(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 121028603 A
(43)申请公布日 2025.11.28

(21)申请号 202511204241.2

(22)申请日 2025.08.26

(71)申请人 清华大学
地址 100084 北京市海淀区清华园

(72)发明人 封硕 杨敬轩 张毅 王子航 姬浩元 陆秋婧

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 张建秀 龙洪

(51)Int.Cl.
G05B 17/02(2006.01)

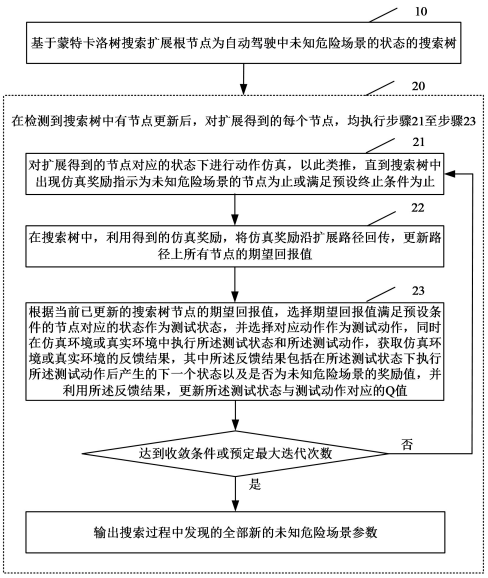
权利要求书2页 说明书13页 附图2页

(54)发明名称

未知危险场景的获取和控制方法、存储介质和电子装置

(57)摘要

一种未知危险场景的获取和控制方法、存储介质和电子装置,所述方法包括:步骤10、扩展根节点为自动驾驶中未知危险场景的状态的搜索树;步骤20、在检测到搜索树中有节点更新后,对扩展的每个节点,均执行步骤21至步骤23,直至达到终止条件为止,输出搜索过程中发现的全部新的未知危险场景参数,其中:步骤21、对扩展得到的节点对应的状态下进行动作仿真,以此类推,直到出现未知危险场景的节点为止或满足预设终止条件为止;步骤22、在搜索树中,利用得到的仿真奖励,更新路径上所有节点的期望回报值;步骤23、选择测试状态和测试动作,获取反馈结果,利用反馈结果,更新测试状态与测试动作对应的Q值,得到获取未知危险场景的目的。



CN 121028603 A

1.一种未知危险场景的获取方法,包括:

步骤10、基于蒙特卡洛树搜索扩展根节点为自动驾驶中未知危险场景的状态s的搜索树;

步骤20、在检测到搜索树中有节点更新后,对扩展得到的每个节点,均执行步骤21至步骤23,直至达到收敛条件或预定最大迭代次数为止,输出搜索过程中发现的全部新的未知危险场景参数,其中:

步骤21、对扩展得到的节点对应的状态下进行动作仿真,以此类推,直到搜索树中出现仿真奖励指示为未知危险场景的节点为止或满足预设终止条件为止;

步骤22、在搜索树中,利用得到的仿真奖励,将仿真奖励沿扩展路径回传,更新路径上所有节点的期望回报值;

步骤23、根据当前已更新的搜索树节点的期望回报值,选择期望回报值满足预设条件的节点对应的状态作为测试状态s^{t},并选择对应动作作为测试动作a^{t},同时在仿真环境或真实环境中执行所述测试状态s^{t}和所述测试动作a^{t},获取仿真环境或真实环境的反馈结果,其中所述反馈结果包括在所述测试状态下执行所述测试动作后产生的下一个状态s'以及是否为未知危险场景的奖励值r^{t},并利用所述反馈结果,更新所述测试状态s^{t}与测试动作a^{t}对应的Q值$Q(s^{t},a^{t})$。

2.根据权利要求1所述的方法,其特征在于:

在步骤23中,Q值$Q(s^{t},a^{t})$的更新操作是根据奖励值r^{t}以及状态 s′下所有动作a′ 中期望回报的最大值$\max_{a'} Q(s',a')$共同确定的。

3.根据权利要求2所述的方法,其特征在于:

在步骤23中,利用折扣因子γ对最大值$\max_{a'} Q(s',a')$进行修正,利用修正后的数值与奖励值r^{t}更新Q值$Q(s^{t},a^{t})$。

4.根据权利要求3所述的方法,其特征在于:

Q值$Q(s^{t},a^{t})$的更新操作的计算表达式如下:

$$Q(s^{t},a^{t})\leftarrow(1-\alpha)Q(s^{t},a^{t})+\alpha\left(r^{t}+\gamma\cdot\max_{a'}Q(s',a')\right);$$

其中,α表示学习率。

5.根据权利要求1所述的方法,其特征在于:

在步骤10中,沿着基于概率置信区间上界PUCT确定的最优动作路径直到到达根节点的叶节点,若叶节点还有未尝试动作,则选择一个未尝试动作执行,扩展该叶节点,生成子节点,基于子节点对应状态下所执行的动作扩展下一级节点。

6.根据权利要求1所述的方法,其特征在于:

在步骤22中,节点i对应的状态s_i执行的每个动作a_i对应的Q值$Q(s_i,a_i)$还根据神经网络模型对状态s_i的价值预测值$v_\theta(s_i)$,其中该价值预测值用于表示状态s_i在当前策略下导致未知危险场景的概率,其中θ表示神经网络模型的模型参数,i为正整数。

7.根据权利要求6所述的方法,其特征在于:

Q值$Q(s_i,a_i)$的计算表达式如下：

$$Q(s_i,a_i)=(1-\lambda)\cdot\frac{W(s_i,a_i)}{N(s_i,a_i)}+\lambda\cdot v_\theta(s_i);$$

其中：

λ为融合系数；

$W(s_i,a_i)$表示状态s_i下动作a_i的累计奖励；

$N(s_i,a_i)$表示动作a_i在状态s_i下被执行次数，i为正整数。

8.一种自动驾驶控制方法，包括：

获取自动驾驶操作的管理数据，其中所述管理数据是基于如权利要求1至7中任一项所述的方法得到的未知危险场景得到的；

根据所述管理数据对自动驾驶行为进行管理。

9.一种存储介质，其特征在于，所述存储介质中存储有计算机程序，其中，所述计算机程序被设置为运行时执行所述权利要求1至8中任一项所述的方法。

10.一种电子装置，包括存储器和处理器，其特征在于，所述存储器中存储有计算机程序，所述处理器被设置为运行所述计算机程序以执行所述权利要求1至8中任一项所述的方法。

