(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110882542 A
(43)申请公布日 2020.03.17

(21)申请号 201911106673.4

(22)申请日 2019.11.13

(71)申请人 广州多益网络股份有限公司
地址 510000 广东省广州市黄埔区伴河路90号
申请人 广东利为网络科技有限公司
多益网络有限公司

(72)发明人 徐波

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 麦小婵 郝传鑫

(51)Int.Cl.
A63F 13/56(2014.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)

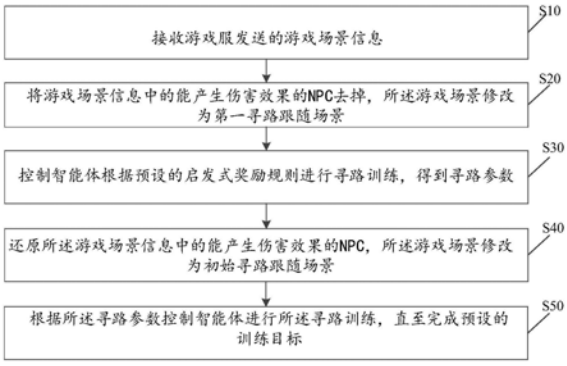
权利要求书3页 说明书9页 附图3页

(54)发明名称

游戏智能体的训练方法、装置、设备及存储介质

(57)摘要

本发明公开了游戏智能体的训练方法，包括：接收游戏服发送的游戏场景信息；将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景；控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景；根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标。本发明实施例还公开了一种游戏智能体的训练装置、设备及存储介质，采用多个实施例有效解决了现有技术强化学习训练效率低下，时间周期长的问题。



CN 110882542 A

1.一种游戏智能体的训练方法,其特征在于,包括:

接收游戏服发送的游戏场景信息;

将游戏场景信息中的能产生伤害效果的NPC去掉,所述游戏场景修改为第一寻路跟随场景;

控制智能体根据预设的启发式奖励规则进行寻路训练,得到寻路参数;其中,所述预设的启发式奖励规则为:当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息;

还原所述游戏场景信息中的能产生伤害效果的NPC,所述游戏场景修改为初始寻路跟随场景;

根据所述寻路参数控制智能体进行所述寻路训练,直至完成预设的训练目标。

2.如权利要求1所述的游戏智能体的训练方法,其特征在于,所述控制智能体根据预设的启发式奖励规则进行寻路训练,得到寻路参数;其中,所述预设的启发式奖励规则为:当智能体的当前位置与智能体的目标位置间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息,具体包括:

在所述第一寻路跟随场景中生成所有可以达到的地点,作为所述寻路训练的备用目标点;

从所述备用目标点中随机选择一个第一备用目标点,采用预设的策略梯度强化学习算法控制所述智能体以所述第一备用目标点为目标位置进行所述寻路训练;

当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息;

将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法,计算当前回合的最大奖励的梯度,通过梯度下降反向传播,得到最优策略;

根据最优策略进行训练得到最优寻路参数并进行保存。

3.如权利要求2所述的游戏智能体的训练方法,其特征在于,所述当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息,具体包括:

当所述智能体未到达所述目标位置时,每一帧的启发式奖励根据第一启发式奖励公式计算;其中,所述第一启发式奖励公式,具体为,$R=\alpha(t)*(D_{pre}-D_{now})-\beta$,$D_{pre}$为前一帧的智能体与目标位置的曼哈顿距离,$D_{now}$为在当前帧的智能体与目标位置的曼哈顿距离,$\alpha(t)$为随游戏帧数以预设的衰减规则不断减小的退火的因子,β为每一帧的惩罚因子;

当所述智能体到达目标位置后,直接获得一个预设的正向奖励。

4.如权利要求1所述的游戏智能体的训练方法,其特征在于,在所述接收游戏服发送的游戏场景信息之后,所述将游戏场景信息中的能产生伤害效果的NPC去掉,所述游戏场景修改为第一寻路跟随场景之前,所述方法还包括:

对所述游戏场景信息进行编码作为所述智能体的输入信息;

所述智能体根据所述输入信息输出对应的执行动作,并将对应的执行动作编码后反馈至所述游戏服。

5.如权利要求4所述的游戏智能体的训练方法,其特征在于,

对所述游戏场景信息编码方式,具体包括:将所述游戏场景信息绘制成对应的二位图

像；

将对应的执行动作编码后反馈至所述游戏服，具体包括：

采用one-hot方式对执行动作进行编码，得到输出行为数据；

将所述输出行为数据反馈至所述游戏服。

6.如权利要求2所述的游戏智能体的训练方法，其特征在于，所述将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法，计算当前回合的最大奖励的梯度，通过梯度下降反向传播，得到最优策略，具体包括：

