(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 113837211 A
(43)申请公布日 2021.12.24

(21)申请号 202010584738.2

(22)申请日 2020.06.23

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 李栋 王滨 刘武龙 庄雨铮

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 张翠华

(51)Int.Cl.
G06K 9/62(2006.01)
G05D 1/02(2020.01)
G05D 1/00(2006.01)

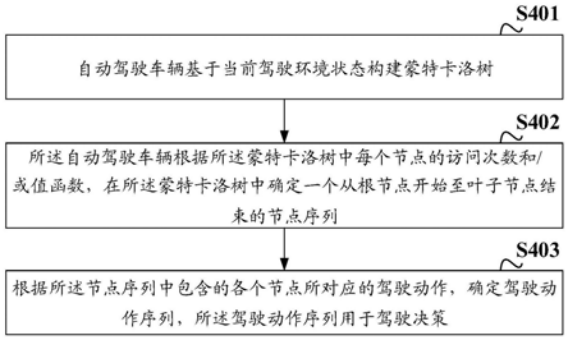
权利要求书3页 说明书12页 附图5页

(54)发明名称

一种驾驶决策方法及装置

(57)摘要

本申请涉及自动驾驶领域，公开了一种驾驶决策方法及装置，用以提高驾驶决策策略的鲁棒性，保障输出的决策结果最优。该方法包括：基于当前驾驶环境状态构建蒙特卡洛树，所述蒙特卡洛树包括根节点和N-1个非根节点，每个节点表示一个驾驶环境状态，其中根节点表示所述当前驾驶环境状态，任一非根节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述非根节点的父节点表示的驾驶环境状态，以及所述非根节点的父节点扩展所述非根节点的驾驶动作预测得到；根据蒙特卡洛树中每个节点的访问次数和/或值函数，在蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；根据所述节点序列中各个节点所对应的驾驶动作，确定驾驶动作序列。



CN 113837211 A

1.一种驾驶决策方法,其特征在于,包括:

基于当前驾驶环境状态构建蒙特卡洛树,其中所述蒙特卡洛树包含N个节点,每个节点表示一个驾驶环境状态,所述N个节点包括根节点和N-1个非根节点,其中所述根节点表示所述当前驾驶环境状态,第一节点表示的驾驶环境状态,通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态,以及第一驾驶动作预测得到,其中,所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作,所述第一节点为所述N-1个非根节点中的任意一个节点,所述N为大于或等于2的正整数;

根据所述蒙特卡洛树中每个节点的访问次数和/或值函数,在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列;

根据所述节点序列中包含的各个节点所对应的驾驶动作,确定驾驶动作序列,所述驾驶动作序列用于驾驶决策;

其中,所述每个节点的访问次数,根据所述节点的子节点的访问次数以及所述节点的初始访问次数确定,所述每个节点的值函数,根据所述节点的子节点的值函数和所述节点的初始值函数确定,其中,所述每个节点的初始访问次数为1、初始值函数根据与所述节点表示的驾驶环境状态匹配的值函数确定。

2.如权利要求1所述的方法,其特征在于,通过所述驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态,以及所述第一驾驶动作,预测得到所述第一节点表示的驾驶环境状态,包括:

通过所述驾驶环境随机模型采用Dropout前向传播,预测基于所述第一节点的父节点表示的驾驶环境状态,执行所述第一驾驶动作后的驾驶环境状态的概率分布;

从所述概率分布采样得到所述第一节点表示的驾驶环境状态。

3.如权利要求1所述的方法,其特征在于,根据与所述节点表示的驾驶环境状态匹配的值函数确定所述节点的初始值函数,包括:

从情节记忆存储库中选取与所述节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态;

根据所述第一数量的目标驾驶环境状态分别对应的值函数,确定所述节点的初始值函数。

4.如权利要求1-3中任一项所述的方法,其特征在于,所述方法还包括:

执行所述驾驶动作序列中的第一个驾驶动作后,获取执行所述第一个驾驶动作后的真实驾驶环境状态;

根据所述当前驾驶环境状态、所述第一个驾驶动作以及执行所述第一个驾驶动作后的真实驾驶环境状态,对所述驾驶环境随机模型进行更新。

5.如权利要求1所述的方法,其特征在于,所述根据所述蒙特卡洛树中每个节点的访问次数和/或值函数,在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列,包括:

根据所述蒙特卡洛树中每个节点的访问次数,按照访问次数最大原则,在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列;或,

根据所述蒙特卡洛树中每个节点的值函数,按照值函数最大原则,在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列;或,

根据所述蒙特卡洛树中每个节点的访问次数和值函数，按照访问次数最大优先、值函数最大次之原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列。

6.如权利要求3所述的方法，其特征在于，所述方法还包括：

当驾驶情节结束后，根据驾驶结果，确定所述驾驶情节中每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值；

将每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值作为其对应的值函数，更新所述情节记忆存储库。

7.一种驾驶决策装置，其特征在于，包括：

构建单元，用于基于当前驾驶环境状态构建蒙特卡洛树，其中所述蒙特卡洛树包含N个节点，每个节点表示一个驾驶环境状态，所述N个节点包括根节点和N-1个非根节点，其中所述根节点表示所述当前驾驶环境状态，第一节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及第一驾驶动作预测得到，其中，所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作，所述第一节点为所述N-1个非根节点中的任意一个节点，所述N为大于或等于2的正整数；

确定单元，用于根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；以及根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用于驾驶决策；

其中，所述每个节点的访问次数，根据所述节点的子节点的访问次数以及所述节点的初始访问次数确定，所述每个节点的值函数，根据所述节点的子节点的值函数和所述节点的初始值函数确定，其中，所述每个节点的初始访问次数为1、初始值函数根据与所述节点表示的驾驶环境状态匹配的值函数确定。

8.如权利要求7所述的装置，其特征在于，所述构建单元通过所述驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及所述第一驾驶动作，预测得到所述第一节点表示的驾驶环境状态时，具体用于：

通过所述驾驶环境随机模型采用Dropout前向传播，预测基于所述第一节点的父节点表示的驾驶环境状态，执行所述第一驾驶动作后的驾驶环境状态的概率分布；从所述概率分布采样得到所述第一节点表示的驾驶环境状态。

9.如权利要求7所述的装置，其特征在于，所述构建单元根据与所述节点表示的驾驶环境状态匹配的值函数确定所述节点的初始值函数时，具体用于：

从情节记忆存储库中选取与所述节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态；根据所述第一数量的目标驾驶环境状态分别对应的值函数，确定所述节点的初始值函数。

10.如权利要求7-9中任一项所述的装置，其特征在于，所述装置还包括：

更新单元，用于执行所述驾驶动作序列中的第一个驾驶动作后，获取执行所述第一个驾驶动作后的真实驾驶环境状态；根据所述当前驾驶环境状态、所述第一个驾驶动作以及执行所述第一个驾驶动作后的真实驾驶环境状态，对所述驾驶环境随机模型进行更新。

11.如权利要求7所述的装置，其特征在于，所述确定模块根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列时，具体用于：

根据所述蒙特卡洛树中每个节点的访问次数，按照访问次数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的值函数，按照值函数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的访问次数和值函数，按照访问次数最大优先、值函数最大次之原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列。

12.如权利要求9所述的装置，其特征在于，所述装置还包括：

更新单元，用于当驾驶情节结束后，根据驾驶结果，确定所述驾驶情节中每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值；将每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值作为其对应的值函数，更新所述情节记忆存储库。

13.一种驾驶决策装置，其特征在于，包括：

所述存储器，存储有计算机程序或指令；

所述处理器，用于调用所述存储器中存储的计算机程序或指令，执行如权利要求1-6中任一项所述的方法。

14.一种计算机存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序或指令，当所述计算机程序或指令被驾驶决策装置执行时，实现如权利要求1-6中任一项所述的方法。

一种驾驶决策方法及装置

技术领域

[0001] 本申请实施例涉及自动驾驶领域，尤其涉及一种驾驶决策方法及装置。

背景技术

[0002] 随着自动驾驶时代的到来，具有自动驾驶功能的智能车辆成为各大厂商研究的重点。目前，自动驾驶包括辅助驾驶和完全自动驾驶，其实现的关键技术有：环境感知、驾驶决策以及控制执行等。其中，驾驶决策根据感知到的交通参与者信息，给出驾驶动作，供车辆执行。