未知危险场景的获取和控制方法、存储介质和电子装置

技术领域

[0001] 本文涉及自动驾驶技术,尤指一种未知危险场景的获取和控制方法、存储介质和电子装置。

背景技术

[0002] 自动驾驶技术近年来取得了显著进展,但其在面对未知危险场景时的安全性仍是一个重大挑战。未知危险场景指的是那些在自然驾驶环境中极少出现,难以通过传统数据训练获得的驾驶场景。这些场景的出现频率虽然低,但一旦发生,可能会对自动驾驶系统的安全性造成严重影响。

[0003] 准确识别和应对这些未知危险场景对保障自动驾驶系统的安全性至关重要。然而,由于这些场景的稀有性,现有方法在高效地进行全面搜索和识别方面存在局限,导致对潜在事故风险的预测能力有限。这不仅增加了自动驾驶系统在实际应用中的风险,也限制了其在更广泛和更复杂环境中的部署。此外,这些未知危险场景难以通过传统数据训练获得,使得如何发现和这些场景成为亟待解决的问题。传统的自动驾驶系统依赖于大量的已知数据进行训练,以识别和响应常见的道路情况。但是,这种方法在面对未知危险场景时无法较好地适用,因为这些场景超出了训练数据的覆盖范围。

[0004] 因此,如何高效搜索和识别未知危险场景是亟待解决的技术问题。

发明内容

[0005] 本申请实施例提供了一种未知危险场景的获取和控制方法、存储介质和电子装置。

[0006] 一种未知危险场景的获取方法,包括:

步骤10、基于蒙特卡洛树搜索扩展根节点为自动驾驶中未知危险场景的状态s的搜索树;

步骤20、在检测到搜索树中有节点更新后,对扩展得到的每个节点,均执行步骤21至步骤23,直至达到收敛条件或预定最大迭代次数为止,输出搜索过程中发现的全部新的未知危险场景参数,其中:

步骤21、对扩展得到的节点对应的状态下进行动作仿真,以此类推,直到搜索树中出现仿真奖励指示为未知危险场景的节点为止或满足预设终止条件为止;

步骤22、在搜索树中,利用得到的仿真奖励,将仿真奖励沿扩展路径回传,更新路径上所有节点的期望回报值;

步骤23、根据当前已更新的搜索树节点的期望回报值,选择期望回报值满足预设条件的节点对应的状态作为测试状态s^t,并选择对应动作作为测试动作a^t,同时在仿真环境或真实环境中执行所述测试状态s^t和所述测试动作a^t,获取仿真环境或真实环境的反馈结果,其中所述反馈结果包括在所述测试状态下执行所述测试动作后产生的下一个状态s'以及是否为未知危险场景的奖励值r^t,并利用所述反馈结果,更新所述测试状态s^t与测试动

作a^t对应的Q值$Q(s^t,a^t)$。

[0007] 一种自动驾驶控制方法,包括:

获取自动驾驶操作的管理数据,其中所述管理数据是基于上文所述的方法得到的未知危险场景得到的;

根据所述管理数据对自动驾驶行为进行管理。

[0008] 一种存储介质,所述存储介质中存储有计算机程序,其中,所述计算机程序被设置为运行时执行上文所述的方法。

[0009] 一种电子装置,包括存储器和处理器,所述存储器中存储有计算机程序,所述处理器被设置为运行所述计算机程序以执行上文所述的方法。

[0010] 本申请实施例,通过构建以未知危险场景状态为根节点的搜索树,利用蒙特卡洛树搜索算法的高效探索能力,系统地覆盖各种可能的动作和状态转移路径,有效发现潜在未知危险场景;通过对新节点进行动作仿真和奖励回传,不断更新节点的期望回报值,使模型更精准地识别未知危险场景,再根据期望回报值选择测试状态和动作,并在仿真或真实环境中执行,验证这些状态和动作是否为未知危险场景,并通过更新Q值来不断优化模型对状态-动作价值的评估,从而能够更准确地挖掘出更多的未知危险场景。

[0011] 本申请的其它特征和优点将在随后的说明书中阐述,并且,部分地从说明书中变得显而易见,或者通过实施本申请而了解。本申请的其他优点可通过在说明书以及附图中所描述的方案来实现和获得。

附图说明

[0012] 附图用来提供对本申请技术方案的理解,并且构成说明书的一部分,与本申请的实施例一起用于解释本申请的技术方案,并不构成对本申请技术方案的限制。

[0013] 图1为本申请实施例提供的自动驾驶中未知危险场景的获取方法的流程示意图;

图2为本申请实施例提供的自动驾驶控制方法的流程示意图。

具体实施方式

[0014] 本申请描述了多个实施例,但是该描述是示例性的,而不是限制性的,并且对于本领域的普通技术人员来说显而易见的是,在本申请所描述的实施例包含的范围内可以有更多的实施例和实现方案。尽管在附图中示出了许多可能的特征组合,并在具体实施方式中进行了讨论,但是所公开的特征的许多其它组合方式也是可能的。除非特意加以限制的情况以外,任何实施例的任何特征或元件可以与任何其它实施例中的任何其他特征或元件结合使用,或可以替代任何其它实施例中的任何其他特征或元件。

[0015] 本申请包括并设想了与本领域普通技术人员已知的特征和元件的组合。本申请已经公开的实施例、特征和元件也可以与任何常规特征或元件组合,以形成独特的发明方案。任何实施例的任何特征或元件也可以与来自其它发明方案的特征或元件组合,以形成另一个独特的发明方案。因此,应当理解,在本申请中示出和/或讨论的任何特征可以单独地或以任何适当的组合来实现。因此,除了根据所附权利要求及其等同替换所做的限制以外,实施例不受其它限制。此外,可以在所附权利要求的保护范围内进行各种修改和改变。

[0016] 此外,在描述具有代表性的实施例时,说明书可能已经将方法和/或过程呈现为特

定的步骤序列。然而,在该方法或过程不依赖于本文所述步骤的特定顺序的程度上,该方法或过程不应限于所述的特定顺序的步骤。如本领域普通技术人员将理解的,其它的步骤顺序也是可能的。因此,说明书中阐述的步骤的特定顺序不应被解释为对权利要求的限制。此外,针对该方法和/或过程的权利要求不应限于按照所写顺序执行它们的步骤,本领域技术人员可以容易地理解,这些顺序可以变化,并且仍然保持在本申请实施例的精神和范围内。

[0017] 在自动驾驶领域中,s(State,状态)表示自动驾驶汽车当前所处的环境状态,a(Action,动作)表示自动驾驶汽车在状态s下可能采取的控制动作。

[0018] 其中状态通常包括但不限于以下参数:车辆自身的状态(如速度、加速度、转向角、位置等);周围环境信息(如其他车辆的位置、行人动态、道路拓扑、交通信号状态等);传感器数据(如摄像头、雷达、激光雷达的实时输入)。

