(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117360552 B
(45)授权公告日 2024.03.26

(21)申请号 202311662491.1
(22)申请日 2023.12.06
(65)同一申请的已公布的文献号
申请公布号 CN 117360552 A
(43)申请公布日 2024.01.09
(73)专利权人 苏州元脑智能科技有限公司
地址 215100 江苏省苏州市吴中经济开发区郭巷街道官浦路1号9幢
(72)发明人 邓琪 李茹杨 张恒 张腾飞
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 周念念
(51)Int.Cl.
B60W 60/00(2020.01)
B60W 30/02(2012.01)

(56)对比文件
CN 116661299 A,2023.08.29
CN 116853243 A,2023.10.10
US 2020033869 A1,2020.01.30
US 2022196414 A1,2022.06.23
Changxi You ect..<Robotics and Autonomous Systems>.2019,第114卷1-18.
欧阳可可.《中国优秀硕士学位论文全文数据库工程科技Ⅱ辑》.2023,(第01期),1-55.

审查员 陈莹莹

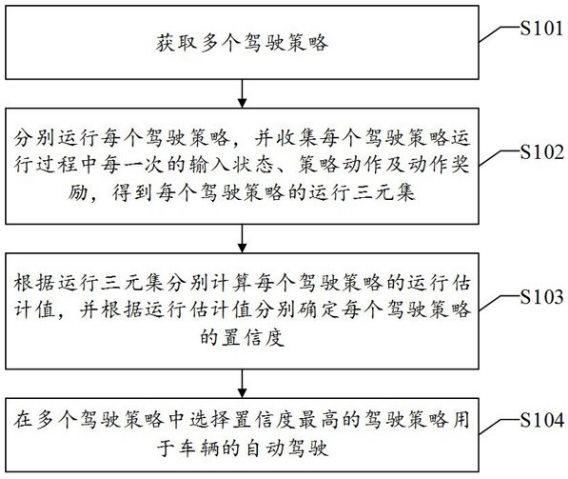
权利要求书4页 说明书24页 附图5页

(54)发明名称

一种车辆控制方法、装置、设备及可读存储介质

(57)摘要

本发明公开了自动驾驶技术领域内的一种车辆控制方法、装置、设备及可读存储介质。本发明能够分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,该策略动作用于控制车辆沿设定轨迹点行驶预设距离,可应对更复杂的驾驶场景;还能够根据每个驾驶策略的运行估计值分别确定每个驾驶策略的置信度,选择置信度最高的驾驶策略用于车辆的自动驾驶,由此可选择可靠性高的、适用于更高复杂度的驾驶场景的驾驶策略进行车辆的自动驾驶。该方案基于驾驶策略的置信度衡量驾驶策略的风险程度,通过风险程度最小的驾驶策略可确保车辆驾驶期间始终执行最优驾驶策略,保障长尾情况下的驾驶性能的稳定性。



CN 117360552 B

1.一种车辆控制方法,其特征在于,包括:

获取多个驾驶策略;

分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集;所述策略动作用于控制车辆沿设定轨迹点行驶预设距离;

根据所述运行三元集分别计算每个驾驶策略的运行估计值,并根据所述运行估计值分别确定每个驾驶策略的置信度;

在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶;

其中,所述分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集,包括:

针对每一驾驶策略,利用当前驾驶策略控制真实车辆进行自动驾驶,并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作;在自动驾驶结束后,汇总各次的输入状态及策略动作,得到训练样本;

利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型;

利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

2.根据权利要求1所述的方法,其特征在于,所述分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集,包括:

针对每一驾驶策略,利用当前驾驶策略控制真实车辆进行自动驾驶,并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态、策略动作及动作奖励;

在自动驾驶结束后,汇总各次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

3.根据权利要求1所述的方法,其特征在于,所述分别运行每个驾驶策略,包括:

分别利用每个驾驶策略控制同一真实车辆进行自动驾驶,以在同一真实车辆上分别运行每个驾驶策略。

4.根据权利要求1所述的方法,其特征在于,所述利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型,包括:

将所述训练样本和当前驾驶策略训练预设的高斯神经网络模型,得到所述虚拟驾驶模型。

5.根据权利要求1所述的方法,其特征在于,所述利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型,包括:

将所述训练样本划分为至少两个子样本集;

利用每个子样本集和当前驾驶策略分别训练一个子模型,得到至少两个子模型;

在所述至少两个子模型中选择模型评估值最低的子模型作为所述虚拟驾驶模型。

6.根据权利要求5所述的方法,其特征在于,所述在所述至少两个子模型中选择模型评估值最低的子模型作为所述虚拟驾驶模型,包括:

计算每个子模型在所述训练样本上的模型评估值;

选择模型评估值最低的子模型作为所述虚拟驾驶模型。

7.根据权利要求1所述的方法,其特征在于,所述利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集,包括:

若当前迭代次数未超出预测总次数,则获取前一次输入状态及前一次策略动作;将前一次输入状态及前一次策略动作输入所述虚拟驾驶模型,以使所述虚拟驾驶模型输出当前输入状态;

使当前驾驶策略根据当前输入状态输出当前策略动作;

使当前驾驶策略对应的奖励函数根据当前策略动作计算当前动作奖励;

将当前输入状态、当前策略动作和当前动作奖励构建为三元组,并将所述三元组作为当前驾驶策略的运行三元集中的一个元素;

将当前输入状态作为前一次输入状态,将当前策略动作作为前一次策略动作,并使当前迭代次数递增一,然后判断当前迭代次数是否超出预测总次数。

8.根据权利要求7所述的方法,其特征在于,所述奖励函数为:$r=\lambda_e\times r_e+\lambda_s\times r_s+\lambda_{ot}\times r_{ot}$;$r$为当前动作奖励,$\lambda_e$为当前驾驶策略的第一奖励系数,$\lambda_s$为当前驾驶策略的第二奖励系数,$\lambda_{ot}$为当前驾驶策略的第三奖励系数,$r_e$为当前车辆效率,$r_s$为当前安全奖励,$r_{ot}$为当前超车奖励。

9.根据权利要求1所述的方法,其特征在于,所述多个驾驶策略中的任意驾驶策略i的第一奖励系数、第二奖励系数和第三奖励系数的计算公式包括:

$\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$;

$\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$;

$\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$;

其中,$\lambda_{e,i}$为驾驶策略i的第一奖励系数,$\lambda_{s,i}$为驾驶策略i的第二奖励系数,$\lambda_{ot,i}$为驾驶策略i的第三奖励系数,$\lambda_{e,max}$为第一奖励系数对应的预设最大值,$\lambda_{e,min}$为第一奖励系数对应的预设最小值,$\lambda_{s,max}$为第二奖励系数对应的预设最大值,$\lambda_{s,min}$为第二奖励系数对应的预设最小值,$\lambda_{ot,max}$为第三奖励系数对应的预设最大值,$\lambda_{ot,min}$为第三奖励系数对应的预设最小值,m为驾驶策略的总个数。

10.根据权利要求1至9任一项所述的方法,其特征在于,所述多个驾驶策略中的任意目标驾驶策略的生成过程包括:

设定奖励函数,并构建包括所述奖励函数的初始策略;

利用强化学习方法训练所述初始策略,得到待优化策略;

利用所述待优化策略构建优化样本;

在成本函数的约束下,以最大奖励为求解目标,构建拉格朗日目标函数;

利用所述优化样本迭代求解所述拉格朗日目标函数,以优化所述待优化策略,得到所述目标驾驶策略。

11.根据权利要求10所述的方法,其特征在于,所述利用所述待优化策略构建优化样本,包括:

将目标状态输入所述待优化策略,以使所述待优化策略输出结束状态和目标窗口;

在所述目标窗口内使所述目标状态为起始点,使所述结束状态为终点,并通过曲线拟合确定所述目标窗口内的各轨迹点;

连接各轨迹点得到运动轨迹,并生成能够控制车辆沿所述运动轨迹行驶的目标策略动

作；

将所述目标状态、所述目标策略动作和所述目标策略动作的奖励值构建为所述优化样本。

12.根据权利要求11所述的方法，其特征在于，所述通过曲线拟合确定所述目标窗口内的各轨迹点，包括：

在所述目标窗口内拟合得到位移变化曲线；

在所述目标窗口内拟合得到速度变化曲线；

匹配所述位移变化曲线和所述速度变化曲线中的各点，以确定所述目标窗口内的各轨迹点。

13.根据权利要求11所述的方法，其特征在于，所述通过曲线拟合确定所述目标窗口内的各轨迹点，包括：

在所述目标窗口内拟合得到位移变化曲线；

在所述目标窗口内拟合速度变化曲线时与所述位移变化曲线进行匹配，以确定所述目标窗口内的各轨迹点。

14.根据权利要求11所述的方法，其特征在于，所述通过曲线拟合确定所述目标窗口内的各轨迹点，包括：

在所述目标窗口内拟合得到速度变化曲线；

在所述目标窗口内拟合位移变化曲线时与所述速度变化曲线进行匹配，以确定所述目标窗口内的各轨迹点。

15.根据权利要求10所述的方法，其特征在于，所述设定奖励函数，包括：

确定所述目标驾驶策略在所述多个驾驶策略中的标识信息；

根据所述标识信息计算第一奖励系数、第二奖励系数和第三奖励系数，并构建所述奖励函数。

16.根据权利要求1至9任一项所述的方法，其特征在于，所述在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶，包括：

使置信度最高的驾驶策略针对车辆当前状态输出可信策略动作；

按照所述可信策略动作确定由多个控制指令构成的指令序列；

按照所述指令序列控制车辆沿设定轨迹点自动行驶预设距离。

17.一种车辆控制装置，其特征在于，包括：

获取模块，用于获取多个驾驶策略；

收集模块，用于分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集；所述策略动作用于控制车辆沿设定轨迹点行驶预设距离；