[0003] 目前，驾驶决策通常是基于强化学习(deep Q network，DQN)算法实现的，通过DQN算法构建的DQN模型，对大量车辆在某一时刻的驾驶环境状态(如自身车辆速度、相邻车辆速度等)，以及车辆驾驶者基于该时刻的驾驶环境状态输出的驾驶动作(如向左变道、向右变道等)进行学习。在驾驶决策时，即可根据车辆当前的驾驶环境状态，通过DQN模型，得到自动驾驶车辆需要执行的驾驶动作。

[0004] 然而，基于DQN模型的驾驶决策，在输出驾驶动作时，未考虑到输出的驾驶动作对车辆后续驾驶的影响，难以保证输出的决策结果最优。

发明内容

[0005] 本申请实施例提供一种驾驶决策方法及装置，用以提高驾驶决策策略的鲁棒性，保障输出的决策结果最优。

[0006] 第一方面，本申请实施例提供了一种驾驶决策方法，该方法包括：基于当前驾驶环境状态构建蒙特卡洛树，其中所述蒙特卡洛树包含N个节点，每个节点表示一个驾驶环境状态，所述N个节点包括根节点和N-1个非根节点，其中所述根节点表示所述当前驾驶环境状态，第一节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及第一驾驶动作预测得到，其中，所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作(也即所述第一驾驶动作为所述第一节点的父节点扩展所述第一节点的驾驶动作)，所述第一节点为所述N-1个非根节点中的任意一个节点，所述N为大于或等于2的正整数；根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用于驾驶决策；其中，所述每个节点的访问次数，根据所述节点的子节点的访问次数以及所述节点的初始访问次数确定，所述每个节点的值函数，根据所述节点的子节点的值函数和所述节点的初始值函数确定，其中，所述每个节点的初始访问次数为1、初始值函数根据与所述节点表示的驾驶环境状态匹配的值函数确定。

[0007] 采用上述方法，通过驾驶环境随机模型，可以预测自动驾驶车辆在未来一段时间内多步驾驶动作可能面临的各种驾驶环境状态，构建蒙特卡洛树，并基于自动驾驶车辆在未来一段时间内多步驾驶动作可能面临的各种驾驶环境状态，确定最有利于车辆行驶的驾

驶动作序列，提高了驾驶决策策略的鲁棒性，有利于保障输出的决策结果最优。

[0008] 在一个可能的设计中，通过所述驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及所述第一驾驶动作，预测得到所述第一节点表示的驾驶环境状态，包括：通过所述驾驶环境随机模型采用Dropout前向传播，预测基于所述第一节点的父节点表示的驾驶环境状态，执行所述第一驾驶动作后的驾驶环境状态的概率分布；从所述概率分布采样得到所述第一节点表示的驾驶环境状态。

[0009] 上述设计中，通过驾驶环境随机模型采用Dropout前向传播的方式，预测基于所述第一节点的父节点表示的驾驶环境状态，执行第一驾驶动作后驾驶环境状态的概率分布；并从概率分布中采样得到第一节点表示的驾驶环境状态，在扩展蒙特卡洛树中节点时充分考虑了驾驶环境状态的不确定性，增加了节点扩展的多样性，使得驾驶决策策略更具鲁棒性。

[0010] 在一个可能的设计中，根据与所述节点表示的驾驶环境状态匹配的值函数确定所述节点的初始值函数，包括：从情节记忆存储库中选取与所述节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态；根据所述第一数量的目标驾驶环境状态分别对应的值函数，确定所述节点的初始值函数。

[0011] 上述设计中，通过为蒙特卡洛树引入情节记忆存储库，能够根据历史经验数据准确快速地估计节点的初始值函数，避免通过低效展开的方式估计节点的初始值函数，减少了估计节点初始值函数带来的计算开销，有利于提高蒙特卡洛树搜索的效率。

[0012] 在一个可能的设计中，所述方法还包括：执行所述驾驶动作序列中的第一个驾驶动作后，获取执行所述第一个驾驶动作后的真实驾驶环境状态；根据所述当前驾驶环境状态、所述第一个驾驶动作以及执行所述第一个驾驶动作后的真实驾驶环境状态，对所述驾驶环境随机模型进行更新。

[0013] 上述设计中，可以不断对驾驶环境随机模型进行训练更新，有利于提高驾驶环境随机模型的准确性。

[0014] 在一个可能的设计中，所述根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列，包括：根据所述蒙特卡洛树中每个节点的访问次数，按照访问次数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的值函数，按照值函数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的访问次数和值函数，按照访问次数最大优先、值函数最大次之原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列。

[0015] 上述设计中，提供多种节点序列的确定方式，有利于满足不同驾驶决策的需求。

[0016] 在一个可能的设计中，所述方法还包括：当驾驶情节结束后，根据驾驶结果，确定所述驾驶情节中每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值；将每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值作为其对应的值函数，更新所述情节记忆存储库。

[0017] 上述设计中，根据累积奖赏回报值更新情节记忆存储库，有利于准确确定蒙特卡洛树中节点的初始值函数，从而保证输出的决策结果的可靠性。

[0018] 第二方面，本申请实施例提供了一种驾驶决策装置，包括用于执行上述第一方面或者第一方面的任一种可能的设计中各个步骤的单元。

[0019] 第三方面，本申请实施例提供了一种驾驶决策装置，包括处理器和存储器，其中所述存储器用于存储计算机程序或指令，所述处理器用于调用所述存储器中存储的计算机程序或指令，执行上述第一方面或者第一方面的任一种可能的设计中所述的方法。

[0020] 第四方面，本申请实施例提供一种计算机可读存储介质，所述计算机可读存储介质具有用于执行上述第一方面或者第一方面的任一种可能的设计中所述的方法的计算机程序或指令。

[0021] 第五方面，本申请实施例还提供一种计算机程序产品，包括计算机程序或指令，当所述计算机程序或指令被执行时，可以实现上述第一方面或者第一方面的任一种可能的设计中所述的方法。

[0022] 第六方面，本申请还提供一种芯片，所述芯片用于实现上述第一方面或者第一方面的任一种可能的设计中所述的方法。

[0023] 上述第二方面至第六方面所能达到的技术效果请参照上述第一方面所能达到的技术效果，这里不再重复赘述。

附图说明

[0024] 图1为本申请实施例提供的自动驾驶系统的示意图；

[0025] 图2为本申请实施例提供的树形结构示意图；

[0026] 图3为本申请实施例提供的Dropout处理效果示意图；

[0027] 图4为本申请实施例提供的驾驶决策过程示意图；

[0028] 图5为本申请实施例提供的自动驾驶车辆所处环境示意图；

[0029] 图6为本申请实施例提供的蒙特卡洛树结构示意图之一；

[0030] 图7为本申请实施例提供的蒙特卡洛树结构示意图之二；

[0031] 图8为本申请实施例提供的驾驶决策装置示意图之一；

[0032] 图9为本申请实施例提供的驾驶决策装置示意图之二。

具体实施方式

[0033] 图1为本申请实施例可以应用的一种自动驾驶系统的示意图，包括环境感知模块、基于高效蒙特卡洛树搜索(monte-carlo tree search，MCTS)的行为决策模块、路径规划模块、控制执行模块，其中基于高效MCTS的行为决策模块是本申请实施例中用于驾驶决策的核心模块。

[0034] 环境感知模块：该模块的作用是感知周围环境状态，获取自车的当前驾驶环境状态。具体地，综合各个传感器所传递的信息，判断自车状态，包括自车的位置、速度、航向角等；判断周围环境状态，包括静态目标(车道线及道路边界、绿化带等)的位置、动态目标(社会车辆及行人等)位置和速度、路面条件、天气状况(温度、湿度、光线)等信息，并将这些信息传递给路径规划模块和基于高效MCTS的行为决策控制模块。