所述预设的策略梯度强化学习算法包括：动作策略输出网络和价值估计网络；

根据以下公式计算算当前回合的最大奖励的梯度：

$$\hat{g}_k = \frac{1}{|\mathcal{D}_k|} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_k} \hat{A}_t$$

其中，$\mathcal{D}_k$为智能体和环境的交互的序列数据τ的集合，每条序列τ长度为最大长度为T，序列中每个时间节点包括状态s_t，动作a_t，当前动作策略下执行动作a_t的概率$\pi_\theta(a_t|s_t)$，该节点的价值估计$\hat{V}_\phi(s_t)$，执行该动作对应奖励r_t，$\hat{A}_t$为该节点的优势估计，$\hat{A}_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} + \gamma^{T-t} \hat{V}_\phi(s_T) - \hat{V}_\phi(s_t)$；$\hat{V}_\phi(s_t)$为在t时刻该节点的价值估计，γ为奖励折扣因子，

对策略输出网络参数进行更新的公式如下，

$$\theta_{k+1} = \theta_k + \alpha_k \hat{g}_k$$

其中，θ_k为第k次迭代的策略网络参数，α_k为策略网络的学习率，$\hat{g}_k$为每次神经网络损失函数反向传播的梯度；

对价值估计网络参数进行更新的公式如下，

$$\phi_{k+1} = \arg\min_\phi \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t\right)^2$$

其中，Φ_k为第k次迭代的价值网络参数，$\hat{R}_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} + \gamma^{T-t} \hat{V}_\phi(s_T)$，为时间t节点的实际状态价值。

7.如权利要求2所述的游戏智能体的训练方法，其特征在于，所述预游戏场景信息，包括：

地形信息、静止的NPC、随机游走的NPC以及游戏中的机关信息。

8.一种游戏智能体的训练装置，其特征在于，包括：

接收模块，用于接收游戏服发送的游戏场景信息；

第一游戏场景修改模块，用于将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景；

第一控制模块，用于控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息；

第二游戏场景修改模块，用于还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景；

第二控制模块，用于根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标。

9.一种游戏智能体的训练设备，其特征在于，包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如权利要求1至7中任意一项所述的游戏智能体的训练方法。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质包括存储的计算机程序，其中，在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如权利要求1至7中任意一项所述的游戏智能体的训练方法。

游戏智能体的训练方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术领域，尤其涉及一种游戏智能体的训练方法、装置、设备及存储介质。

背景技术

[0002] 传统的寻路跟随算法包括复制轨迹和重新规划路线等方法，这些方法在简单的游戏场景可以取得很好的效果。但是，随着对游戏场景的环境越来越复杂，要求这些算法能迅速响应这些复杂环境的变化，这些传统的寻路跟随算法越来越无法满足对应的要求。

[0003] 随着深度学习技术的爆发，基于深度学习的强化学习(Reinforcement Learning)技术也得到了飞速发展。深度强化学习技术可以利用不断地试错和学习以及对人类行为的模仿，自发地产生对环境的响应行为，避免了人工设计规则的问题。强化学习利用智能体和环境的不断交互，不断获得回报，通过最大化回报的方式进行学习，目前在游戏中获得比较理想的效果。然而，强化学习的过程需要不断与环境交互，不断地试错，尤其在环境的反馈比较稀疏时候，智能体获得反馈频率很低的情况下，强化学习训练需要耗费大量的时间，训练效率非常低下。游戏跟随正是这样反馈频率很低的场合，需要智能体一直探索试错，直到到达目的地，才能得到正向的反馈。如果路途中间踩到陷阱或者遇到致命的静止和移动的NPC，没有及时躲避可能直接结束回合，无法获得正向反馈。在复杂的游戏场景中通过随机探索试错到达目的地的概率极其低，进而造成强化学习训练效率低下，时间周期长。

发明内容

[0004] 本发明实施例提供一种游戏智能体的训练方法、装置、设备及存储介质，能有效解决现有技术强化学习训练效率低下，时间周期长的问题。

[0005] 本发明一实施例提供一种游戏智能体的训练方法，包括：

[0006] 接收游戏服发送的游戏场景信息；

[0007] 将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景；

[0008] 控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息；

[0009] 还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景；

[0010] 根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标。

[0011] 作为上述方案的改进，所述控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息，具体包括：

[0012] 在所述第一寻路跟随场景中生成所有可以达到的地点，作为所述寻路训练的备用

目标点；

[0013] 从所述备用目标点中随机选择一个第一备用目标点，采用预设的策略梯度强化学习算法控制所述智能体以所述第一备用目标点为目标位置进行所述寻路训练；

[0014] 当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息；

[0015] 将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法，计算当前回合的最大奖励的梯度，通过梯度下降反向传播，得到最优策略；