评估模块，用于根据所述运行三元集分别计算每个驾驶策略的运行估计值，并根据所述运行估计值分别确定每个驾驶策略的置信度；

应用模块，用于在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶；

其中，所述收集模块包括：

样本准备单元，用于针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾

驶,并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作;在自动驾驶结束后,汇总各次的输入状态及策略动作,得到训练样本;

训练单元,用于利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型;

生成单元,用于利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

18.一种电子设备,其特征在于,包括:

存储器,用于存储计算机程序;

处理器,用于执行所述计算机程序,以实现如权利要求1至16任一项所述的方法。

19.一种可读存储介质,其特征在于,用于保存计算机程序,其中,所述计算机程序被处理器执行时实现如权利要求1至16任一项所述的方法。

一种车辆控制方法、装置、设备及可读存储介质

技术领域

[0001] 本发明涉及自动驾驶技术领域，特别涉及一种车辆控制方法、装置、设备及可读存储介质。

背景技术

[0002] 自动驾驶技术在提高各种驾驶场景下的车辆安全性和机动性方面具有巨大潜力。然而，现实世界的驾驶场景通常是长尾分布式的，对于出现概率较小的风险案例，驾驶系统会由于数据不足而缺乏对环境的了解，无法及时作出合理响应。自动驾驶车辆在现实中可能遇到的风险案例无穷无尽，这些案例可能具有多种特征，例如封路、交通事故、违反交通规则等，即使进行数百万英里的实际路测也无法一一遍历。即便是对于一个训练有素的驾驶策略，在实际驾驶过程中仍然可能会出现故障。

[0003] 由于真实自动驾驶过程并不是特定个别场景的简单切换，驾驶策略可能会被要求同时处理多种未见场景，这对驾驶策略提出了更高的要求。当前通过强化学习得到的自动驾驶策略要么过于激进要么过于保守，导致自动驾驶策略实际上难以产生可靠的自动驾驶动作。并且，当前自动驾驶策略用于产生车辆级别的控制命令，如：每个时刻的车辆转向、加速指令等，这种单步控制的自动驾驶策略难以实现复杂度更高的高级驾驶行为。

[0004] 因此，如何选择可靠性高的、适用于更高复杂度驾驶场景的自动驾驶策略，是本领域技术人员需要解决的问题。

发明内容

[0005] 有鉴于此，本发明的目的在于提供一种车辆控制方法、装置、设备及可读存储介质，以选择可靠性高的、适用于更高复杂度驾驶场景的自动驾驶策略。其具体方案如下：

[0006] 第一方面，本发明提供了一种车辆控制方法，包括：

[0007] 获取多个驾驶策略；

[0008] 分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集；所述策略动作用于控制车辆沿设定轨迹点行驶预设距离；

[0009] 根据所述运行三元集分别计算每个驾驶策略的运行估计值，并根据所述运行估计值分别确定每个驾驶策略的置信度；

[0010] 在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶。

[0011] 可选地，所述分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集，包括：

[0012] 针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾驶，并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态、策略动作及动作奖励；

[0013] 在自动驾驶结束后，汇总各次的输入状态、策略动作及动作奖励，得到当前驾驶策略的运行三元集。

[0014] 可选地,所述分别运行每个驾驶策略,包括:

[0015] 分别利用每个驾驶策略控制同一真实车辆进行自动驾驶,以在同一真实车辆上分别运行每个驾驶策略。

[0016] 可选地,所述分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集,包括:

[0017] 针对每一驾驶策略,利用当前驾驶策略控制真实车辆进行自动驾驶,并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作;在自动驾驶结束后,汇总各次的输入状态及策略动作,得到训练样本;

[0018] 利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型;

[0019] 利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

[0020] 可选地,所述利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型,包括:

[0021] 将所述训练样本和当前驾驶策略训练预设的高斯神经网络模型,得到所述虚拟驾驶模型。

[0022] 可选地,所述利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型,包括:

[0023] 将所述训练样本划分为至少两个子样本集;

[0024] 利用每个子样本集和当前驾驶策略分别训练一个子模型,得到至少两个子模型;

[0025] 在所述至少两个子模型中选择模型评估值最低的子模型作为所述虚拟驾驶模型。

[0026] 可选地,所述在所述至少两个子模型中选择模型评估值最低的子模型作为所述虚拟驾驶模型,包括:

[0027] 计算每个子模型在所述训练样本上的模型评估值;

[0028] 选择模型评估值最低的子模型作为所述虚拟驾驶模型。

[0029] 可选地,所述利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集,包括:

[0030] 若当前迭代次数未超出预测总次数,则获取前一次输入状态及前一次策略动作;将前一次输入状态及前一次策略动作输入所述虚拟驾驶模型,以使所述虚拟驾驶模型输出当前输入状态;