[0035] 基于高效MCTS的行为决策模块：基于自车的当前驾驶环境状态，输出未来多步的驾驶动作序列。

[0036] 路径规划模块:根据自车和社会车辆状态以及确定的驾驶动作序列等,规划出一条用于完成驾驶动作序列的带有位置和速度信息的路径轨迹。

[0037] 控制执行模块:该模块通过控制自车的方向盘和油门刹车踏板等,输出合理的控制量,使得自车行驶的轨迹能够跟踪所规划的路径轨迹。

[0038] 本申请实施例提供的驾驶决策方案,可以适用于应用有如图1所示的自动驾驶系统的车载设备、自动驾驶车辆等。在本申请实施例的后续说明中,以应用有如图1所示的自动驾驶系统的自动驾驶车辆为例进行说明。

[0039] 在介绍本申请实施例之前,首先对本申请中的部分用语进行解释说明,以便于本领域技术人员理解。

[0040] 1)、树形结构(树),是一种数据结构,是由n个节点组成的一个具有层次关系的集合,n通常为不小于1的正整数。把它叫做“树”是因为它看起来像一棵倒挂的树,也就是说它是根朝上,而叶朝下的。其中树中没有子节点的节点为树的叶子节点,树的最顶端的节点为树的根节点。例如:在图2所示的树形结构中,节点11位于树的第一层;节点21、节点22和节点23位于树的第二层;节点31、节点32、节点33和节点34位于树的第三层,节点11为树的根节点,节点31、节点32、节点33、节点34和节点23为树的叶子节点。在本申请实施例中所涉及的蒙特卡洛树也是一种树形结构。

[0041] 2)、蒙特卡洛树搜索(MCTS),是一种用于某些决策过程的启发式搜索算法,通常在游戏中使用,主要目的是在一个给出的游戏状态下,选择出胜率最高的下一步。蒙特卡洛树搜索的主要概念是搜索,即沿着蒙特卡洛树(也可以称为博弈树)向下的一组遍历过程。单次遍历的路径会从根节点(也即当前博弈状态)延伸到没有完全展开的节点,未完全展开的节点表示其子节点至少有一个未访问到(或未被扩展出)。遇到未完全展开的节点时,它的一个未访问子节点将会被扩展出来,扩展出来的子节点会按照一定的策略计算其对应的初始值函数,其中初始值函数也可以称之为初始价值或模拟结果。例如对于下棋游戏来说可以在扩展出的子节点表示的棋局的基础上,按照快速走子策略(rollout policy)等,一直走到游戏结束,得到一个模拟结果(如输或赢),也即得到该子节点的初始值函数(如输为0/1,赢为1/1)。得到该子节点的初始值函数后,将会反向传播回当前蒙特卡洛树的根节点,将该子节点的初始值函数(模拟结果)加到该子节点所有的祖先节点中,例如该子节点的初始值函数是0/1(代表游戏输了),那么就把这个节点的所有祖先节点的值函数加0/1。一旦到达蒙特卡洛树的搜索时间或最大搜索步长(即从根节点开始最大扩展的节点数量),则搜索终止,基于蒙特卡洛树中每个节点的值函数进行决策。需要说明的是,在蒙特卡洛树中每个节点表示一种状态(也可以称为局面,如棋局),并针对每个节点记录有该节点的统计数据,如节点的值函数、访问次数等,其中针对节点记录的值函数,也可以称为节点的价值,可以是一个具体的数值。

[0042] 3)、Dropout,指以概率P舍弃模型中的部分神经元,舍弃的神经元的输出被设置为零。如图3所示,为某一模型经过Dropout处理的示意图。其中,图3(A)为经过Dropout处理前模型的结构示意图,图3(B)为经过Dropout处理后模型的结构示意图,可见经过Dropout处理后的模型部分神经元被临时舍弃,输出为零。

[0043] 另外,需要理解的是,在本申请实施例中,至少一个还可以描述为一个或多个,多个可以是两个、三个、四个或者更多个,本申请不做限制。在本申请实施例中,“/”可以表示

前后关联的对象是一种“或”的关系，例如，A/B可以表示A或B；“和/或”可以用于描述关联对象存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况，其中A，B可以是单数或者复数。在本申请实施例中，“示例性的”或者“例如”等词用于表示例子、例证或说明，被描述为“示例性的”或者“例如”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。使用“示例性的”或者“例如”等词旨在以具体方式呈现相关概念，便于理解。

[0044] 本申请实施例旨在通过驾驶环境随机模型，预测自动驾驶车辆在未来一段时间内多步驾驶动作可能面临的各种驾驶环境状态，构建蒙特卡洛树，并基于自动驾驶车辆在未来一段时间内多步驾驶动作可能面临的各种驾驶环境状态，确定最有利于车辆行驶的驾驶动作序列，用以提高驾驶决策策略的鲁棒性，保障输出的决策结果最优。同时，还可以通过情节记忆存储库估计蒙特卡洛树中节点的初始值函数，从而减少估计节点初始值函数带来的计算开销，提高蒙特卡洛树搜索的效率。

[0045] 下面以换道决策场景为例，结合附图详细说明本申请实施例。

[0046] 图4为本申请实施例提供的一种驾驶决策过程示意图，该过程包括：

[0047] S401：自动驾驶车辆基于当前驾驶环境状态构建蒙特卡洛树。

[0048] 其中，所述蒙特卡洛树包含N个节点，每个节点表示一个驾驶环境状态，所述N个节点包括根节点和N-1个非根节点，其中所述根节点表示所述当前驾驶环境状态，第一节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及第一驾驶动作预测得到，其中，所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作，所述第一节点为所述N-1个非根节点中的任意一个节点，所述N为大于或等于2的正整数。

[0049] 在换道决策场景中，自动驾驶车辆的环境状态可以包括自动驾驶车辆的速度、位于自动驾驶车辆前方(自动驾驶车辆行进方向)的社会车辆相对于自动驾驶车辆的相对速度和相对距离、位于自动驾驶车辆后方(自动驾驶车辆行进方向的反方向)的社会车辆的到达时长(即追上自动驾驶车辆所需的时长)等。自动驾驶车辆的驾驶环境状态可以由包含当前时刻在内的T个历史时刻的环境状态构成，T为大于或等于1的正整数。示例的：以当前时刻为10:00:00为例，自动驾驶车辆的当前驾驶环境状态可以由自动驾驶车辆在9:56:30、9:57:00、9:57:30、9:58:00、9:58:30、9:59:00、9:59:30、10:00:00共8个历史时刻的环境状态构成。

[0050] 对于自动驾驶车辆的环境状态，可以通过自动驾驶车辆的车载传感器获取。示例的：自动驾驶车辆的车载传感器包括车速传感器、加速度传感器、测距传感器(如雷达测距传感器)等，车速传感器能够测量自动驾驶车辆的速度、加速度传感器能够测量自动驾驶车辆的加速度、距离传感器可以测量自动驾驶车辆与社会车辆的相对距离。自动驾驶车辆可以根据自车与社会车辆的相对距离变化，确定社会车辆相对于自车的相对速度；并可以根据社会车辆与自车的相对距离以及社会车辆相对于自车的相对速度，确定自车后方社会车辆的到达时长。

[0051] 在一种可能的实施中，自动驾驶车辆对于社会车辆相对于自车的相对速度或相对距离的获知，还可以通过与社会车辆通信实现。示例的：自动驾驶车辆可以接收社会车辆发送的速度以及位置信息，并根据自车的速度以及位置信息，确定社会车辆相对于自车的相

对速度以及相对距离。

[0052] 以某一时刻自动驾驶车辆所处环境如图5所示为例，自动驾驶车辆可以获取自车前方三个车道上社会车辆相对于自车的相对距离(△D)和相对速度(△V)，获取自车后方三个车道上社会车辆的到达时长(TTA)，以及自车的速度(V)。其中图5中L、M、R分别表示自动驾驶车辆的左、中、右三个车道的社会车辆。在本申请实施例中，可以将当前时刻对应的T个历史时刻的环境状态作为当前驾驶环境状态，在进行驾驶决策时，自动驾驶车辆可以自动驾驶车辆当前驾驶环境状态作为蒙特卡洛树的根节点，构建蒙特卡洛树。假设图5为自动驾驶车辆当前所处环境，则自动驾驶车辆的当前驾驶环境状态(S_t)可以表示为S_t＝(△$D_{i,t-T:t}$，△$V_{i,t-T:t}$，$V_{t-T:t}$，$TTA_{j,t-T:t}$)，其中i∈(L,M,R)，j∈(L,R)，t表示当前时刻，t-T:t表示当前时刻对应的T个历史时刻。