[0016] 根据最优策略进行训练得到最优寻路参数并进行保存。

[0017] 作为上述方案的改进，所述当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息，具体包括：

[0018] 当所述智能体未到达所述目标位置时，每一帧的启发式奖励根据第一启发式奖励公式计算；其中，所述第一启发式奖励公式，具体为，$R=\alpha(t)*(D_{pre}-D_{now})-\beta$，$D_{pre}$为前一帧的智能体与目标位置的曼哈顿距离，$D_{now}$为在当前帧的智能体与目标位置的曼哈顿距离，$\alpha(t)$为随游戏帧数以预设的衰减规则不断减小的退火的因子，β为每一帧的惩罚因子；

[0019] 当所述智能体到达目标位置后，直接获得一个预设的正向奖励。

[0020] 作为上述方案的改进，在所述接收游戏服发送的游戏场景信息之后，所述将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景之前，所述方法还包括：

[0021] 对所述游戏场景信息进行编码作为所述智能体的输入信息；

[0022] 所述智能体根据所述输入信息输出对应的执行动作，并将对应的执行动作编码后反馈至所述游戏服。

[0023] 作为上述方案的改进，对所述游戏场景信息编码方式，具体包括：将所述游戏场景信息绘制成对应的二位图像；

[0024] 将对应的执行动作编码后反馈至所述游戏服，具体包括：

[0025] 采用one-hot方式对执行动作进行编码，得到输出行为数据；

[0026] 将所述输出行为数据反馈至所述游戏服。

[0027] 作为上述方案的改进，所述将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法，计算当前回合的最大奖励的梯度，通过梯度下降反向传播，得到最优策略，具体包括：

[0028] 所述预设的策略梯度强化学习算法包括：动作策略输出网络和价值估计网络；

[0029] 根据以下公式计算算当前回合的最大奖励的梯度：

[0030] $$\hat{g}_k = \frac{1}{|\mathcal{D}_k|}\sum_{\tau\in\mathcal{D}_k}\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_k}\hat{A}_t$$

[0031] 其中，$\mathcal{D}_k$为智能体和环境的交互的序列数据τ的集合，每条序列τ长度为最大长度为T，序列中每个时间节点包括状态s_t，动作a_t，当前动作策略下执行动作a_t的概率$\pi_\theta(a_t|s_t)$，该节点的价值估计$\hat{V}_\phi(s_t)$，执行该动作对应奖励r_t，$\hat{A}_t$为该节点的优势估计，

$\widehat{A}_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} + \gamma^{T-t} \widehat{V}_\phi(s_T) - \widehat{V}_\phi(s_t)$；$\widehat{V}_\phi(s_t)$为在t时刻该节点的价值估计，$\gamma$为奖励折扣因子，

[0032] 对策略输出网络参数进行更新的公式如下，

[0033] $$\theta_{k+1} = \theta_k + \alpha_k \hat{g}_k$$

[0034] 其中，θ_k为第k次迭代的策略网络参数，α_k为策略网络的学习率，$\hat{g}_k$为每次神经网络损失函数反向传播的梯度；

[0035] 对价值估计网络参数进行更新的公式如下，

[0036] $$\phi_{k+1} = \operatorname{argmin}_\phi \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \widehat{R}_t\right)^2$$

[0037] 其中，ϕ_k为第k次迭代的价值网络参数，$\hat{R}_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} + \gamma^{T-t} \hat{V}_\phi(s_T)$，为时间t节点的实际状态价值。

[0038] 作为上述方案的改进，所述预游戏场景信息，包括：

[0039] 地形信息、静止的NPC、随机游走的NPC以及游戏中的机关信息。

[0040] 本发明另一实施例对应提供了一种游戏智能体的训练装置，包括：

[0041] 接收模块，用于接收游戏服发送的游戏场景信息；

[0042] 第一游戏场景修改模块，用于将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景；

[0043] 第一控制模块，用于控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息；

[0044] 第二游戏场景修改模块，用于还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景；

[0045] 第二控制模块，用于根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标。

[0046] 本发明另一实施例提供了一种游戏智能体的训练设备，包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现上述发明实施例所述的游戏智能体的训练方法。