[0031] 使当前驾驶策略根据当前输入状态输出当前策略动作;

[0032] 使当前驾驶策略对应的奖励函数根据当前策略动作计算当前动作奖励;

[0033] 将当前输入状态、当前策略动作和当前动作奖励构建为三元组,并将所述三元组作为当前驾驶策略的运行三元集中的一个元素;

[0034] 将当前输入状态作为前一次输入状态,将当前策略动作作为前一次策略动作,并使当前迭代次数递增一,然后判断当前迭代次数是否超出预测总次数。

[0035] 可选地,所述奖励函数为:$r=\lambda_e\times r_e+\lambda_s\times r_s+\lambda_{ot}\times r_{ot}$;$r$为当前动作奖励,$\lambda_e$为当前驾驶策略的第一奖励系数,$\lambda_s$为当前驾驶策略的第二奖励系数,$\lambda_{ot}$为当前驾驶策略的第三奖励系数,$r_e$为当前车辆效率,$r_s$为当前安全奖励,$r_{ot}$为当前超车奖励。

[0036] 可选地,所述多个驾驶策略中的任意驾驶策略i的第一奖励系数、第二奖励系数和第三奖励系数的计算公式包括:

[0037] $\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$;

[0038] $\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$;

[0039] $\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$;

[0040] 其中,$\lambda_{e,i}$为驾驶策略i的第一奖励系数,$\lambda_{s,i}$为驾驶策略i的第二奖励系数,$\lambda_{ot,i}$为驾驶策略i的第三奖励系数,$\lambda_{e,max}$为第一奖励系数对应的预设最大值,$\lambda_{e,min}$为第一奖励系数对应的预设最小值,$\lambda_{s,max}$为第二奖励系数对应的预设最大值,$\lambda_{s,min}$为第二奖励系数对应的预设最小值,$\lambda_{ot,max}$为第三奖励系数对应的预设最大值,$\lambda_{ot,min}$为第三奖励系数对应的预设最小值,m为驾驶策略的总个数。

[0041] 可选地,所述多个驾驶策略中的任意目标驾驶策略的生成过程包括:

[0042] 设定奖励函数,并构建包括所述奖励函数的初始策略;

[0043] 利用强化学习方法训练所述初始策略,得到待优化策略;

[0044] 利用所述待优化策略构建优化样本;

[0045] 在成本函数的约束下,以最大奖励为求解目标,构建拉格朗日目标函数;

[0046] 利用所述优化样本迭代求解所述拉格朗日目标函数,以优化所述待优化策略,得到所述目标驾驶策略。

[0047] 可选地,所述利用所述待优化策略构建优化样本,包括:

[0048] 将目标状态输入所述待优化策略,以使所述待优化策略输出结束状态和目标窗口;

[0049] 在所述目标窗口内使所述目标状态为起始点,使所述结束状态为终点,并通过曲线拟合确定所述目标窗口内的各轨迹点;

[0050] 连接各轨迹点得到运动轨迹,并生成能够控制车辆沿所述运动轨迹行驶的目标策略动作;

[0051] 将所述目标状态、所述目标策略动作和所述目标策略动作的奖励值构建为所述优化样本。

[0052] 可选地,所述通过曲线拟合确定所述目标窗口内的各轨迹点,包括:

[0053] 在所述目标窗口内拟合得到位移变化曲线;

[0054] 在所述目标窗口内拟合得到速度变化曲线;

[0055] 匹配所述位移变化曲线和所述速度变化曲线中的各点,以确定所述目标窗口内的各轨迹点。

[0056] 可选地,所述通过曲线拟合确定所述目标窗口内的各轨迹点,包括:

[0057] 在所述目标窗口内拟合得到位移变化曲线;

[0058] 在所述目标窗口内拟合速度变化曲线时与所述位移变化曲线进行匹配,以确定所述目标窗口内的各轨迹点。

[0059] 可选地,所述通过曲线拟合确定所述目标窗口内的各轨迹点,包括:

[0060] 在所述目标窗口内拟合得到速度变化曲线;

[0061] 在所述目标窗口内拟合位移变化曲线时与所述速度变化曲线进行匹配,以确定所述目标窗口内的各轨迹点。

[0062] 可选地,所述拉格朗日目标函数含用于约束所述成本函数的正则化项。

[0063] 可选地,在任意一次迭代中执行如下优化步骤:

[0064] 固定拉格朗日乘子和成本函数,通过最大化所述拉格朗日目标函数优化优化当前

待优化策略；

[0065] 和/或

[0066] 固定当前待优化策略和成本函数，通过最小化所述拉格朗日目标函数优化拉格朗日乘子；

[0067] 和/或

[0068] 固定当前待优化策略和拉格朗日乘子，通过最大化所述拉格朗日目标函数优化成本函数。

[0069] 可选地，所述设定奖励函数，包括：

[0070] 确定所述目标驾驶策略在所述多个驾驶策略中的标识信息；

[0071] 根据所述标识信息计算第一奖励系数、第二奖励系数和第三奖励系数，并构建所述奖励函数。

[0072] 可选地，所述在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶，包括：

[0073] 使置信度最高的驾驶策略针对车辆当前状态输出可信策略动作；

[0074] 按照所述可信策略动作确定由多个控制指令构成的指令序列；

[0075] 按照所述指令序列控制车辆沿设定轨迹点自动行驶预设距离。

[0076] 第二方面，本发明提供了一种车辆控制装置，包括：

[0077] 获取模块，用于获取多个驾驶策略；

[0078] 收集模块，用于分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集；所述策略动作用于控制车辆沿设定轨迹点行驶预设距离；

[0079] 评估模块，用于根据所述运行三元集分别计算每个驾驶策略的运行估计值，并根据所述运行估计值分别确定每个驾驶策略的置信度；

[0080] 应用模块，用于在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶。

[0081] 可选地，所述收集模块具体用于：针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾驶，并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态、策略动作及动作奖励；在自动驾驶结束后，汇总各次的输入状态、策略动作及动作奖励，得到当前驾驶策略的运行三元集。

[0082] 可选地，所述收集模块具体用于：分别利用每个驾驶策略控制同一真实车辆进行自动驾驶，以在同一真实车辆上分别运行每个驾驶策略。

[0083] 可选地，所述收集模块包括：

[0084] 样本准备单元，用于针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾驶，并收集所述真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作；在自动驾驶结束后，汇总各次的输入状态及策略动作，得到训练样本；

[0085] 训练单元，用于利用所述训练样本和当前驾驶策略训练得到虚拟驾驶模型；

[0086] 生成单元，用于利用当前驾驶策略和所述虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励，得到当前驾驶策略的运行三元集。

[0087] 可选地，所述训练单元具体用于：将所述训练样本和当前驾驶策略训练预设的高

斯神经网络模型,得到所述虚拟驾驶模型。

[0088] 可选地,所述训练单元具体用于:将所述训练样本划分为至少两个子样本集;利用每个子样本集和当前驾驶策略分别训练一个子模型,得到至少两个子模型;在所述至少两个子模型中选择模型评估值最低的子模型作为所述虚拟驾驶模型。

[0089] 可选地,所述训练单元具体用于:计算每个子模型在所述训练样本上的模型评估值;选择模型评估值最低的子模型作为所述虚拟驾驶模型。

[0090] 可选地,所述生成单元具体用于:若当前迭代次数未超出预测总次数,则获取前一次输入状态及前一次策略动作;将前一次输入状态及前一次策略动作输入所述虚拟驾驶模型,以使所述虚拟驾驶模型输出当前输入状态;使当前驾驶策略根据当前输入状态输出当前策略动作;使当前驾驶策略对应的奖励函数根据当前策略动作计算当前动作奖励;将当前输入状态、当前策略动作和当前动作奖励构建为三元组,并将所述三元组作为当前驾驶策略的运行三元集中的一个元素;将当前输入状态作为前一次输入状态,将当前策略动作作为前一次策略动作,并使当前迭代次数递增一,然后判断当前迭代次数是否超出预测总次数。

[0091] 可选地,所述奖励函数为:$r=\lambda_e\times r_e+\lambda_s\times r_s+\lambda_{ot}\times r_{ot}$;$r$为当前动作奖励,$\lambda_e$为当前驾驶策略的第一奖励系数,$\lambda_s$为当前驾驶策略的第二奖励系数,$\lambda_{ot}$为当前驾驶策略的第三奖励系数,$r_e$为当前车辆效率,$r_s$为当前安全奖励,$r_{ot}$为当前超车奖励。

[0092] 可选地,所述多个驾驶策略中的任意驾驶策略i的第一奖励系数、第二奖励系数和第三奖励系数的计算公式包括:$\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$;$\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$;$\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$;其中,$\lambda_{e,i}$为驾驶策略$i$的第一奖励系数,$\lambda_{s,i}$为驾驶策略$i$的第二奖励系数,$\lambda_{ot,i}$为驾驶策略$i$的第三奖励系数,$\lambda_{e,max}$为第一奖励系数对应的预设最大值,$\lambda_{e,min}$为第一奖励系数对应的预设最小值,$\lambda_{s,max}$为第二奖励系数对应的预设最大值,$\lambda_{s,min}$为第二奖励系数对应的预设最小值,$\lambda_{ot,max}$为第三奖励系数对应的预设最大值,$\lambda_{ot,min}$为第三奖励系数对应的预设最小值,$m$为驾驶策略的总个数。

[0093] 可选地,还包括:驾驶策略生成模块,该模块用于生成所述多个驾驶策略中的任意目标驾驶策略;其中,驾驶策略生成模块包括:

[0094] 初始单元,用于设定奖励函数,并构建包括所述奖励函数的初始策略;

[0095] 训练单元,用于利用强化学习方法训练所述初始策略,得到待优化策略;