[0053] 在换道决策场景下，自动驾驶车辆可选的驾驶动作包括向左换道、保持直行、向右换道中的一种或多种。对于驾驶环境随机模型的训练，可以通过预先采集的大量车辆在执行某一驾驶动作前和后的驾驶环境状态实现。具体的，可以预先采集大量车辆在执行某一驾驶动作前和后的驾驶环境状态，以及执行的驾驶动作作为样本对，构建对驾驶环境随机模型进行训练的训练集。其中训练集中的每个样本对可以表示为(S_t+A_t，S_{t+1})，其中A_t表示车辆执行的一个驾驶动作，可以为向左换道、保持直行、向右换道中的一种，S_t表示车辆执行A_t前的驾驶环境状态，S_{t+1}表示车辆执行A_t后的驾驶环境状态。

[0054] 需要理解的是，在本申请实施例中车辆执行驾驶动作前的驾驶环境状态，通常是指车辆开始执行该驾驶动作时车辆的驾驶环境状态，车辆执行驾驶动作后的驾驶环境状态，通常是指车辆执行该驾驶动作结束时车辆的驾驶环境状态。示例的：车辆在10:10:00-10:10:10执行向左变道的驾驶动作，车辆在10:10:00时的驾驶环境状态，可以作为车辆在执行向左变道的驾驶动作前的驾驶环境状态，车辆在10:10:10时的驾驶环境状态，可以作为车辆在执行向左变道的驾驶动作后的驾驶环境状态。

[0055] 在对驾驶环境随机模型($f_\theta(s,a,z)$)进行训练时，可以将样本对中的S_t+A_t输入到驾驶环境随机模型，得到驾驶环境随机模型输出的基于S_t执行A_t后的驾驶环境状态(S_{t+1}')。根据驾驶环境模型输出的S_{t+1}'与样本对中真实的S_{t+1}，通过损失函数(loss function)可以计算驾驶环境随机模型的损失(loss)，loss越高表示驾驶环境随机模型输出的S_{t+1}'与真实的S_{t+1}的差异越大，驾驶环境随机模型根据loss调整驾驶环境随机模型中的参数，如采用随机梯度下降法更新驾驶环境随机模型中神经元的参数，那么对驾驶环境随机模型的训练过程就变成了尽可能缩小这个loss的过程。通过训练集中的样本对不断对驾驶环境随机模型进行训练，当这个loss缩小至预设范围，即可得到训练完成的驾驶环境随机模型。其中驾驶环境随机模型的隐变量z可以用于表征模型的不确定性，驾驶环境随机模型可选为深度神经网络、贝叶斯神经网络等。

[0056] 基于训练完成的驾驶环境随机模型，自动驾驶车辆基于当前驾驶环境状态，可以预测自动驾驶车辆在未来一段时间内多步驾驶动作可能面临的各种驾驶环境状态，构建蒙特卡洛树。具体的，自动驾驶车辆可以从蒙特卡洛树的根节点开始逐层选择节点，当选择的目标节点存在未被预测的一个或多个可选驾驶动作时，选择一个目标驾驶动作，通过驾驶环境随机模型预测自动驾驶车辆基于目标节点执行目标驾驶动作后的驾驶环境状态，作为所述目标节点的一个子节点表示的驾驶环境状态，扩展蒙特卡洛树。

[0057] 在一种可能的实施中，为了充分考虑驾驶环境中其它社会车辆的不确定性，在通过驾驶环境随机模型预测自动驾驶车辆基于目标节点执行目标驾驶动作后的驾驶环境状态时，通过驾驶环境随机模型，采用Dropout前向传播的方式，得到自动驾驶车辆基于目标节点执行目标驾驶动作后的驾驶环境状态的概率分布，并从概率分布采样获取驾驶环境状态，作为自动驾驶车辆基于目标节点执行目标驾驶动作后的驾驶环境状态。

[0058] 示例的：自动驾驶车辆已扩展的蒙特卡洛树如图6所示，自动驾驶车辆从蒙特卡洛树的根节点11开始逐层选择节点(如逐层遍历选取等)，选择到节点21，节点21未被预测的可选驾驶动作(A)包括向左变道、保持直接和向右变道，自动驾驶车辆可以在向左变道、保持直接和向右变道中随机选择一个目标驾驶动作。例如：在向左变道、保持直接和向右变道中随机选择到目标驾驶动作向左变道(A_t)。选择目标驾驶动作后，自动驾驶车辆基于驾驶环境随机模型，通过多次采用Dropout前向传播的方式，预测在节点21的基础上，执行目标驾驶动作(A_t)后的多个可能的驾驶环境状态，根据所述多个可能的驾驶环境状态，计算执行所述A_t后的驾驶环境状态预测值的均值(μ)和方差(δ^2)。基于所述均值和方差通过高斯采样得到执行所述A_t后的驾驶环境状态值的概率分布$s'=N(\mu,\delta^2)$，并从概率分布中采样获取(如随机抽取一个)执行所述A_t后的驾驶环境状态，扩展节点21的一个子节点(节点31)。

[0059] 在扩展出目标节点的一个子节点后，自动驾驶车辆初始化所述子节点的统计数据访问次数(N)和值函数(Q)。也即需要确定所述子节点的初始访问次数和初始值函数。对于扩展出的子节点的初始访问次数，自动驾驶车辆将所述子节点的初始访问次数设置为1，并从所述子节点开始向根节点回溯，更新所述子节点对应的节点路径(搜索路径)上每个节点的访问次数。例如：将扩展出的子节点对应的节点路径上的每个节点的访问次数+1，即N＝N'+1，其中N为更新后的访问次数，N'为更新前的访问次数，也即蒙特卡洛树中每个节点的访问次数为该节点所有子节点的访问次数与该节点的初始访问次数的和。需要理解的是，在本申请中节点对应的节点路径(搜索路径)指由节点的所有祖先节点构成的节点路径。示例的：自动驾驶车辆扩展节点21的子节点(节点31)后，将节点31的初始访问次数更新为1，将节点31节点路径(搜索路径)上的节点21和节点11的访问次数分别+1，完成对节点31对应的节点路径上每个节点的访问次数的更新。

[0060] 对于扩展出的子节点的初始值函数，自动驾驶车辆可以根据情节记忆存储库(EM)确定所述子节点的初始值函数。如果所述子节点表示的驾驶环境状态在情节记忆存储库有记载，直接输出情节记忆存储库中存储的所述驾驶环境状态对应的值函数，作为所述子节点的初始值函数；否则可以从情节记忆存储库中选取与所述子节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态；根据所述第一数量的目标驾驶环境状态分别对应的值函数，确定所述子节点的初始值函数。其中所述第一数量(K)可以为3、5等。例如：自动驾驶车辆可以从情节记忆存储库中选取与扩展出的子节点表示的驾驶环境状态匹配度最高的K个目标驾驶环境状态，将所述K个目标驾驶环境状态分别对应的值函数的均值，作为所述子节点的初始值函数。

[0061] 确定扩展出的子节点的初始值函数后，自动驾驶车辆从所述子节点开始向根节点回溯，更新所述子节点对应的节点路径上每个节点的值函数。示例的：自动驾驶车辆可以根据$Q=Q'+Q_L$对所述子节点对应的节点路径上每个节点的值函数进行更新，其中Q为节点更新后的值函数、Q'为节点更新前的值函数、Q_L为所述子节点的初始值函数，即蒙特卡洛树中

每个节点的值函数为该节点所有子节点的值函数与该节点的初始值函数的和。在另一种可能的实施中，也可以根据Q＝Q'+(Q$_L$-Q')/N对所述子节点对应的节点路径上每个节点的值函数进行更新，其中Q为节点更新后的值函数、Q'为节点更新前的值函数、Q$_L$为所述子节点的初始值函数、N为节点更新后的访问次数。