[0047] 本发明另一实施例提供了一种存储介质，所述计算机可读存储介质包括存储的计算机程序，其中，在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行上述发明实施例所述的游戏智能体的训练方法。

[0048] 与现有技术相比，本发明实施例公开的游戏智能体的训练方法、装置、设备及存储介质，通过接收游戏服发送的游戏场景信息，将游戏场景信息中的能产生伤害效果的NPC去掉得到第一寻路跟随场景，在第一寻路跟随场景中控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数，当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息，再将游戏场景修改为初始寻路跟随场景，

根据寻路参数控制智能体再次进行寻路训练，直至完成预设的训练目标。由上分析可知，通过智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息，给予智能体一个启发式的方向，提高寻路效率，从而提高学习训练效率，减少时间周期。

附图说明

[0049] 图1是本发明训练服与游戏服交互的示意图。

[0050] 图2是本发明一实施例提供的一种游戏智能体的训练方法的流程示意图；

[0051] 图3是本发明一实施例提供的智能体当前位置和目标位置启发式奖励示意图；

[0052] 图4是本发明一实施例提供的一种游戏智能体的训练装置的结构示意图；

[0053] 图5是本发明一实施例提供的一种游戏智能体的训练设备的结构示意图。

具体实施方式

[0054] 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作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。

[0055] 参见图1，在游戏服和训练服之间创建通信连接，游戏服创建游戏环境，实现游戏环境部分的逻辑，训练服负责游戏训练部分逻辑。游戏服把每一帧的游戏数据发送到训练服，训练服对每一帧的数据进行分析训练，返回每一帧对应的动作到游戏服。游戏服创建的环境把每一帧的当前环境的状态(即游戏场景信息)发给训练服，训练服的智能体对环境进行决策分析，得到这一帧应该执行的动作返回给游戏服，游戏服执行这个动作后，把下一帧的状态和执行这个动作对应的奖励信息返回给训练服，一直循环上述操作。

[0056] 参见图2，是本发明一实施例提供的一种游戏智能体的训练方法的流程示意图。

[0057] 本发明实施例提供了一种游戏智能体的训练方法，包括：

[0058] S10，接收游戏服发送的游戏场景信息。其中，游戏场景信息包括：地形信息、静止的NPC、随机游走的NPC以及游戏中的机关信息。

[0059] 具体地，训练服接收游戏服发送的游戏场景信息，并对游戏场景信息进行处理。

[0060] S20，将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景。在本实施例中，产生伤害效果的NPC包括：游戏环境中随机游走的致命怪物，静止的致命怪物，跳跃通过河流时候的滚石和机关等。

[0061] 具体地，把对智能体具有伤害效果的NPC逻辑删除，形成一个没有伤害机制的第一寻路跟随场景。

[0062] S30，控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息。

[0063] 具体地，S301，在所述第一寻路跟随场景中生成所有可以达到的地点，作为所述寻路训练的备用目标点。

[0064] S302，从所述备用目标点中随机选择一个第一备用目标点，采用预设的策略梯度

强化学习算法控制所述智能体以所述第一备用目标点为目标位置进行所述寻路训练。

[0065] 在本实施例中,采用策略梯度强化学习算法进行寻路训练,从生成的所有可以到达的点随机出一个备用目标点作为目标位置,智能体随机初始化在任意可以到达的位置,达到目标位置则完成任务,超时没有到达则任务失败。

[0066] S303,当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息。从而提高寻路效率,减少时间周期。

[0067] S304,将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法,计算当前回合的最大奖励的梯度,通过梯度下降反向传播,得到最优策略。

[0068] S305,根据最优策略进行训练得到最优寻路参数并进行保存。

[0069] S40,还原所述游戏场景信息中的能产生伤害效果的NPC,所述游戏场景修改为初始寻路跟随场景。

[0070] 具体地,把游戏场景信息还原到原始状态,保留游戏场景中的能产生伤害效果的NPC,让智能体在之前的寻路参数初始化下继续训练复杂场景的寻路跟随。由于智能体已经经过简单寻路场景(即第一寻路跟随场景)下的寻路训练作为启发式训练,智能体具有一定的方向性决策,能提高智能体在复杂场景训练的探索效率,进而在复杂场景中进一步训练后,提高的复杂场景(即初始寻路跟随场景)下的寻路能力。