[0096] 样本构建单元,用于利用所述待优化策略构建优化样本;

[0097] 函数构建单元,用于在成本函数的约束下,以最大奖励为求解目标,构建拉格朗日目标函数;

[0098] 优化单元,用于利用所述优化样本迭代求解所述拉格朗日目标函数,以优化所述待优化策略,得到所述目标驾驶策略。

[0099] 可选地,所述样本构建单元具体用于:将目标状态输入所述待优化策略,以使所述待优化策略输出结束状态和目标窗口;在所述目标窗口内使所述目标状态为起始点,使所述结束状态为终点,并通过曲线拟合确定所述目标窗口内的各轨迹点;连接各轨迹点得到运动轨迹,并生成能够控制车辆沿所述运动轨迹行驶的目标策略动作;将所述目标状态、所述目标策略动作和所述目标策略动作的奖励值构建为所述优化样本。

[0100] 可选地,所述样本构建单元具体用于:在所述目标窗口内拟合得到位移变化曲线;

在所述目标窗口内拟合得到速度变化曲线;匹配所述位移变化曲线和所述速度变化曲线中的各点,以确定所述目标窗口内的各轨迹点。

[0101] 可选地,所述样本构建单元具体用于:在所述目标窗口内拟合得到位移变化曲线;在所述目标窗口内拟合速度变化曲线时与所述位移变化曲线进行匹配,以确定所述目标窗口内的各轨迹点。

[0102] 可选地,所述样本构建单元具体用于:在所述目标窗口内拟合得到速度变化曲线;在所述目标窗口内拟合位移变化曲线时与所述速度变化曲线进行匹配,以确定所述目标窗口内的各轨迹点。

[0103] 可选地,所述拉格朗日目标函数含用于约束所述成本函数的正则化项。

[0104] 可选地,在任意一次迭代中执行如下优化步骤:固定拉格朗日乘子和成本函数,通过最大化所述拉格朗日目标函数优化优化当前待优化策略;和/或固定当前待优化策略和成本函数,通过最小化所述拉格朗日目标函数优化拉格朗日乘子;和/或固定当前待优化策略和拉格朗日乘子,通过最大化所述拉格朗日目标函数优化成本函数。

[0105] 可选地,所述初始单元具体用于:确定所述目标驾驶策略在所述多个驾驶策略中的标识信息;根据所述标识信息计算第一奖励系数、第二奖励系数和第三奖励系数,并构建所述奖励函数。

[0106] 可选地,所述应用模块具体用于:使置信度最高的驾驶策略针对车辆当前状态输出可信策略动作;按照所述可信策略动作确定由多个控制指令构成的指令序列;按照所述指令序列控制车辆沿设定轨迹点自动行驶预设距离。

[0107] 第三方面,本发明提供了一种电子设备,包括:

[0108] 存储器,用于存储计算机程序;

[0109] 处理器,用于执行所述计算机程序,以实现前述公开的车辆控制方法。

[0110] 第四方面,本发明提供了一种可读存储介质,用于保存计算机程序,其中,所述计算机程序被处理器执行时实现前述公开的车辆控制方法。

[0111] 通过以上方案可知,本发明提供了一种车辆控制方法,包括:获取多个驾驶策略;分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集;所述策略动作用于控制车辆沿设定轨迹点行驶预设距离;根据所述运行三元集分别计算每个驾驶策略的运行估计值,并根据所述运行估计值分别确定每个驾驶策略的置信度;在所述多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶。

[0112] 可见,本发明的有益效果为:能够分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,该策略动作用于控制车辆沿设定轨迹点行驶预设距离,不同于车辆级别的控制命令,利用此策略动作可应对更复杂的驾驶场景;该方案还能够根据每个驾驶策略的运行估计值分别确定每个驾驶策略的置信度,然后选择置信度最高的驾驶策略用于车辆的自动驾驶,由此可选择可靠性高的、适用于更高复杂度的驾驶场景的驾驶策略进行车辆的自动驾驶。该方案基于驾驶策略的置信度衡量驾驶策略的风险程度,通过置信度最高的(即风险程度最小的)驾驶策略可确保车辆驾驶期间始终执行最优驾驶策略,保障长尾情况下的驾驶性能的稳定性。

[0113] 相应地,本发明提供的一种车辆控制装置、设备及可读存储介质,也同样具有上述

技术效果。

附图说明

[0114] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0115] 图1为本发明公开的一种车辆控制方法流程图;

[0116] 图2为本发明公开的一种驾驶策略生成方法流程图;

[0117] 图3为本发明公开的一种车辆控制装置示意图;

[0118] 图4为本发明公开的一种电子设备示意图;

[0119] 图5为本发明提供的一种服务器结构图;

[0120] 图6为本发明提供的一种终端结构图;

[0121] 图7为本发明公开的一种驾驶策略生成过程示意图;