[0019] 其中,动作可以为加速或减速(油门/刹车控制);转向(方向盘角度调整);车道保持或变道决策;紧急制动或避障策略。

[0020] 其中,$P(s,a)$表示预测的事故概率,具体来说,基于先验信息(如历史数据、仿真模型或规则库),在给定状态s下执行动作a时,自动驾驶汽车发生事故的预测概率。

[0021] $P(s,a)$表示真实事故概率,即在状态s下执行动作a时,通过真实测试或高保真仿真得出的实际事故概率,用于验证和校准自动驾驶模型确定的,$P(s,a)$的准确性。

[0022] 根据$P(s,a)$和$P^*(s,a)$可将状态动作空间划分为四个区域,分别为已知危险状态动作区域、已知安全状态动作区域、未知危险状态动作区域以及未知安全状态动作区域,如表1所示。

	危险状态动作	安全状态动作
已知	$P(s,a)>0, P^*(s,a)>0$	$P(s,a)=0, P^*(s,a)=0$
未知	$P(s,a)=0, P^*(s,a)>0$	$P(s,a)>0, P^*(s,a)=0$

表1

从表1可知,在$P(s,a)$=0且$P^*(s,a)$>0时,未知危险状态动作构成未知危险场景空间,表示为$\Omega_{UU}=(s,a)\in S\times A{:}P(s,a)=0,P^*(s,a)>0$。其中,在未知危险场景中,状态对应的动作在自动驾驶模型中被识别为安全驾驶行为,但在真实环境中却是危险驾驶行为。

[0023] 在自动驾驶安全性测试过程中,未知危险场景起着重要的作用。为了探索自动驾驶汽车的未知危险场景,未知危险场景生成方法需要根据已发现的少量未知危险状态动作对未知危险场景空间Ω_{UU}进行广泛且高效的探索,以便利用探索得到的信息训练未知危险状态动作识别模型以及未知危险状态动作生成模型,进而完成对未知危险状态动作的加速测试。

[0024] 本申请实施例提出,基于蒙特卡洛树搜索未知危险场景。

[0025] 具体的,通过对自动驾驶环境中的背景车辆行为空间进行搜索,以识别可能导致未知危险场景的动作序列。

[0026] 其中,蒙特卡洛树搜索过程为一个基于环境观测的状态转移问题。设状态变量为驾驶环境中6个方向车辆(左前、前、右前、左后、后、右后)的位置、速度等信息,动作变量为背景车辆的加速度、纵向位置等操作,搜索目标是找到能诱发未知危险场景的状态-动作序列。由于不同场景的时长不一致,导致搜索空间的复杂度较高,因此如何进行高效的搜索是

亟待解决的问题。

[0027] 本申请实施例提供如下解决方案,包括:

实施例一

图1为本申请实施例一提供的未知危险场景的获取方法的流程示意图。如图1所示,所述方法包括:

步骤10、基于蒙特卡洛树搜索扩展根节点为自动驾驶中未知危险场景的状态的搜索树。

[0028] 其中,蒙特卡洛树搜索是一种结合了随机模拟和树搜索的算法,用于在决策过程中寻找最优策略。它通过反复模拟可能的行动路径来评估每个决策的价值,尤其适用于具有不确定性和复杂状态空间的问题。

[0029] 其中,根节点是树结构的起始节点,对应于目标驾驶场景的初始状态,搜索树是一种数据结构,用于表示从初始状态出发,通过不同动作选择所到达的各种可能状态及其关系。

[0030] 通过构建以未知危险场景状态为根节点的搜索树,为后续的探索提供了一个系统性的框架,能够全面地覆盖从当前状态出发的各种可能动作和状态转移路径,从而有效发现潜在的未知危险场景。

[0031] 步骤20、在检测到搜索树中有节点更新后,对扩展得到的每个节点,均执行步骤21至步骤23,直至达到收敛条件或预定最大迭代次数为止,输出搜索过程中发现的全部新的未知危险场景参数。

[0032] 其中,节点更新是指在搜索树中添加新的节点或修改现有节点的信息,如期望回报值等;收敛条件是指搜索过程达到稳定状态的条件,即搜索结果不再显著变化或满足一定的精度要求;最大迭代次数是指预先设定的搜索过程的最大循环次数,用于限制搜索时间。确保搜索过程在达到一定的稳定性和覆盖范围后停止,避免无限循环,同时能够输出所有发现的未知危险场景参数,为后续的分析和应用提供完整的数据。

[0033] 步骤21、对扩展得到的节点对应的状态下进行动作仿真,以此类推,直到搜索树中出现仿真奖励指示为未知危险场景的节点为止或满足预设终止条件为止。

[0034] 动作仿真是指在虚拟环境中模拟执行某个动作,以观察其产生的结果和影响。

[0035] 仿真奖励是指根据仿真结果给予的反馈,用于指示某个状态或动作是否为未知危险场景。

[0036] 预设终止条件是指预先设定的停止搜索的条件,如达到最大搜索深度或时间等。

[0037] 通过对新节点对应的状态进行动作仿真,逐步扩展搜索树,发现潜在的未知危险场景,并在满足终止条件时停止搜索,避免不必要的计算和资源浪费。

[0038] 步骤22、在搜索树中,利用得到的仿真奖励,将仿真奖励沿扩展路径回传,更新路径上所有节点的期望回报值。

[0039] 期望回报值是指对某个节点在未来可能获得的奖励的估计,反映了该节点的价值。

[0040] 通过回传仿真奖励并更新期望回报值,能够使搜索树中的各个节点更准确地反映其对应状态的价值,为后续的动作选择提供更可靠的依据,有助于提高搜索效率和准确性。

[0041] 步骤23、根据当前已更新的搜索树节点的期望回报值,选择期望回报值满足预设

条件的节点对应的状态作为测试状态s^t,并选择对应动作作为测试动作a^t,同时在仿真环境或真实环境中执行所述测试状态s^t和所述测试动作a^t,获取仿真环境或真实环境的反馈结果,其中所述反馈结果包括在所述测试状态下执行所述测试动作后产生的下一个状态s'以及是否为未知危险场景的奖励值r^t,并利用所述反馈结果,更新所述测试状态s^t与测试动作a^t对应的Q值$Q(s^t,a^t)$。

[0042] 其中,预设条件是指预先设定的选择标准,用于确定哪些节点具有较高的价值或潜力;测试状态是指被选中用于进一步测试的状态;测试动作是指被选中用于在测试状态下执行的动作;反馈结果是指从仿真或真实环境中获得的关于测试状态和动作的结果信息,包括下一个状态和奖励值等;Q 值是指表示在某个状态下执行某个动作后可能获得的预期回报,是衡量状态 - 动作对价值的重要指标。