[0062] 在本申请实施例中，可以限制扩展蒙特卡洛树的最大步数，即限定从蒙特卡洛树的根节点开始最大扩展的节点数量。如限制扩展蒙特卡洛树的最大步数为20步，当到达最大步数时停止扩展蒙特卡洛树。另外，在本申请实施例中，每扩展一次蒙特卡洛树的一个叶子节点(即扩展蒙特卡洛树中某个节点的一个子节点)后，如果扩展蒙特卡洛树的步数未满足扩展蒙特卡洛树的最大步数，返回从蒙特卡洛树的根节点开始逐层选择节点的步骤，继续扩展蒙特卡洛树的叶子节点。其中从蒙特卡洛树的根节点开始逐层选择节点，其选择方式可以为根据上限置信区间算法(upper confidence bound apply to tree，UCT)的选择策略，逐层选择驾驶动作对应的节点(st+1)。被选驾驶动作(A$_t$)较其他可选驾驶动作，满足最大化该驾驶动作对应值函数(Q)和探索加权项$C\sqrt{N''}/N$之和，使得被选驾驶动作在最大化值函数的节点和低访问次数节点之间得到平衡，保证驾驶动作选择的最优性。其中Q为选择节点的值函数，所述探索加权项中，C为探索项的权重系数，N”为选择节点的访问次数，N为所述选择节点对应可选动作(A)的叶子节点的访问次数。

[0063] S402：所述自动驾驶车辆根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从根节点开始至叶子节点结束的节点序列。

[0064] 在本申请实施例中，可以按照访问次数最大原则，值函数最大原则，或访问次数最大优先、值函数最大次之原则中的一个，在蒙特卡洛树中确定一个从根节点开始至叶子节点结束的节点序列。

[0065] 示例的，以按照访问次数最大原则，在蒙特卡洛树中确定一个从根节点开始至叶子节点结束的节点序列为例，参照图7所示，从节点11(根节点)开始，节点11最大访问次数的子节点为节点21，节点21最大访问次数的子节点为节点31，节点31最大访问次数的子节点为节点41，确定节点序列为节点11-节点21-节点31-节点41。

[0066] S403：根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用于驾驶决策。

[0067] 在本申请实施例中，蒙特卡洛树中节点所对应的驾驶动作，为该节点的父节点扩展该节点的驾驶动作，其中根节点没有所对应的驾驶动作。示例的：参照图7所示，节点22的父节点11扩展节点22的驾驶动作为保持直行，则节点22所对应的驾驶动作为保持直行。以确定的节点序列为节点11-节点21-节点31-节点41为例，则驾驶动作序列为向左变道-向左变道-保持直行。

[0068] 为了保证驾驶环境随机模型的可靠性，在一种可能的实施中，当自动驾驶车辆执行驾驶动作序列中第一个驾驶动作后，可以将自动驾驶车辆执行所述第一个驾驶动作前的驾驶环境状态+所述第一驾驶动作和执行所述第一个驾驶动作后的驾驶环境状态作为一个新样本对(S$_t$+A$_t$，S$_{t+1}$)，补充到训练集中，对驾驶环境随机模型进行更新。

[0069] 示例的：驾驶环境随机模型可通过下述方式更新。从补充到训练集的新样本中采样最小批样本$\{(s_i, a_i, s_i')\}_{i=1}^{M}$，其中M为最小批样本集中的样本数量，(s，a，s')表示一个样

本对，如一个(S_t+A_t，S_{t+1})。计算驾驶环境随机模型的损失函数，并根据随机梯度下降法更新驾驶环境随机模型，使得驾驶环境随机模型的预测值与实际感知结果误差最小。其中，损失函数可以为 $L(\theta,\phi)=\frac{1}{M}\sum_{i=1}^{M}||f_\theta(s_i,a_i,z_i)-s_i'||_2^2+\lambda D_{KL}[q_\phi(z;s_i,s_i')||p(z)]$，由两项组成。第一项 $||f_\theta(s_i,a_i,z_i)-s_i'||_2^2$ 采用均方误差，表示驾驶环境随机模型对于自动驾驶车辆执行驾驶动作后的驾驶环境状态的预测值$f_\theta(s_i,a_i,z_i)$与真实观测值s'_i之间的逼近误差。第二项 $\lambda D_{KL}[q_\phi(z;s_i,s_i')||p(z)]$是对隐变量z引入的正则化项，约束隐变量z的估计分布$q_\phi(z|s_i,s'_i)$与先验假设分布p(z)之间的KL散度，从而以防止过拟合，λ为调节正则化强度的比例系数。

[0070] 当驾驶情节结束后，如自动驾驶车辆到达目的地或者自动驾驶车辆出现意外，停止驾驶后，自动驾驶车辆可以根据是否到达目的地等得到一个回报值，并可以根据驾驶情节中每一驾驶动作执行后真实的驾驶环境状态，确定情节轨迹序列(按时间先后顺序)。对于反向情节轨迹序列(情节轨迹序列由后往前)中的每一步，自动驾驶车辆根据R＝γR'+r计算反向情节轨迹序列中每一驾驶环境状态(S_t)对应的驾驶环境状态动作对(S_t，A_t)的折扣累积回报，其中A_t为自动驾驶车辆在S_t的基础上执行的驾驶动作，R为驾驶环境状态动作对(S_t，A_t)的回报值(也即累积奖赏回报值)，R'为下一驾驶环境状态动作对的回报值，γ为折扣因子，r为在驾驶环境状态(S_t)下执行驾驶动作A_t后所得的奖赏函数。若驾驶环境状态动作对(S_t，A_t)已存在于情节记忆存储库(EM)中，则将EM中的值函数更新为R与存储值之间的较大者；否则，直接将新的样本对(S_t，A_t)和R写入EM，R即为(S_t，A_t)中的驾驶环境状态((S_t)对应的值函数。

[0071] 上述主要从方法流程的角度对本申请提供的方案进行了介绍。可以理解的是，为了实现上述功能，装置可以包括执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0072] 在采用集成的单元的情况下，图8示出了本申请实施例中所涉及的驾驶决策装置的可能的示例性框图，该驾驶决策装置800可以以软件的形式存在。驾驶决策装置800可以包括：构建单元801、确定单元802以及更新单元803。

[0073] 具体地，在一个实施例中，构建单元801，用于基于当前驾驶环境状态构建蒙特卡洛树，其中所述蒙特卡洛树包含N个节点，每个节点表示一个驾驶环境状态，所述N个节点包括根节点和N-1个非根节点，其中所述根节点表示所述当前驾驶环境状态，第一节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及第一驾驶动作预测得到，其中，所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作，所述第一节点为所述N-1个非根节点中的任意一个节点，所述N为大于或等于2的正整数；

[0074] 确定单元802，用于根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；以及根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用

于驾驶决策；

[0075] 其中，所述每个节点的访问次数，根据所述节点的子节点的访问次数以及所述节点的初始访问次数确定，所述每个节点的值函数，根据所述节点的子节点的值函数和所述节点的初始值函数确定，其中，所述每个节点的初始访问次数为1、初始值函数根据与所述节点表示的驾驶环境状态匹配的值函数确定。

[0076] 在一种可能的设计中，所述构建单元801通过所述驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及所述第一驾驶动作，预测得到所述第一节点表示的驾驶环境状态时，具体用于：通过所述驾驶环境随机模型采用Dropout前向传播，预测基于所述第一节点的父节点表示的驾驶环境状态，执行所述第一驾驶动作后驾驶环境状态的概率分布；从所述概率分布采样得到所述第一节点表示的驾驶环境状态。

[0077] 在一种可能的设计中，所述构建单元801根据与所述节点表示的驾驶环境状态匹配的值函数确定所述节点的初始值函数时，具体用于：从情节记忆存储库中选取与所述节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态；根据所述第一数量的目标驾驶环境状态分别对应的值函数，确定所述节点的初始值函数。