[0122] 图8为本发明公开的一种驾驶策略选择过程示意图。

具体实施方式

[0123] 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。

[0124] 目前,驾驶策略可能会被要求同时处理多种未见场景,这对驾驶策略提出了更高的要求。当前通过强化学习得到的自动驾驶策略要么过于激进要么过于保守,导致自动驾驶策略实际上难以产生可靠的自动驾驶动作。并且,当前自动驾驶策略用于产生车辆级别的控制命令,如:每个时刻的车辆转向、加速指令等,这种单步控制的自动驾驶策略难以实现复杂度更高的高级驾驶行为。为此,本发明提供了一种车辆控制方案,能够选择可靠性高的、适用于更高复杂度驾驶场景的自动驾驶策略用于车辆的自动驾驶。

[0125] 参见图1所示,本发明实施例公开了一种车辆控制方法,包括:

[0126] S101、获取多个驾驶策略。

[0127] 在本实施例中,不同驾驶策略的驾驶偏好不同,如:有的驾驶策略比较激进,有的驾驶策略比较保守。驾驶策略的驾驶偏好可通过改变其所对应奖励函数的奖励系数实现。奖励函数的计算公式可以为:$r=\lambda_e\times r_e+\lambda_s\times r_s+\lambda_{ot}\times r_{ot}$;$r$为当前动作奖励,$\lambda_e$为当前驾驶策略的第一奖励系数,$\lambda_s$为当前驾驶策略的第二奖励系数,$\lambda_{ot}$为当前驾驶策略的第三奖励系数,$r_e$为当前车辆效率,$r_s$为当前安全奖励,$r_{ot}$为当前超车奖励。通过调整第一奖励系数、第二奖励系数和/或第三奖励系数的取值,可更改相应驾驶策略的驾驶偏好。需要说明的是,本实施例中的奖励函数针对策略动作计算其对应的动作奖励,动作奖励也就是:车辆在策略动作的控制下沿设定轨迹点行驶预设距离的累计奖励,因此本实施例中的奖励函数不同于计算单一车辆控制命令的奖励的函数。

[0128] 其中,第一奖励系数、第二奖励系数和第三奖励系数的取值可按照下述进行调整

和更改。在一种实施方式中，多个驾驶策略中的任意驾驶策略i的第一奖励系数、第二奖励系数和第三奖励系数的计算公式包括：$\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$；$\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$；$\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$；其中，$\lambda_{e,i}$为驾驶策略$i$的第一奖励系数，$\lambda_{s,i}$为驾驶策略$i$的第二奖励系数，$\lambda_{ot,i}$为驾驶策略$i$的第三奖励系数，$\lambda_{e,max}$为第一奖励系数对应的预设最大值，$\lambda_{e,min}$为第一奖励系数对应的预设最小值，$\lambda_{s,max}$为第二奖励系数对应的预设最大值，$\lambda_{s,min}$为第二奖励系数对应的预设最小值，$\lambda_{ot,max}$为第三奖励系数对应的预设最大值，$\lambda_{ot,min}$为第三奖励系数对应的预设最小值，$m$为驾驶策略的总个数。$i$可看作当前驾驶策略在多个驾驶策略中的标识信息，该标识信息可以用数字表示，例如：该标识信息可以是多个驾驶策略按激进程度或保守程序由大至小或由小至大排列确定的驾驶策略的排列序号。

[0129] S102、分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集。

[0130] 其中，策略动作用于控制车辆沿设定轨迹点行驶预设距离。通过策略动作可确定一系列有排列顺序的控制指令，按照这些控制指令可以有序控制车辆沿设定轨迹点行驶预设距离。其中，一个控制指令对应一个轨迹点。输入状态为输入给驾驶策略被其处理的车辆的状态，包括：车辆位置(包括纵向位置和横向位置)、航向角、速度和加速度等信息。动作奖励是驾驶策略针对输入状态所输出的策略动作对应的奖励值。

[0131] 需要说明的是，每个驾驶策略的运行三元集可以是在真实车辆上应用相应驾驶策略得到，也可以是在虚拟环境中运行相应驾驶策略得到。在一种实施方式中，分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集，包括：针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾驶，并收集真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态、策略动作及动作奖励；在自动驾驶结束后，汇总各次的输入状态、策略动作及动作奖励，得到当前驾驶策略的运行三元集。由此得到的运行三元集中的元素均为真实驾驶数据。

[0132] 在一种实施方式中，分别运行每个驾驶策略，包括：分别利用每个驾驶策略控制同一真实车辆进行自动驾驶，以在同一真实车辆上分别运行每个驾驶策略。由此使每个驾驶策略在同一真实车辆上得到真实驾驶数据构成的运行三元集。

[0133] 在一种实施方式中，分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集，包括：针对每一驾驶策略，利用当前驾驶策略控制真实车辆进行自动驾驶，并收集真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作；在自动驾驶结束后，汇总各次的输入状态及策略动作，得到训练样本；利用训练样本和当前驾驶策略训练得到虚拟驾驶模型；利用当前驾驶策略和虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励，得到当前驾驶策略的运行三元集。由此得到的运行三元集中的元素均为虚拟驾驶数据。