[0043] 根据期望回报值选择测试状态和动作,并在仿真或真实环境中执行,能够验证这些状态和动作是否为未知危险场景,并通过更新 Q 值来不断优化模型对状态-动作价值的评估,提高搜索的针对性和有效性。

[0044] 上述技术方案所到达的技术效果包括:

高效探索未知危险场景:通过构建以未知危险场景状态为根节点的搜索树,利用蒙特卡洛树搜索 算法的高效探索能力,能够系统地覆盖各种可能的动作和状态转移路径,从而有效发现潜在的未知危险场景,解决了现有方法难以全面覆盖复杂未知危险情境的问题。

[0045] 提高安全性评估的准确性:通过对新节点进行动作仿真和奖励回传,不断更新节点的期望回报值,使模型能够更准确地评估每个状态的价值,从而更精准地识别未知危险场景,提高了自动驾驶系统安全性评估的准确性。

[0046] 优化决策过程:根据期望回报值选择测试状态和动作,并在仿真或真实环境中执行,能够验证这些状态和动作是否为未知危险场景,并通过更新 Q 值来不断优化模型对状态 - 动作价值的评估,从而为自动驾驶系统的决策提供更可靠的依据,优化了决策过程,提高了系统在面对未知危险场景时的应对能力。

[0047] 提供全面的场景覆盖:从根节点的目标驾驶场景状态出发,逐步扩展搜索树,直到出现未知危险场景或满足终止条件,能够全面覆盖从当前状态出发的各种可能驾驶场景,为自动驾驶系统的安全评估和改进提供了全面的数据支持,有助于提升自动驾驶的安全性和可靠性。

[0048] 下面对本申请实施例提供的方法进行说明:

在一个示例性实施例中,在步骤23中,Q值$Q(s^t,a^t)$的更新操作是根据奖励值r^t以及状态 s′下所有动作a′ 中期望回报的最大值$\max_{a'} Q(s',a')$共同确定的。

[0049] 奖励值r^t是在仿真环境中执行某个动作后获得的反馈,用于指示该动作是否导致了未知危险场景。例如,若执行某个动作后出现了未知危险场景,则对应的奖励值会被设置为1;否则,奖励值为0。

[0050] 期望回报的最大值$\max_{a'} Q(s',a')$ 是指在状态s′下,对所有可能的动作a′计算其期望回报值,取其中的最大值。期望回报值反映了执行某个动作后可能获得的长期收益的估

计,综合了该动作及其后续动作可能带来的奖励。

[0051] 结合奖励值和期望回报的最大值来更新Q值,能够同时考虑当前动作的即时效果(奖励值)以及后续动作的潜在价值(期望回报的最大值),从而使Q值更准确地反映状态-动作对的真实价值,为后续的动作选择提供更可靠的依据。

[0052] 上述示例性实施例,通过根据奖励值以及状态s′下所有动作a′中期望回报的最大值共同确定Q值的更新操作,能够使模型在评估状态-动作对的价值时,不仅关注当前的即时奖励,还考虑未来的潜在收益。这有助于更准确地识别那些可能在短期内看似安全,但长期来看存在潜在危险的场景,从而进一步提高自动驾驶系统在面对复杂未知危险场景时的应对能力,增强系统的安全性和可靠性。同时,这种更新方式也能够加快模型的学习速度,使其更快地收敛到最优策略,提高搜索效率。

[0053] 进一步的,在步骤23中,利用折扣因子γ对最大值$\max_{a'} Q(s',a')$进行修正,利用修正后的数值与奖励值r^t更新Q值$Q(s^t,a^t)$。

[0054] 其中,折扣因子是一个介于0和1之间的参数,用于平衡当前奖励和未来奖励的重要性。通过将未来奖励进行折扣处理,可以反映出未来奖励的不确定性以及及时性的考量。

[0055] 将期望回报的最大值乘以折扣因子,得到修正后的数值。这使得模型在更新Q值时,不仅考虑当前的奖励,还会综合考虑未来可能获得的奖励,但对未来的奖励进行一定的折扣,以体现其不确定性。引入折扣因子对期望回报的最大值进行修正在Q值更新操作中,能够使模型在学习过程中更好地平衡短期利益(当前奖励)与长期利益(未来奖励)。在面对复杂的未知危险场景搜索时,这样的更新方式能帮助模型更稳健地评估状态-动作对的价值,避免过度关注短期的奖励而忽视了长期可能存在的风险或收益,从而提高搜索过程中的决策质量和对未知危险场景的识别能力。

[0056] 通过在Q值更新时利用折扣因子对最大值进行修正,使得模型在评估状态-动作对的价值时,能够综合考虑当前和未来的奖励,并且合理地权衡两者的重要性。这有助于提高模型对未来潜在危险场景的预判能力,避免因过度关注即时奖励而忽视潜在的未知危险。同时,这种更新方式增强了模型的稳定性和适应性,使其在不同的驾驶场景和条件下都能更有效地发现未知危险场景,进一步提升自动驾驶系统在复杂环境下的安全性和可靠性。

[0057] 在一个示例性实施例中,Q值$Q(s^t,a^t)$的更新操作的计算表达式如下:

$$Q(s^t,a^t)\leftarrow(1-\alpha)Q(s^t,a^t)+\alpha\left(r^t+\gamma\cdot\max_{a'}Q(s',a')\right);$$

其中,α表示学习率。

[0058] 通过上述 Q 值更新方式,模型在评估状态-动作对的价值时,能够综合考虑当前的即时奖励和未来的潜在价值。这种更新方式不仅提高了模型对未知危险场景的识别能力,还增强了模型的稳定性和适应性。具体表现为:

提高搜索效率: 通过合理设置学习率和折扣因子,模型能够更快地收敛到最优策略,减少不必要的计算和资源浪费。

[0059] 增强安全性评估: 通过综合考虑当前和未来的奖励,模型能够更准确地识别潜在的未知危险场景,提高自动驾驶系统的安全性评估准确性。

[0060] 适应复杂环境: 通过动态调整 Q 值,模型能够更好地适应不同的驾驶场景和条

件,提高其在复杂环境下的应对能力。

[0061] 综上所述,通过引入具体的 Q 值更新公式,进一步细化了 Q 值的更新机制,优化了模型对状态-动作对价值的评估,从而提高了对未知危险场景的搜索效率和准确性。