[0071] S50,根据所述寻路参数控制智能体进行所述寻路训练,直至完成预设的训练目标。其中,预设的训练目标为:按时到达目标位置。

[0072] 具体地,智能体加载之前在第一寻路跟随场景训练的寻路参数,在初始寻路跟随场景生成所有可以达到的点,作为训练的备用目标点,采用上述策略梯度强化学习算法继续导航训练,智能体随机初始化在任意可以到达的位置,从生成的所有可以到达点中随机选取一个点作为目标位置,训练智能体寻路到该目标位置,到达目标位置则完成任务,超时没有到达目标位置则失败,继续训练,直到按时到达目标位置。

[0073] 综上所述,通过接收游戏服发送的游戏场景信息,将游戏场景信息中的能产生伤害效果的NPC去掉得到第一寻路跟随场景,在第一寻路跟随场景中控制智能体根据预设的启发式奖励规则进行寻路训练,得到寻路参数,当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息,再将游戏场景修改为初始寻路跟随场景,根据寻路参数控制智能体再次进行寻路训练,直至完成预设的训练目标。由上分析可知,通过智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息,给予智能体一个启发式的方向,提高寻路效率,从而提高学习训练效率,减少时间周期。

[0074] 作为上述方案的改进,所述当智能体的当前位置与第一备用目标点的位置间的曼哈顿距离减少后,则向所述智能体发送对应的启发式奖励信息,具体包括:

[0075] 当所述智能体未到达所述目标位置时,每一帧的启发式奖励根据第一启发式奖励公式计算;其中,所述第一启发式奖励公式,具体为,$R=\alpha(t)*(D_{pre}-D_{now})-\beta$,$D_{pre}$为前一帧的智能体与目标位置的曼哈顿距离,$D_{now}$为在当前帧的智能体与目标位置的曼哈顿距离,$\alpha(t)$为随游戏帧数以预设的衰减规则不断减小的退火的因子,β为每一帧的惩罚因子。其中,惩罚因子与当场游戏的时间总长度(游戏帧数)和游戏胜利时候的最终奖励有关,参考值为最终胜利奖励值除以游戏时间总长度,一般小于这个值。预设的衰减规则可以为线性衰减,也

可以指数衰减，也可以为间隔固定时间衰减固定数值，让奖励幅度随着时间衰减，衰减程度根据在不同的场合调节对应的值，比如线性衰减，可以初始为0.5倍衰减，后续不断优化这个值到最佳。

[0076] 当所述智能体到达目标位置后，直接获得一个预设的正向奖励。

[0077] 在本实施例中，参见图3，带状条表示能够通过的区域，0点为智能体所在的位置，A,B,C,D四点分别为不同的目标点位置，当目标点在不同位置时，和0点的曼哈顿距离的示意图。

[0078] 直接用当前位置和目标的位置的曼哈顿距离作为启发式奖励的参考依据，当距离减少时候，给予智能体奖励，虽然实际上寻路时候的轨迹大多情况因为有障碍阻挡曼哈顿距离不是最短距离，但是这个奖励是作为强化学习的一个启发式奖励，在智能体探索期间，可以给智能体一个启发式的方向，能够提高搜索效率。同时，启发式奖励一直随着时间在慢慢衰减，直到为零。

[0079] 作为上述方案的改进，S304，所述将对应的所述启发式奖励信息反馈至预设的策略梯度强化学习算法，计算当前回合的最大奖励的梯度，通过梯度下降反向传播，得到最优策略，具体包括：

[0080] 所述预设的策略梯度强化学习算法包括：动作策略输出网络和价值估计网络，其中，训练服和游戏服不断交互的状态转移数据(包括游戏场景信息等)和对应的奖励数据，反馈给动作策略输出网络和价值估计网络进行学习，实现最大化每个回合的奖励。从交互序列中获得序列数据，为了获得当前回合的最大奖励，计算当前回合的最大奖励的梯度，通过梯度下降反向传播，得到最优策略。

[0081] 根据以下公式计算当前回合的最大奖励的梯度：

[0082] $$\hat{g}_k = \frac{1}{|\mathcal{D}_k|}\sum_{\tau\in\mathcal{D}_k}\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t)|_{\theta_k}\hat{A}_t$$

[0083] 其中，$\mathcal{D}_k$为智能体和环境的交互的序列数据τ的集合，每条序列τ长度为最大长度为T，序列中每个时间节点包括状态s_t，动作a_t，当前动作策略下执行动作a_t的概率$\pi_\theta(a_t|s_t)$，该节点的价值估计$\hat{V}_\phi(s_t)$，执行该动作对应奖励r_t，$\hat{A}_t$为该节点的优势估计，$\hat{A}_t = \sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} + \gamma^{T-t}\hat{V}_\phi(s_T) - \hat{V}_\phi(s_t)$；$\hat{V}_\phi(s_t)$为在t时刻该节点的价值估计，γ为奖励折扣因子，

