T_s经过离散化得到无人艇轨迹跟踪的预测模型:

[0064] $$\begin{cases}x(k+1)=x(k)+T_sU\cos(\psi(k))\\ y(k+1)=y(k)+T_sU\sin(\psi(k))\\ \psi(k+1)=\psi(k)+T_sr(k)\\ r(k+1)=r(k)+T_s\left(-\dfrac{r(k)}{T}-\dfrac{\alpha}{T}r^3(k)+\dfrac{K}{T}(\delta(k)+w(k))\right)\end{cases}$$

[0065] 由预测模型通过当前时刻的位置x(k),y(k),航向角ψ(k),角速度r(k)以及操舵角

δ(k)可以推测下一时刻的位置x(k+1),y(k+1),航向角ψ(k+1),角速度r(k+1)。

[0066] 式中,w(k)为系统的扰动变量。

[0067] 在控制器设计中,状态变量χ=(x,y,ψ,r),输出量Y=(x,y,ψ),控制量u=δ。

[0068] 离散状况下无人艇轨迹跟踪非线性系统的预测模型可表示为:

[0069] $\chi(k+1)=f(\chi(k),u(k),w(k))$

[0070] 考虑k时刻对k+i时刻状态变量χ的预测值可表示为χ(k+i|k),其对应的系统输出值Y(k+i|k)=Cχ(k+i|k),k+i时刻输入系统参考轨迹为$Y_{ref}(k+i|k)$,作用于系统的控制量δ(k+i|k)。

[0071] 考虑k时刻开始由预测模型预测未来N个时刻的状态序列χ(k)、输出序列Y(k)、控制序列u(k)以及轨迹参考序列$Y_{ref}(k)$可表示为:

[0072] $\chi(k)=(\chi(k+1|k),.....,\chi(k+N|k))^T$

[0073] $Y(k)=(Y(k+1|k),.....,Y(k+N|k))^T$

[0074] $u(k)=(u(k|k),.....,u(k+N-1|k))^T$

[0075] $Y_{ref}(k)=(Y_{ref}(k+1|k),.....,Y_{ref}(k+N|k))^T$

[0076] 由此根据上述部分可建立无人艇轨迹跟踪控制的性能指标:

[0077] $$J(k)=\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2+\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

[0078] 式中,Q,R为性能指标函数的权值矩阵。

[0079] 步骤4.如图3所示,利用强化学习的深度确定性策略梯度(DDPG)算法求解MPC滚动优化过程中的最优控制序列,控制序列的第一个控制量作用于无人艇系统上。

[0080] (1)构建DDPG算法的Actor-Critic网络,包括4个网络结构,Actor策略网络$\mu(a|\theta^{\pi})$,Critic价值网络$Q(s,a|\theta^{Q})$,Actor目标策略网络$\mu(a|\theta^{\pi'})$,Critic目标价值网络$Q(s,a|\theta^{Q'})$。当前网络和目标网络的网络结构一致。

[0081] (2)初始化网络网络模型参数θ^{π},θ^{Q},当前网络的参数复制到目标网络

[0082] $\theta^{\pi'}\leftarrow\theta^{\pi},\theta^{Q'}\leftarrow\theta^{Q}$。

[0083] (3)选择无人艇某时刻的状态序列χ(k)作为DDPG算法的状态s_t,控制序列u(k)作为执行的动作a_t。

[0084] (4)选择无人艇轨迹跟踪控制的性能指标函数的负数作为DDPG算法的奖励回报:

[0085] $$r_t=-\sum_{i=1}^{N}\|Y(k+i|k)-Y_{ref}(k+i|k)\|_Q^2-\sum_{i=0}^{N-1}\|u(k+i|k)\|_R^2$$

[0086] (5)初始化无人艇的状态,根据当前无人艇的状态s_t,由策略网络根据当前的策略$\mu(a|\theta^{\pi})$给出状态s_t下的执行动作a_t,同时价值网络给出在状态s_t下执行动作a_t的价值$Q(s,a|\theta^{Q})$,将动作序列的第一个动作即控制序列的第一个控制量作用于无人艇系统上使其与环境交互,得到下一时刻的状态s_{t+1}并得到及时的奖励r_t,将每个过程产生的数据样本(s_t,a_t,r_t,s_{t+1})储存在经验池中用于训练策略网络和价值网络。

[0087] (6)从经验池中随机选取M个数据样本(s_i,a_i,r_i,s_{i+1}),对于价值网络的训练,构造价值网络训练的损失函数:

[0088] $$L=\frac{1}{M}\sum_i\left(r_i+\gamma Q\left(s_{i+1},\mu\left(s_{i+1}|\theta^{\pi'}\right)|\theta^{Q'}\right)-Q\left(s_i,a_i|\theta^{Q}\right)\right)^2$$

[0089] 对于策略网络的训练采用策略梯度算法更新训练策略网络的参数,计算策略网络的策略梯度:

[0090] $$\nabla_{\theta^{\pi}}\mu \approx \frac{1}{M}\sum_{i}\left(\nabla_{a}Q\left(s,a\mid\theta^{Q}\right)\mid_{s=si,a=\mu(si)}\cdot\nabla_{\theta^{\pi}}\mu\left(s\mid\theta^{\pi}\right)\mid_{s=si}\right)$$

[0091] 对于目标网络采用软更新的方式对其参数进行更新训练:

[0092] $$\theta^{Q'}=\tau\theta^{Q}+(1-\tau)\theta^{Q'}$$

[0093] $$\theta^{\pi'}=\tau\theta^{\pi}+(1-\tau)\theta^{\pi'}$$

[0094] (7)经过多次迭代训练,策略网络产生最优策略$\mu\left(s\mid\theta^{\pi^{*}}\right)$,将训练好的策略网络作为MPC滚动优化的求解器。

[0095] 本发明所述的无人艇轨迹跟踪方法应用于无人艇的自主航行上,有效提高了无人艇轨迹跟踪过程中抗干扰性和鲁棒性,相较于传统的MPC控制算法其自主性和实时性更强,跟踪误差更小。

[0096] 以上所述仅为本发明的优选实施例而已,并不用于限制本发明,对于本领域的技术人员来说,本发明可以有各种更改和变化。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

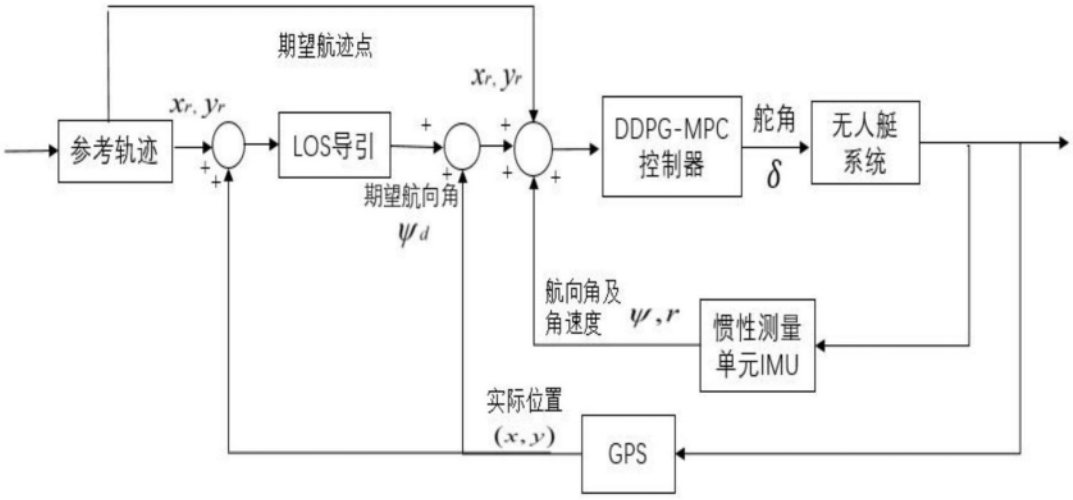

期望航迹点
x_r, y_r
参考轨迹
x_r, y_r
LOS导引
期望航向角
ψ_d
x_r, y_r
DDPG-MPC
控制器
舵角
δ
无人艇
系统
航向角及
角速度 ψ, r
惯性测量
单元IMU
实际位置
(x, y)
GPS