[0062] 在一个示例性实施例中,在步骤10中,沿着基于PUCT(Probabilistic Upper Confidence Tree bound,概率置信区间上界)确定的最优动作路径直到到达根节点的叶节点,若叶节点还有未尝试动作,则选择一个未尝试动作执行,扩展该叶节点,生成子节点,基于子节点对应状态下所执行的动作扩展下一级节点。

[0063] 其中,基于PUCT确定的最优动作的计算表达式如下:

$$a^{*}=\mathrm{argmax}_{a}\left(Q(s,a)+c_{\mathrm{puct}}\cdot P(s,a)\cdot\frac{\sqrt{\sum_{b}N(s,b)}}{1+N(s,a)}\right);$$

其中:

a^{*}为最优动作;

$Q(s,a)$表示状态s下动作a的期望回报;

c_{puct}为探索系数,取值为常数,用于控制探索未知危险状态动作与利用已知高回报动作之间的权重,其中取值越大,越倾向于探索未知危险状态动作;

$P(s,a)$为策略网络对动作a的先验概率;

$N(s,a)$表示动作a在状态s下被执行次数;其中,$\sum_{b}N(s,b)$表示在状态s 下,所有可能的动作b 被执行次数之和。

[0064] 其中,argmax 表示使函数取得最大值的输入值,其英文全称是argument of the maximum。

[0065] 例如,状态s 下有 3 个可能动作 a1,a2,a3,且它们被执行次数分别为 N(s,a1)=5、N(s,a2)=3、N(s,a3)=2,则$\sum_{b}N(s,b)$=5+3+2=10。

[0066] 在上述表达式中,$\sqrt{\sum_{b}N(s,b)}$作为分子,$1+N(s,a)$作为分母;其中:

分子能够体现状态 s 的整体探索热度,其中平方根的作用是减缓数值增长,避免总访问次数过大时,探索项失去平衡。

[0067] 分母能够体现动作 a 的已探索程度,其中加 1 是为了防止 N(s,a)=0 时出现除零错误。

[0068] 之所以分子是总访问次数,而分母是单动作访问次数,原因在于,鼓励对访问次数少的动作(即未充分探索的动作)进行更多探索。

[0069] 对于某个动作 a,如果被执行次数N(s,a) 较少,则分母 1+N(s,a) 较小,分数值更大,算法倾向于选择该动作。在不断访问动作 a后,N(s,a) 变大,对应的分母1+N(s,a)增大,分数值 降低,减少对该动作的重复探索。

[0070] 例如,假设状态 s 下总访问次数$\sum_{b}N(s,b)$ = 100,且有两个动作:

动作a1 的N(s,a1)=10,则其探索项中分子与分母的比值约为0.91;

动作 a2 的 N(s,a2)=2,则其探索项为分子与分母的比值约为3.33。

[0071] 从数值的大小可知,访问次数更少的a2 会被优先探索。

[0072] 另外,选择分子为总访问次数的平方根的原因在于:

从数学原理上分析,平方根函数$\sqrt{x}$的增长速度慢于线性函数x,可以避免总访问次数过大时,探索项主导整个表达式;

从实际效果来看,随着总访问次数的增加,探索项的权重增长逐渐放缓,确保算法在后期更多依赖已知的 Q(s,a)(即利用)。

[0073] 例如,若使用$\sum_b N(s,b)$,当总访问次数为 1000 时,探索项可能高达 1000,远超过Q(s,a),导致算法完全偏向探索,忽略已有知识。在使用平方根后,探索项为31.6,更易于平衡。

[0074] 基于PUCT算法确定的最优动作路径,是从根节点到叶节点的一条路径,该路径上的动作具有最高的潜在价值。通过沿着这条路径搜索,模型能够快速找到最有希望的未知危险场景。通过选择最优动作路径,模型能够集中资源探索最有潜力的分支,减少不必要的计算,提高搜索效率。

[0075] 当到达叶节点时,如果该叶节点还有未尝试的动作,模型会选择一个未尝试的动作执行,并扩展该叶节点生成新的子节点。这种扩展机制确保了搜索树的不断扩展和更新。通过扩展叶节点,模型能够逐步覆盖更多的状态空间,发现更多的潜在未知危险场景,提高搜索的全面性。

[0076] 基于子节点对应状态下所执行的动作,进一步扩展下一级节点。这有助于构建更深入的搜索树,探索更复杂的状态转移路径。子节点的生成不仅增加了搜索树的深度,还为模型提供了更多的信息来评估不同动作的长期影响,从而提高对未知危险场景的识别能力。

[0077] 通过基于PUCT算法的最优动作路径选择和叶节点扩展机制,模型能够高效地探索未知危险场景。具体表现为:

提高搜索效率: PUCT算法通过平衡探索和利用的关系,帮助模型快速找到最有潜力的未知危险场景,减少不必要的计算。

[0078] 增强探索能力: 通过扩展叶节点和生成新的子节点,模型能够逐步覆盖更多的状态空间,发现更多的潜在未知危险场景。

[0079] 提高识别准确性: 通过构建更深入的搜索树,模型能够评估不同动作的长期影响,从而更准确地识别未知危险场景。

[0080] 综上所述,通过引入PUCT算法和叶节点扩展机制,优化了蒙特卡洛树搜索的过程,提高了对未知危险场景的搜索效率和识别准确性。

[0081] 在一个示例性实施例中,在步骤22中,节点i对应的状态s_i执行的每个动作a_i对应的Q值$Q(s_i,a_i)$还根据神经网络模型对状态s_i的价值预测值$v_\theta(s_i)$,其中该价值预测值用于表示状态s_i在当前策略下导致未知危险场景的概率,其中θ表示神经网络模型的模型参数,i为正整数。

[0082] 其中,神经网络模型是一种基于人工神经网络的机器学习模型,能够学习和预测复杂数据模式。其中,神经网络模型输出的对某个状态价值的预测,反映了在当前策略下,该状态导致未知危险场景的可能性大小。价值预测值越高,表示该状态越有可能导致未知危险场景。

[0083] 基于上述方式确定期望回报的原因和优势如下：

首先，神经网络模型能够学习复杂的数据模式和非线性关系，这为其在处理复杂驾驶场景时提供了独特优势。引入神经网络模型能够更准确地评估每个节点对应的状态s_i的价值，进而提高搜索路径中节点i的期望回报的准确性。

[0084] 其次，神经网络模型能够基于大量的历史数据进行训练，从而准确评估各种驾驶场景下的风险。这种基于数据驱动的方法，能够更好地适应自动驾驶场景的多样性和复杂性，提供更可靠的决策支持。

[0085] 此外，结合神经网络模型的价值预测值和传统的搜索算法，可以实现更高效的搜索过程。神经网络模型能够快速评估大量可能的状态，帮助搜索算法更快地收敛到最优的决策路径，从而提高搜索效率。