[0084] 对策略输出网络参数进行更新的公式如下，

[0085] $$\theta_{k+1} = \theta_k + \alpha_k\hat{g}_k$$

[0086] 其中，θ_k为第k次迭代的策略网络参数，α_k为策略网络的学习率，$\hat{g}_k$为每次神经网络损失函数反向传播的梯度；

[0087] 对价值估计网络参数进行更新的公式如下，

[0088] $$\phi_{k+1} = \operatorname{argmin}_{\phi} \frac{1}{|\mathcal{D}_k|T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(V_\phi(s_t) - \hat{R}_t\right)^2$$

[0089] 其中，ϕ_k为第k次迭代的价值网络参数，$\hat{R}_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} + \gamma^{T-t} \hat{V}_\phi(s_T)$，为时间t节点的实际状态价值。

[0090] 可选的，强化学习网络中的两个学习网络(策略网络和价值估计网络)为共享特征表示层的深度神经网络，输入包括当前游戏画面、玩家当前位置坐标和目标位置坐标。其中，共享特征表示层包括三层卷积层和两层全连接层。三层卷积层把当前帧的游戏画面进行特征提取，得到的特征与智能体的当前位置坐标和目标点位置坐标进行组合，形成新的特征，通过两层全连接层，得到策略网络和价值估计网络的共享特征。

[0091] 作为上述方案的改进，在所述接收游戏服发送的游戏场景信息之后，所述将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景之前，所述方法还包括：

[0092] 对所述游戏场景信息进行编码作为所述智能体的输入信息。

[0093] 在本实施例中，对所述游戏场景信息编码方式，具体包括：将所述游戏场景信息绘制成对应的二位图像。

[0094] 所述智能体根据所述输入信息输出对应的执行动作，并将对应的执行动作编码后反馈至所述游戏服。

[0095] 在本实施例中，智能体的操作包括上下左右行走和跳跃五个操作，作为智能体的输出。

[0096] 在本实施例中，采用one-hot方式对执行动作进行编码，得到输出行为数据，将所述输出行为数据反馈至所述游戏服。

[0097] 参见图4，是本发明一实施例提供的一种游戏智能体的训练装置的结构示意图。

[0098] 本发明实施例对应提供了一种游戏智能体的训练装置，包括：

[0099] 接收模块10，用于接收游戏服发送的游戏场景信息。

[0100] 第一游戏场景修改模块20，用于将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景。

[0101] 第一控制模块30，用于控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数；其中，所述预设的启发式奖励规则为：当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后，则向所述智能体发送对应的启发式奖励信息。

[0102] 第二游戏场景修改模块40，用于还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景。

[0103] 第二控制模块50，用于根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标。

[0104] 综上所述，通过接收游戏服发送的游戏场景信息，将游戏场景信息中的能产生伤害效果的NPC去掉得到第一寻路跟随场景，在第一寻路跟随场景中控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数，当智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息，再将游戏场景修改为初始

寻路跟随场景，根据寻路参数控制智能体再次进行寻路训练，直至完成预设的训练目标。由上分析可知，通过智能体的当前位置与智能体的目标位置之间的曼哈顿距离减少后向智能体发送对应的启发式奖励信息，给予智能体一个启发式的方向，提高寻路效率，从而提高学习训练效率，减少时间周期。

[0105] 参见图5，是本发明一实施例提供的游戏智能体的训练设备的示意图。该实施例的游戏智能体的训练设备包括：处理器、存储器以及存储在所述存储器中并可在所述处理器上运行的计算机程序。所述处理器执行所述计算机程序时实现上述各个游戏智能体的训练方法实施例中的步骤。或者，所述处理器执行所述计算机程序时实现上述各装置实施例中各模块/单元的功能。

[0106] 示例性的，所述计算机程序可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器中，并由所述处理器执行，以完成本发明。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述所述计算机程序在所述游戏智能体的训练设备中的执行过程。

[0107] 所述游戏智能体的训练设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述游戏智能体的训练设备可包括，但不仅限于，处理器、存储器。本领域技术人员可以理解，所述示意图仅仅是游戏智能体的训练设备的示例，并不构成对游戏智能体的训练设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述游戏智能体的训练设备还可以包括输入输出设备、网络接入设备、总线等。