[0078] 在一种可能的设计中，更新单元803，用于执行所述驾驶动作序列中的第一个驾驶动作后，获取执行所述第一个驾驶动作后的真实驾驶环境状态；根据所述当前驾驶环境状态、所述第一个驾驶动作以及执行所述第一个驾驶动作后的真实驾驶环境状态，对所述驾驶环境随机模型进行更新。

[0079] 在一种可能的设计中，所述确定模块根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列时，具体用于：根据所述蒙特卡洛树中每个节点的访问次数，按照访问次数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的值函数，按照值函数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的访问次数和值函数，按照访问次数最大优先、值函数最大次之原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列。

[0080] 在一种可能的设计中，更新单元803，还用于当驾驶情节结束后，根据驾驶结果，确定所述驾驶情节中每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值；将每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值作为其对应的值函数，更新所述情节记忆存储库。

[0081] 基于上述驾驶决策方法，本申请实施例还提供一种驾驶决策装置，如图9所示，所述驾驶决策装置900包括存储器901和处理器902，所述存储器901和所述处理器902之间相互连接，可选的，所述存储器901与所述处理器902之间可以通过总线相互连接；所述总线可以是外设部件互连标准(peripheral component interconnect，PCI)总线或扩展工业标准结构(extended industry standard architecture，EISA)总线等。所述总线可以分为地址总线、数据总线、控制总线等。为便于表示，图9中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0082] 所述驾驶决策装置900在实现驾驶决策方法时：

[0083] 所述存储器，存储有计算机程序或指令；

[0084] 所述处理器，用于调用所述存储器中存储的计算机程序或指令，执行下述方法：基于当前驾驶环境状态构建蒙特卡洛树，其中所述蒙特卡洛树包含N个节点，每个节点表示一个驾驶环境状态，所述N个节点包括根节点和N-1个非根节点，其中所述根节点表示所述当前驾驶环境状态，第一节点表示的驾驶环境状态，通过驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及第一驾驶动作预测得到，其中，所述第一驾驶动作为所述第一节点的父节点在扩展所述第一节点的过程中确定的驾驶动作，所述第一节点为所述N-1个非根节点中的任意一个节点，所述N为大于或等于2的正整数；根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用于驾驶决策；

[0085] 其中，所述每个节点的访问次数，根据所述节点的子节点的访问次数以及所述节点的初始访问次数确定，所述每个节点的值函数，根据所述节点的子节点的值函数和所述节点的初始值函数确定，其中，所述每个节点的初始访问次数为1、初始值函数根据与所述节点表示的驾驶环境状态匹配的值函数确定。

[0086] 在一种可能的设计中，通过所述驾驶环境随机模型基于所述第一节点的父节点表示的驾驶环境状态，以及所述第一驾驶动作，预测得到所述第一节点表示的驾驶环境状态，包括：通过所述驾驶环境随机模型采用Dropout前向传播，预测基于所述第一节点的父节点表示的驾驶环境状态，执行所述第一驾驶动作后的驾驶环境状态的概率分布；从所述概率分布采样得到所述第一节点表示的驾驶环境状态。

[0087] 在一种可能的设计中，根据与所述节点表示的驾驶环境状态匹配的值函数确定所述节点的初始值函数，包括：从情节记忆存储库中选取与所述节点表示的驾驶环境状态匹配度最高的第一数量的目标驾驶环境状态；根据所述第一数量的目标驾驶环境状态分别对应的值函数，确定所述节点的初始值函数。

[0088] 在一种可能的设计中，所述方法还包括：执行所述驾驶动作序列中的第一个驾驶动作后，获取执行所述第一个驾驶动作后的真实驾驶环境状态；根据所述当前驾驶环境状态、所述第一个驾驶动作以及执行所述第一个驾驶动作后的真实驾驶环境状态，对所述驾驶环境随机模型进行更新。

[0089] 在一种可能的设计中，所述根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列，包括：根据所述蒙特卡洛树中每个节点的访问次数，按照访问次数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的值函数，按照值函数最大原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列；或，根据所述蒙特卡洛树中每个节点的访问次数和值函数，按照访问次数最大优先、值函数最大次之原则，在所述蒙特卡洛树中确定一个从所述根节点开始至叶子节点结束的节点序列。

[0090] 在一种可能的设计中，所述方法还包括：当驾驶情节结束后，根据驾驶结果，确定所述驾驶情节中每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值；将每一驾驶动作执行后的真实驾驶环境状态对应的累积奖赏回报值作为其对应的值函数，更新所述情节记忆存储库。

[0091] 作为本实施例的另一种形式，提供一种计算机可读存储介质，其上存储有程序或指令，该程序或指令被执行时可以执行上述方法实施例中的驾驶决策方法。

[0092] 作为本实施例的另一种形式，提供一种包含指令的计算机程序产品，该指令被执行时可以执行上述方法实施例中的驾驶决策方法。

[0093] 作为本实施例的另一种形式，提供一种芯片，所述芯片，可以实现上述方法实施例中的驾驶决策方法。

[0094] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0095] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0096] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0097] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0098] 尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0099] 显然，本领域的技术人员可以对本申请实施例进行各种改动和变型而不脱离本申请实施例的精神和范围。这样，倘若本申请实施例的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