[0086] 最后，结合神经网络模型和搜索算法的方法还能够提高搜索的鲁棒性。在面对未知或复杂场景时，神经网络模型能够提供额外的指导，帮助搜索算法更好地应对不确定性，提高决策的可靠性。

[0087] 综上所述，通过引入神经网络模型对节点i对应的状态s_i的价值进行预测，不仅提高了搜索路径中节点i的期望回报的准确性，还增强了搜索过程的效率和鲁棒性，为自动驾驶系统在复杂环境中的安全决策提供了有力支持。

[0088] 上述示例性实施例，通过结合仿真奖励和神经网络模型的价值预测值来更新Q值，能够更准确地评估每个状态的价值。具体效果包括：

提高评估准确 ：神经网络模型能够学习到复杂的状态特征和价值关系，提供更精确的价值预测值。结合仿真奖励和价值预测值，可以更全面地评估状态的价值，避免单纯依赖仿真奖励可能导致的评估偏差。

[0089] 增强泛化能力：神经网络模型具有良好的泛化能力，能够在有限的训练数据基础上，对未见过的状态进行较为准确的价值预测。这有助于在搜索过程中更快地识别潜在的未知危险场景，减少对大规模仿真数据的依赖。

[0090] 提升学习效率：通过引入神经网络模型的价值预测，可以加速Q值的收敛过程，使模型更快地学习到有效的策略，提高搜索效率。

[0091] 综上所述，通过引入神经网络模型的价值预测，进一步优化了Q值的更新机制，提高了对未知危险场景的识别能力和搜索效率，增强了自动驾驶系统的安全性评估性能。

[0092] 其中，Q值$Q(s_i,a_i)$的计算表达式如下：

$$Q(s_i,a_i)=(1-\lambda)\cdot\frac{W(s_i,a_i)}{N(s_i,a_i)}+\lambda\cdot v_\theta(s_i);$$

其中：

λ为融合系数；

$W(s_i,a_i)$表示状态s_i下动作a_i的累计奖励；

$N(s_i,a_i)$表示动作a_i在状态s_i下被执行次数；

$v_\theta(s_i)$表示由神经网络模型对状态s_i的价值预测值，其中θ表示神经网络模型的模型参数。

[0093] 在上述计算表达式中，利用累计奖励 $W(s_i,a_i)$和访问次数 $N(s_i,a_i)$的技术优势包

括：

准确评估动作价值：累计奖励 $W(s_i,a_i)$记录了在状态s_i下执行动作a_i所获得的总奖励，而访问次数 $N(s_i,a_i)$则反映了该动作的被执行次数。通过将累计奖励除以访问次数，可以得到一个平均奖励值，这有助于更准确地评估该动作的价值。这种基于历史数据的统计方法能够有效减少因少量高奖励事件导致的偏差，提高评估的稳定性和可靠性。

[0094] 平衡探索与利用：在强化学习中，探索新的动作和利用已知的高奖励动作之间需要找到一个平衡点。通过结合累计奖励和访问次数，系统能够在探索新动作的同时，充分利用已有的信息，避免因过度探索而导致的效率低下或因过度利用而导致的局部最优。

[0095] 动态适应性：随着迭代的进行，累计奖励和访问次数会不断更新，这使得期望回报的计算能够动态适应系统的探索过程。这种动态适应性有助于系统在不断变化的环境中保持良好的性能。

[0096] 在上述计算表达式中，融合系数 λ 的作用及其优势包括：

平衡历史数据与模型预测：融合系数 λ 用于平衡基于历史数据的统计评估（累计奖励与访问次数的比值）和基于神经网络模型的预测值。通过调整 λ，可以在依赖历史数据和依赖模型预测之间找到一个合适的平衡点。这有助于系统在不同的迭代阶段或不同的场景下灵活调整评估策略。

[0097] 减少模型偏差：神经网络模型的预测可能会受到数据噪声和模型误差的影响。通过引入融合系数，可以减少对单一模型预测的依赖，降低模型偏差对决策的影响，提高系统的鲁棒性。

[0098] 提高评估准确性：历史数据提供了可靠的统计信息，而模型预测则提供了对未知状态的前瞻性评估。通过融合这两种信息源，可以更全面地评估动作的价值，提高评估的准确性和可靠性。

[0099] 自适应调整：根据不同的迭代阶段或不同的场景，可以动态调整融合系数 λ。例如，在探索阶段，可以适当降低 λ，增加对历史数据的重视，以加快对新场景的探索；在利用阶段，可以提高 λ，更多地依赖模型的预测，以提高决策的效率和准确性。

[0100] 基于上述说明可知，通过结合累计奖励 $W(s_i,a_i)$和访问次数 $N(s_i,a_i)$，能够准确评估动作的价值，平衡探索与利用，并动态适应环境变化；另外，融合系数 λ 的引入进一步增强了系统的灵活性和鲁棒性，使其能够在不同的情况下做出更优的决策。这种设计不仅提高了系统的性能和可靠性，还为自动驾驶系统在复杂动态环境中的应用提供了坚实的技术基础。

[0101] 以下通过具体应用示例说明基于蒙特卡洛树搜索的扩展搜索方法：

在应用示例中，$Q(s,a)$表示在状态s下执行动作 a的期望回报，其中$\forall s \in \mathcal{S}, a \in \mathcal{A}$， $\mathcal{S}$为驾驶场景的状态空间，$\mathcal{A}$为驾驶场景的动作空间；其中，已知的未知危险场景的期望回报设为1。

[0102] 初始化搜索树，其中根节点为目标驾驶场景为状态s_0下执行动作a_0，具体流程如下：

为探索目标驾驶场景中可能存在的未知危险场景，对历史记录的每个未知危险场景均执行如下操作，包括：

对于当前场景中的当前状态s,执行如下操作,包括:

步骤A: 初始化变量:设置初始奖励值r = 0和时间步t=0,为单次仿真循环做准备。

[0103] 其中,奖励(r)用于反馈当前动作是否触发了未知危险场景,其中r=1表示触发危险,r=0表示安全。迭代次数(t)用于记录当前仿真步数,防止无限循环。

[0104] 步骤B:判断当前状态 s 在目标驾驶场景中的动作a_0对应的奖励取值是否为0,以及,判断当前的迭代次数 t 是否达到预设的最大迭代次数 T。

[0105] 如果当前状态 s 在目标驾驶场景中的动作a_0对应的奖励取值0,且第t次迭代次数未达到最大迭代次数T,则执行步骤C;否则,执行步骤D。

[0106] 步骤C、循环执行步骤C1至步骤C8;具体如下:

步骤C1: 蒙特卡洛树搜索选择阶段:

具体的,选择最优动作,从而平衡“利用已知高回报动作”与“探索未知动作”。例如,若动作$a1$的$Q(s,a1)=0.8$,但访问次数$N(s,a1)=100$,而动作$a2$的$Q(s,a2)=0.5$, $N(s,a2)=10$,则PUCT算法可能选择$a2$,因其探索潜力更大

步骤C2: 蒙特卡洛树搜索扩展阶段

当叶节点(未充分探索的子状态)访问次数超过阈值时,扩展新的子节点,丰富搜索树。其中,叶节点为搜索树中未被完全探索的末端状态节点。

[0107] 例如,若当前状态s的子节点中,某个动作路径仅被访问了5次(低于阈值10次),则继续模拟;若超过阈值,则扩展新的子节点。

[0108] 步骤C3: 蒙特卡洛树搜索模拟阶段

具体的,从叶节点执行默认策略(如随机动作或基于规则的策略),快速评估潜在回报。其中,默认策略为简化版的策略,用于快速生成仿真结果(例如随机选择加速或变道)。

[0109] 步骤C4: 蒙特卡洛树搜索回传阶段

具体的,基于步骤C3得到的奖励,从叶节点沿搜索路径会反向传播到根节点,更新搜索路径上所有节点的期望回报。

[0110] 步骤C5:利用蒙特卡洛树搜索得到的期望回报,选择任一节点对应的状态为测试状态,以及采样任一动作作为测试动作a。

[0111] 步骤C6:执行采样动作,并观测产生的状态s'与奖励。

[0112] 步骤C7:更新测试状态 s和测试动作 a 对应的期望回报;若r的取值为1,输出新发现的未知危险场景;

例如,若执行动作a后获得奖励r=1,则大幅提升Q(s,a),引导后续策略优先选择该动作。在自动驾驶危险场景搜索中,若执行动作a(如“急刹车”)后到达状态 s′(如“前方车辆停止”),则$\max_{a'} Q(s',a')$ 表示在状态 s′ 下所有可能动作(如“变道”“加速”“保持”)中,Q值的最大值。该值反映了状态 s′ 的长期安全性评估。

[0113] 步骤C8:将状态s更新为状态s',将t的取值更新为t+1,然后执行步骤C1。

[0114] 具体的,通过学习率α和折扣因子γ更新Q值,能够动态平衡即时奖励与未来价值,提升搜索效率。

[0115] 步骤D:判定是否终止针对当前状态s的搜索;其中,终止条件可以为连续失败次数或期望回报的稳定性,其中稳定性是指Q值波动小于阈值(如连续10次迭代变化量<0.01)。例如,若连续100次迭代未发现新危险场景,且Q值趋于稳定,则确定终止针对当前状态s的搜索。

[0116] 综上所述,基于蒙特卡洛树搜索的自动驾驶未知危险场景搜索方法,通过扩展已有未知危险场景的邻近状态,能够在多轮仿真中有效识别潜在的未知危险场景,从而在较短时间内发现可能的未知危险场景。

[0117] 此外,还可以通过尝试不同的未知危险场景或动作序列,可以增强探索能力。例如,在步骤C3中,动作仿真所使用的动作序列可以是不同的;在步骤C5中,不同的未知危险场景中可供选择的测试动作的动作序列可以是不同的。

[0118] 图2为本申请实施例提供的自动驾驶控制方法的流程示意图。如图2所示,所述方法包括:

步骤201、获取自动驾驶操作的管理数据,其中所述管理数据是基于上文所述的方法得到的未知危险场景得到的。

[0119] 管理数据包含了在自动驾驶过程中可能遇到的未知危险场景的具体参数和特征,例如,特定道路条件下车辆的行驶状态、周围车辆的行为模式等。

[0120] 通过获取这些管理数据,能够为后续的自动驾驶行为管理提供基础依据,帮助自动驾驶系统提前了解潜在的危险场景,从而制定相应的应对策略。

[0121] 步骤202、根据所述管理数据对自动驾驶行为进行管理。

[0122] 例如,如果管理数据显示在某个特定的未知危险场景下,车辆的刹车距离可能增加,那么自动驾驶系统可以提前降低车速,增加与前车的距离,以确保行车安全。

[0123] 通过对自动驾驶行为的管理,能够有效提高自动驾驶系统在面对未知危险场景时的应对能力,降低事故风险,提升行车安全性。

[0124] 本申请实施例提供的方法,利用未知危险场景的管理数据对自动驾驶行为进行管理,提高了自动驾驶系统的安全性和适应性,使其能够更好地应对复杂多变的路况。

[0125] 例如,通过获取和利用未知危险场景的管理数据,驾驶系统可以提前对潜在危险做出预判和反应,减少因突发情况而导致的事故。管理数据为自动驾驶系统的决策提供了更全面的信息,使其能够在复杂路况下做出更合理的决策,如合理规划路线、调整车速等。驾驶系统可以根据不同的未知危险场景数据,动态调整自身的控制策略,提高对各种复杂路况和突发情况的适应能力。

[0126] 本申请实施例还提供一种存储介质,所述存储介质中存储有计算机程序,其中,所述计算机程序被设置为运行时执行上文所述的方法。

[0127] 一种电子装置,包括存储器和处理器,所述存储器中存储有计算机程序,所述处理器被设置为运行所述计算机程序以执行上文所述的方法。

[0128] 本领域普通技术人员可以理解,上文中所公开方法中的全部或某些步骤、系统、装置中的功能模块/单元可以被实施为软件、固件、硬件及其适当的组合。在硬件实施方式中,在以上描述中提及的功能模块/单元之间的划分不一定对应于物理组件的划分;例如,一个物理组件可以具有多个功能,或者一个功能或步骤可以由若干物理组件合作执行。某些组件或所有组件可以被实施为由处理器,如数字信号处理器或微处理器执行的软件,或者被

实施为硬件,或者被实施为集成电路,如专用集成电路。这样的软件可以分布在计算机可读介质上,计算机可读介质可以包括计算机存储介质(或非暂时性介质)和通信介质(或暂时性介质)。如本领域普通技术人员公知的,术语“计算机存储介质”包括在用于存储信息(诸如计算机可读指令、数据结构、程序模块或其他数据)的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外,本领域普通技术人员公知的是,通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据,并且可包括任何信息递送介质。