[0108] 所称处理器可以是中央处理单元(Central Processing Unit，CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor，DSP)、专用集成电路(Application Specific Integrated Circuit，ASIC)、现成可编程门阵列(Field-Programmable Gate Array，FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等，所述处理器是所述游戏智能体的训练设备的控制中心，利用各种接口和线路连接整个游戏智能体的训练设备的各个部分。

[0109] 所述存储器可用于存储所述计算机程序和/或模块，所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块，以及调用存储在存储器内的数据，实现所述游戏智能体的训练设备的各种功能。所述存储器可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等；存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外，存储器可以包括高速随机存取存储器，还可以包括非易失性存储器，例如硬盘、内存、插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)、至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0110] 其中，所述游戏智能体的训练设备集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明实现上述实施例方法中的全部或部分流程，也可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算

机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。

[0111] 需说明的是，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外，本发明提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

[0112] 以上所述是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也视为本发明的保护范围。

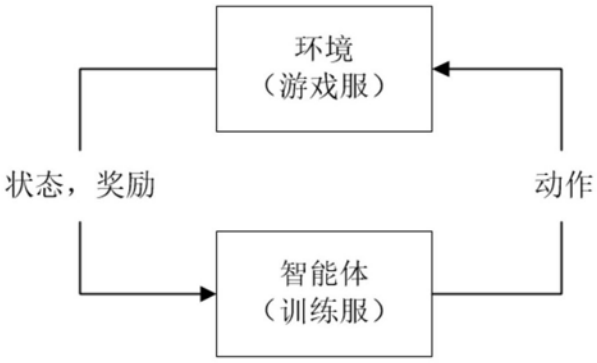

环境
（游戏服）
状态，奖励
动作
智能体
（训练服）

图1

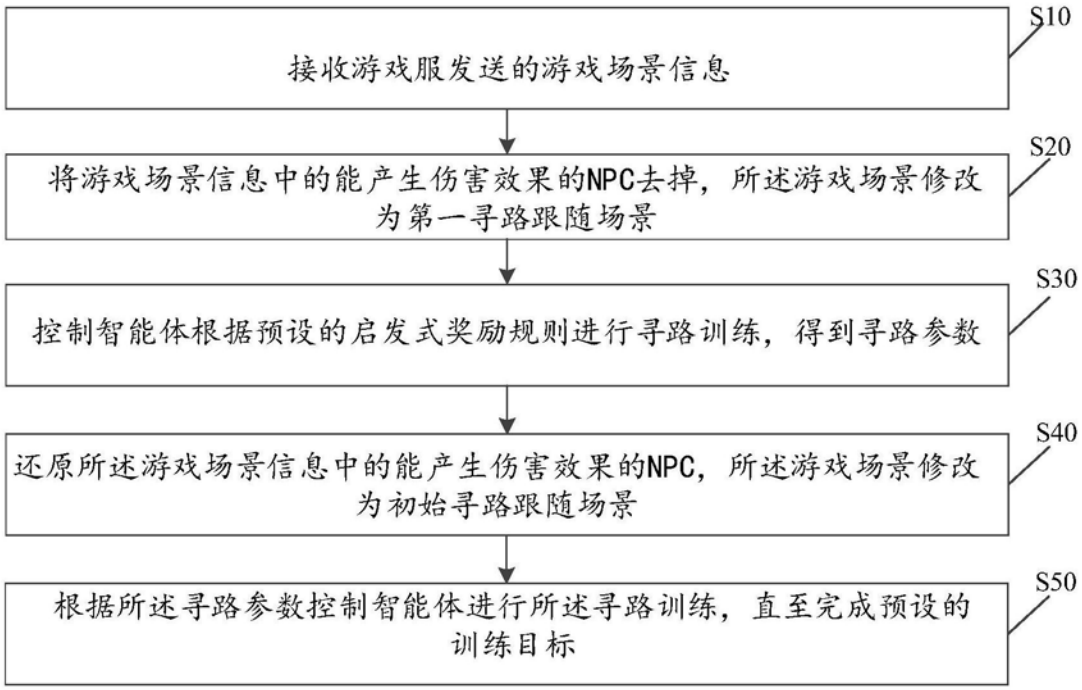

S10
接收游戏服发送的游戏场景信息
S20
将游戏场景信息中的能产生伤害效果的NPC去掉，所述游戏场景修改为第一寻路跟随场景
S30
控制智能体根据预设的启发式奖励规则进行寻路训练，得到寻路参数
S40
还原所述游戏场景信息中的能产生伤害效果的NPC，所述游戏场景修改为初始寻路跟随场景
S50
根据所述寻路参数控制智能体进行所述寻路训练，直至完成预设的训练目标