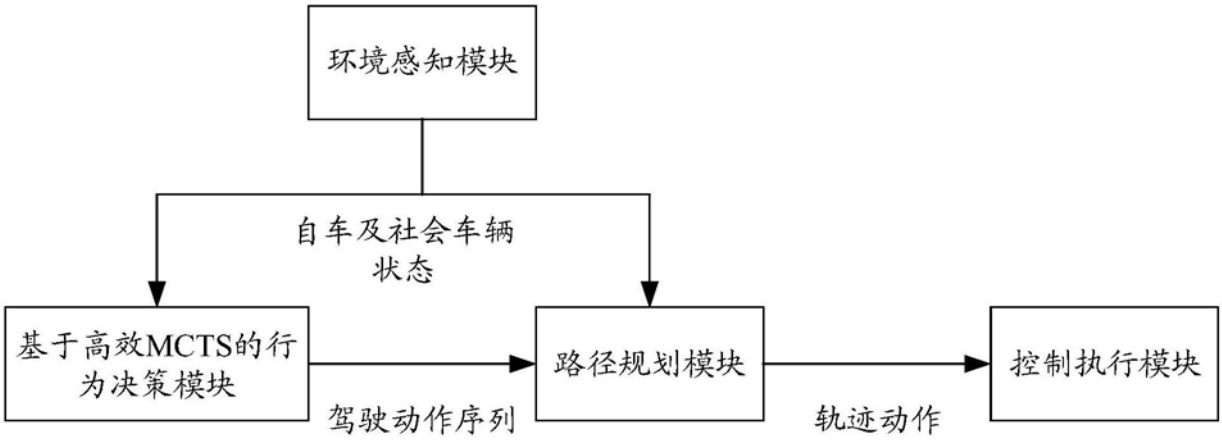

环境感知模块
自车及社会车辆状态
基于高效MCTS的行为决策模块
驾驶动作序列
路径规划模块
轨迹动作
控制执行模块

图1

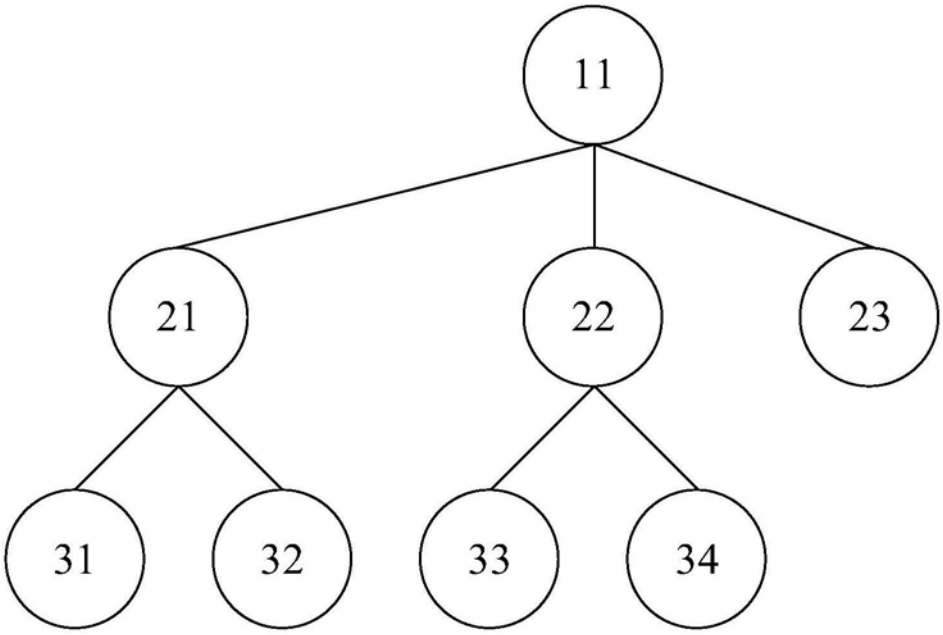

11
21
22
23
31
32
33
34

图2

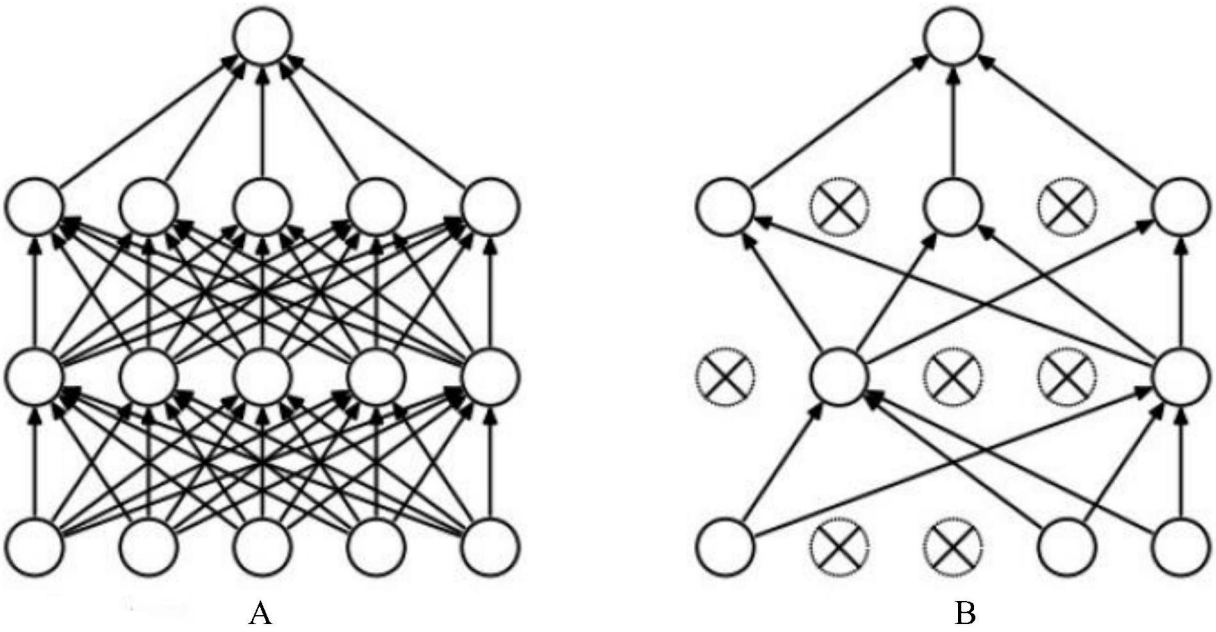

A
B

图3

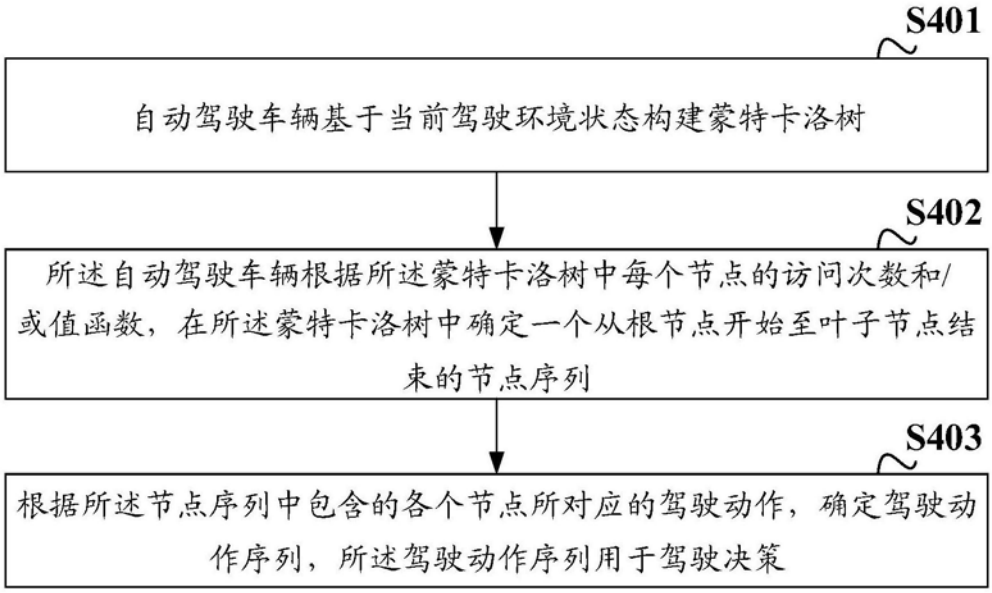

S401
自动驾驶车辆基于当前驾驶环境状态构建蒙特卡洛树
S402
所述自动驾驶车辆根据所述蒙特卡洛树中每个节点的访问次数和/或值函数，在所述蒙特卡洛树中确定一个从根节点开始至叶子节点结束的节点序列
S403
根据所述节点序列中包含的各个节点所对应的驾驶动作，确定驾驶动作序列，所述驾驶动作序列用于驾驶决策