图1

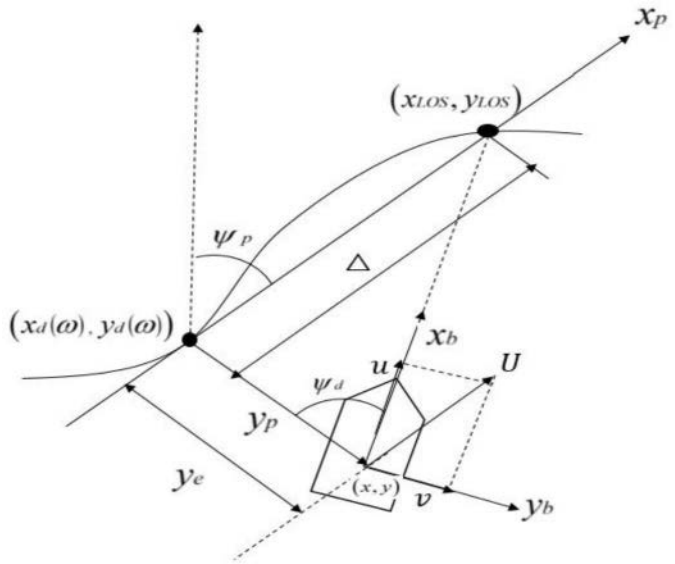

x_p
(x_{LOS}, y_{LOS})
ψ_p
Δ
$(x_d(\omega), y_d(\omega))$
x_b
u
U
ψ_d
y_p
y_e
(x, y)
v
y_b

图2

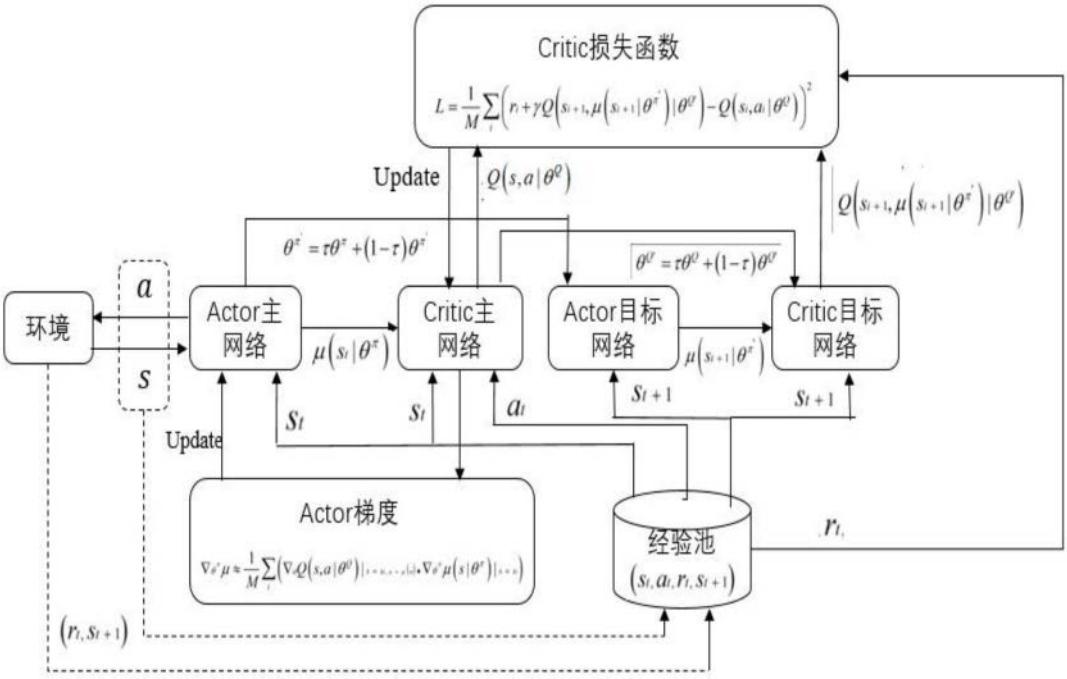

Critic损失函数
$L=\frac{1}{M}\sum_i\left(r_i+\gamma Q\left(s_{i+1},\mu\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right)-Q\left(s_i,a_i|\theta^{Q}\right)\right)^2$
Update
$Q\left(s,a|\theta^{Q}\right)$
$Q\left(s_{i+1},\mu\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right)$
$\theta^{\mu'}=\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$
$\theta^{Q'}=\tau\theta^{Q}+(1-\tau)\theta^{Q'}$
环境
a
s
Actor主网络
$\mu\left(s_t|\theta^{\mu}\right)$
Critic主网络
Actor目标网络
$\mu\left(s_{t+1}|\theta^{\mu'}\right)$
Critic目标网络
s_t
s_t
a_t
s_{t+1}
s_{t+1}
Update
Actor梯度
$\nabla_{\theta^{\mu}}\mu\approx\frac{1}{M}\sum_i\left(\nabla_a Q\left(s,a|\theta^{Q}\right)|_{s=s_i,a=\mu(s_i)}\nabla_{\theta^{\mu}}\mu\left(s|\theta^{\mu}\right)|_{s_i}\right)$
经验池
$\left(s_t,a_t,r_t,s_{t+1}\right)$
r_t
$\left(r_t,s_{t+1}\right)$

图3