图2

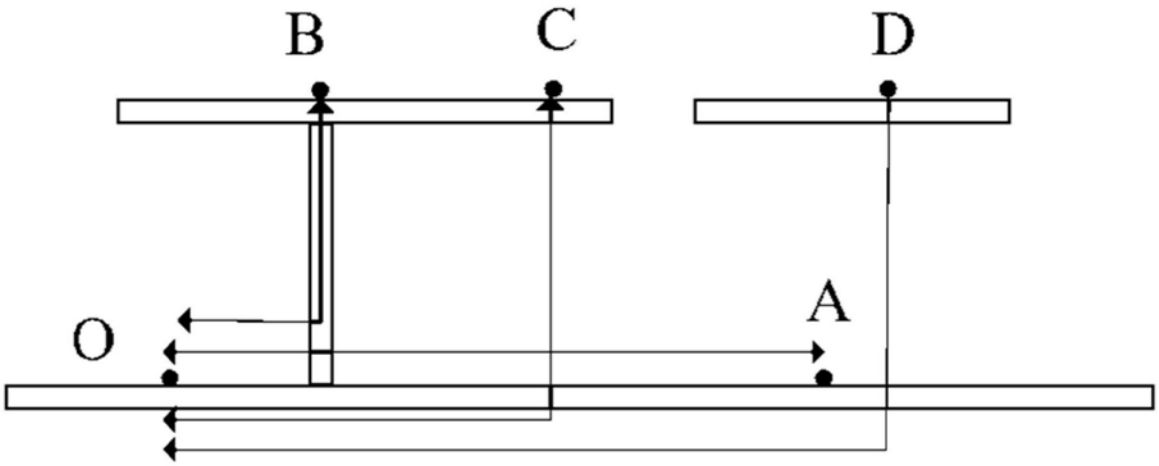

B
C
D
A
O

图3

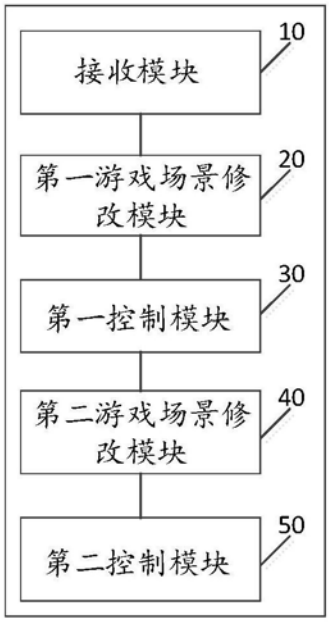

接收模块
10
第一游戏场景修改模块
20
第一控制模块
30
第二游戏场景修改模块
40
第二控制模块
50

图4

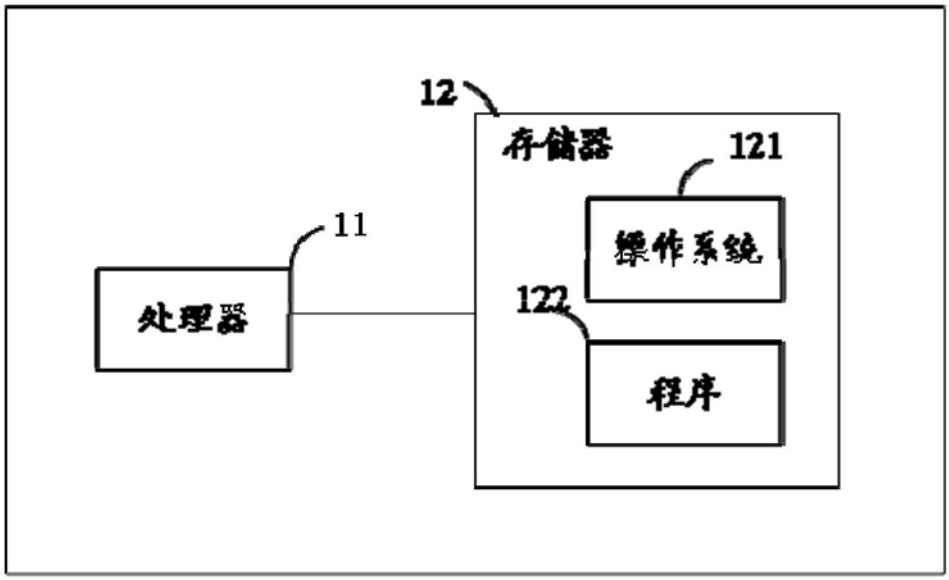


图5