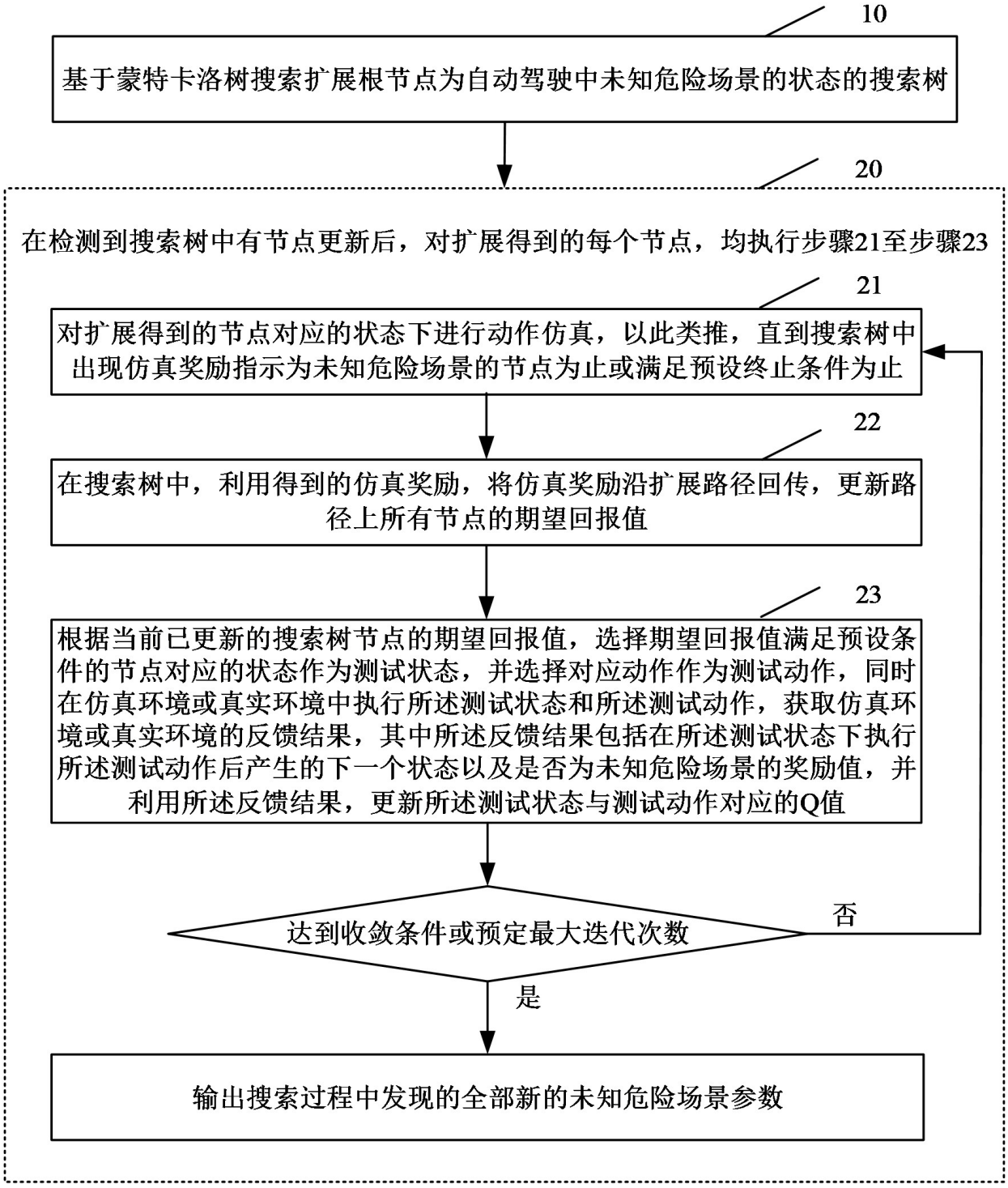

10
基于蒙特卡洛树搜索扩展根节点为自动驾驶中未知危险场景的状态的搜索树
20
在检测到搜索树中有节点更新后，对扩展得到的每个节点，均执行步骤21至步骤23
21
对扩展得到的节点对应的状态下进行动作仿真，以此类推，直到搜索树中出现仿真奖励指示为未知危险场景的节点为止或满足预设终止条件为止
22
在搜索树中，利用得到的仿真奖励，将仿真奖励沿扩展路径回传，更新路径上所有节点的期望回报值
23
根据当前已更新的搜索树节点的期望回报值，选择期望回报值满足预设条件的节点对应的状态作为测试状态，并选择对应动作作为测试动作，同时在仿真环境或真实环境中执行所述测试状态和所述测试动作，获取仿真环境或真实环境的反馈结果，其中所述反馈结果包括在所述测试状态下执行所述测试动作后产生的下一个状态以及是否为未知危险场景的奖励值，并利用所述反馈结果，更新所述测试状态与测试动作对应的Q值
达到收敛条件或预定最大迭代次数
否
是
输出搜索过程中发现的全部新的未知危险场景参数

图1

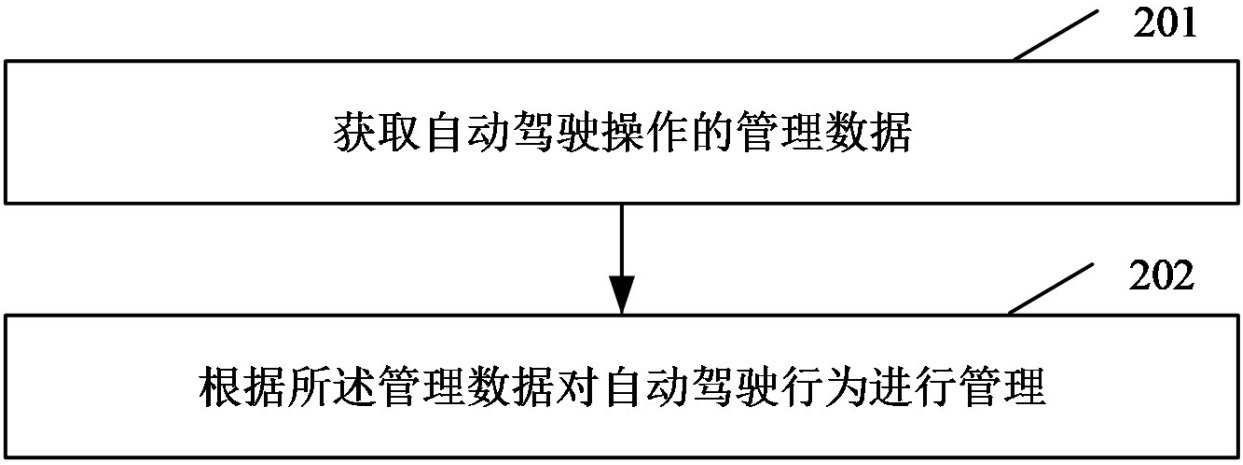


图2