图4

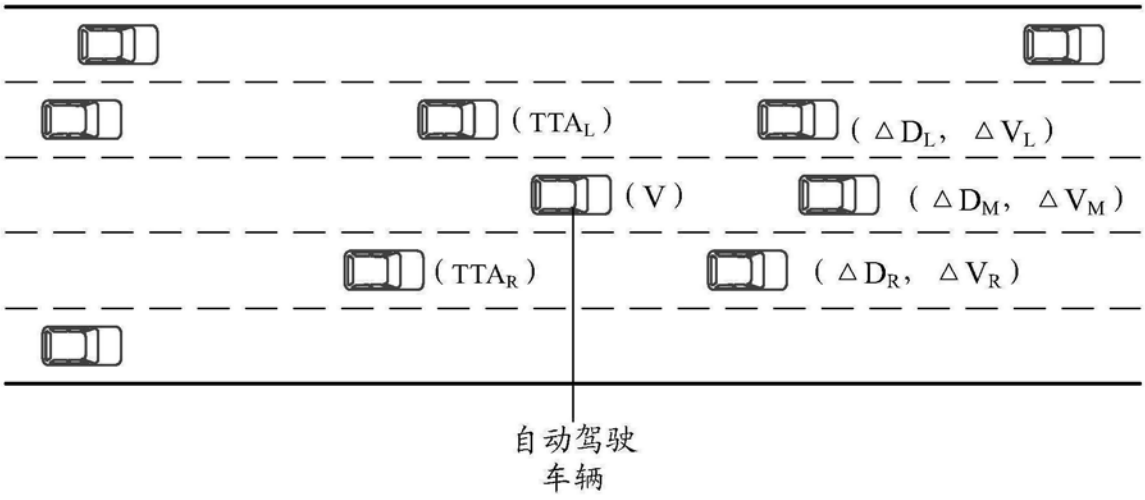

(TTA_L)
(△D_L, △V_L)
(V)
(△D_M, △V_M)
(TTA_R)
(△D_R, △V_R)
自动驾驶
车辆

图5

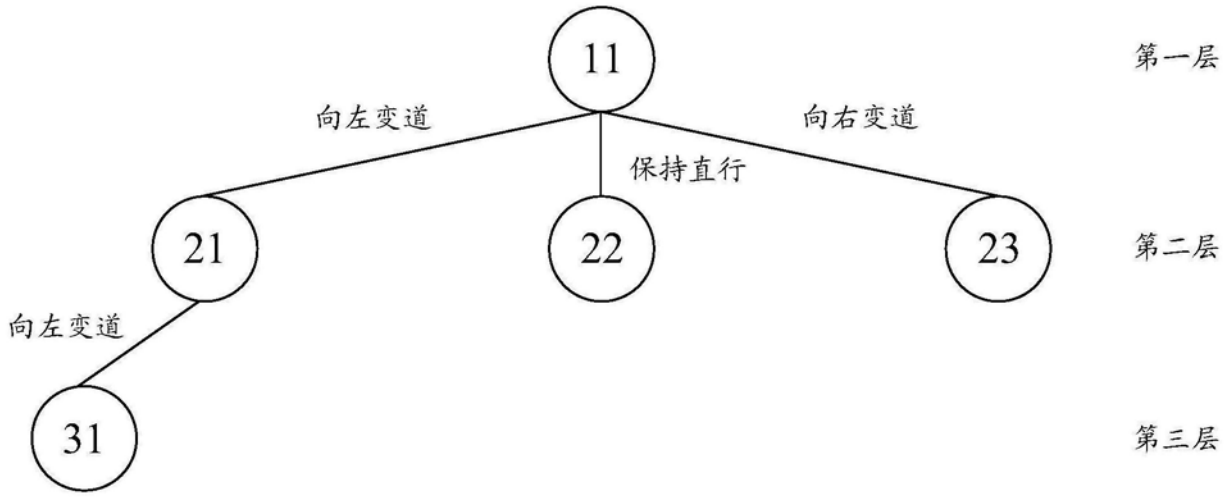

11
第一层
向左变道
向右变道
保持直行
21
22
23
第二层
向左变道
31
第三层

图6

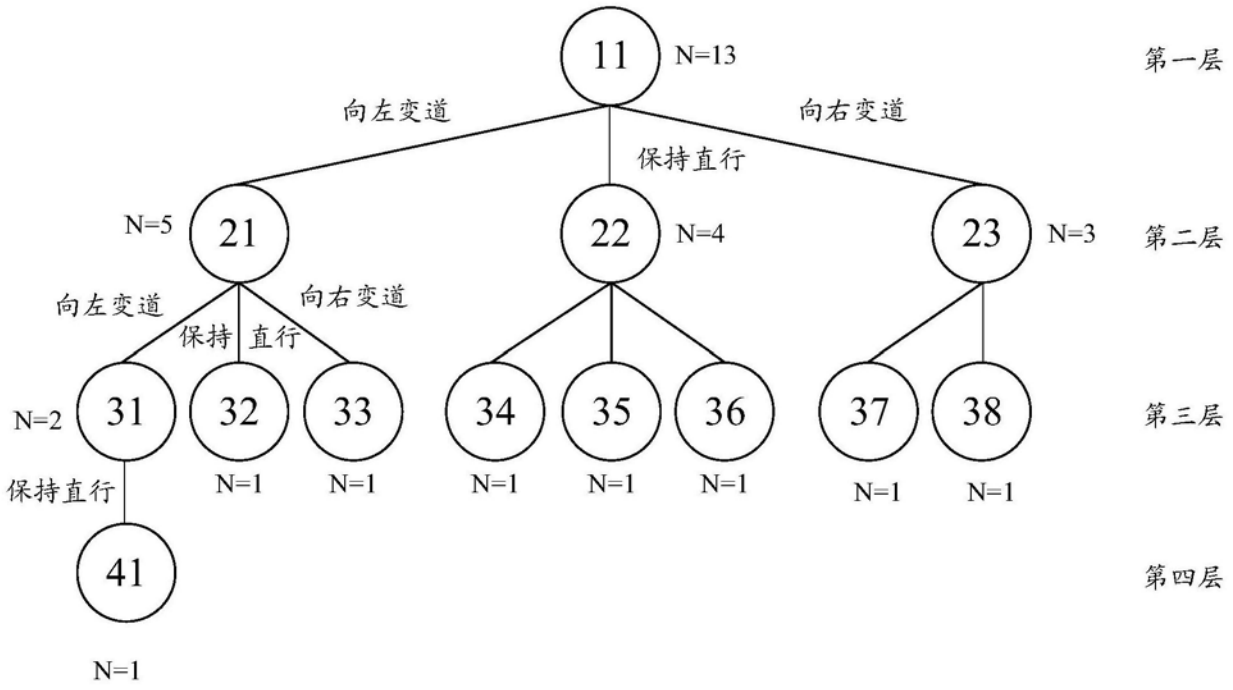

11
N=13
第一层
向左变道
保持直行
向右变道
N=5
21
22
N=4
23
N=3
第二层
向左变道
保持
直行
向右变道
N=2
31
32
33
34
35
36
37
38
第三层
保持直行
N=1
N=1
N=1
N=1
N=1
N=1
N=1
N=1
41
第四层
N=1

图7

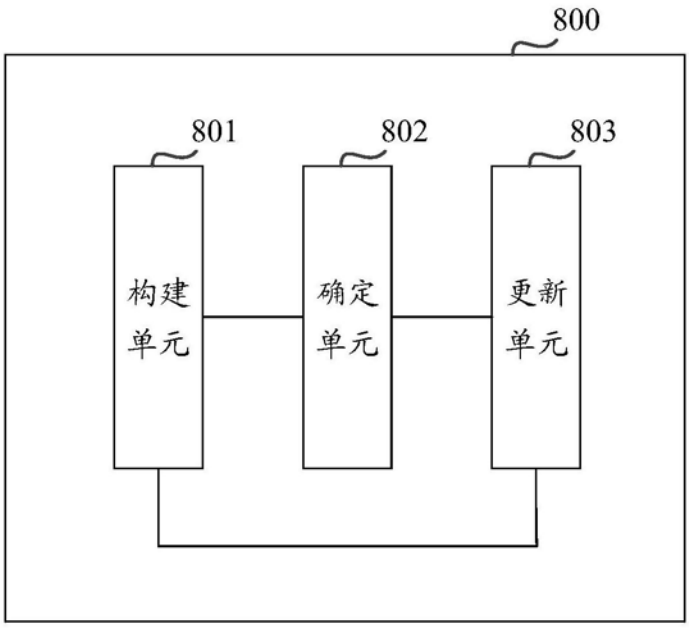

800
801
802
803
构建
单元
确定
单元
更新
单元

图8

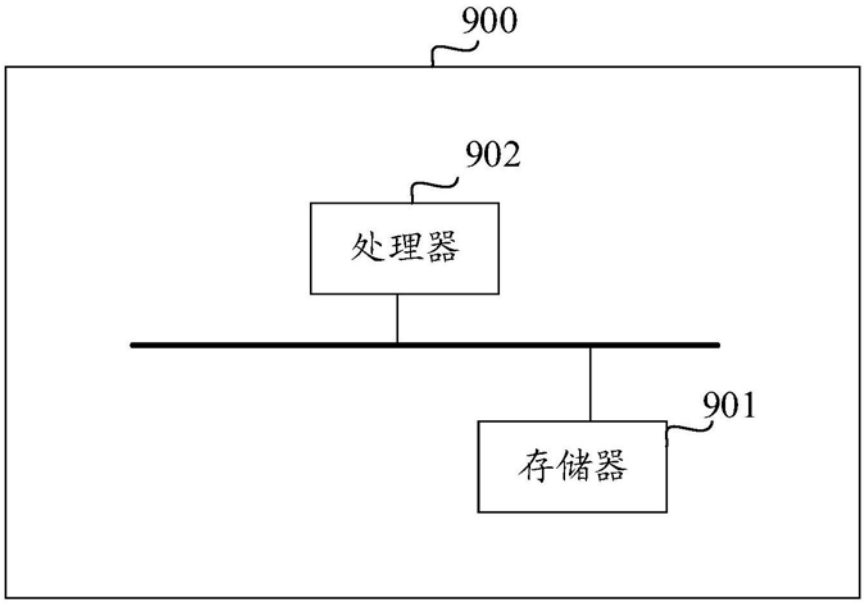


图9