[0134] 需要说明的是，单一驾驶策略的运行三元集中的元素可以为虚拟驾驶数据和/或真实驾驶数据。

[0135] 在一种实施方式中，利用训练样本和当前驾驶策略训练得到虚拟驾驶模型，包括：将训练样本和当前驾驶策略训练预设的高斯神经网络模型，得到虚拟驾驶模型。在一种实施方式中，利用训练样本和当前驾驶策略训练得到虚拟驾驶模型，包括：将训练样本划分为

至少两个子样本集;利用每个子样本集和当前驾驶策略分别训练一个子模型,得到至少两个子模型;在至少两个子模型中选择模型评估值最低的子模型作为虚拟驾驶模型。其中,在至少两个子模型中选择模型评估值最低的子模型作为虚拟驾驶模型,包括:计算每个子模型在训练样本上的模型评估值;选择模型评估值最低的子模型作为虚拟驾驶模型。模型评估值用于描述模型的性能,模型评估值越大模型性能越好,模型评估值越小模型性能越差。本实施例为了评估出每一驾驶策略对应的最不安全情况,选择模型评估值最低的子模型作为虚拟驾驶模型,由此能够使驾驶策略的运行三元集含该驾驶策略在最不安全情况下产生的策略动作及动作奖励,之后据此评估每一驾驶策略的运行估计值,有利于选择出安全性最可靠的驾驶策略。虚拟驾驶模型通过有监督的训练方法训练获得。

[0136] 进一步地,利用当前驾驶策略和虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集,包括:若当前迭代次数未超出预测总次数,则获取前一次输入状态及前一次策略动作;将前一次输入状态及前一次策略动作输入虚拟驾驶模型,以使虚拟驾驶模型输出当前输入状态;使当前驾驶策略根据当前输入状态输出当前策略动作;使当前驾驶策略对应的奖励函数根据当前策略动作计算当前动作奖励;将当前输入状态、当前策略动作和当前动作奖励构建为三元组,并将三元组作为当前驾驶策略的运行三元集中的一个元素;将当前输入状态作为前一次输入状态,将当前策略动作作为前一次策略动作,并使当前迭代次数递增一,然后判断当前迭代次数是否超出预测总次数。若当前迭代次数未超出预测总次数,则停止当前流程,并输出当前驾驶策略的运行三元集。可见,虚拟驾驶模型的输入数据包括:前一次输入状态和前一次策略动作;输出数据为:后一次输入状态(当前输入状态)。

[0137] S103、根据运行三元集分别计算每个驾驶策略的运行估计值,并根据运行估计值分别确定每个驾驶策略的置信度。

[0138] 其中,每个驾驶策略的运行估计值可直接作为相应驾驶策略的置信度。

[0139] S104、在多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶。

[0140] 在本实施例中,在多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶,包括:使置信度最高的驾驶策略针对车辆当前状态输出可信策略动作;按照可信策略动作确定由多个控制指令构成的指令序列;按照指令序列控制车辆沿设定轨迹点自动行驶预设距离。

[0141] 可见,本实施例能够分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励,该策略动作用于控制车辆沿设定轨迹点行驶预设距离,不同于车辆级别的控制命令,利用此策略动作可应对更复杂的驾驶场景;该方案还能够根据每个驾驶策略的运行估计值分别确定每个驾驶策略的置信度,然后选择置信度最高的驾驶策略用于车辆的自动驾驶,由此可选择可靠性高的、适用于更高复杂度的驾驶场景的驾驶策略进行车辆的自动驾驶。该方案基于驾驶策略的置信度衡量驾驶策略的风险程度,通过置信度最高的(即风险程度最小的)驾驶策略可确保车辆驾驶期间始终执行最优驾驶策略,保障长尾情况下的驾驶性能的稳定性。

[0142] 请参见图2,多个驾驶策略中的任意目标驾驶策略的生成过程包括:

[0143] S201、设定奖励函数,并构建包括奖励函数的初始策略。

[0144] 需要说明的是,不同驾驶策略的奖励函数的奖励系数取值不同,其取值方式可按

照如下公式确定:标识信息为i的驾驶策略的三个奖励系数的计算公式包括:$\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$;$\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$;$\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$。相应地奖励函数的计算公式为:$r=\lambda_{e,i}\times r_e+\lambda_{s,i}\times r_s+\lambda_{ot,i}\times r_{ot}$。在一种实施方式中,设定奖励函数,包括:确定目标驾驶策略在多个驾驶策略中的标识信息;根据标识信息计算第一奖励系数、第二奖励系数和第三奖励系数,并构建奖励函数。

[0145] S202、利用强化学习方法训练初始策略,得到待优化策略。

[0146] S203、利用待优化策略构建优化样本。

[0147] 在一种实施方式中,利用待优化策略构建优化样本,包括:将目标状态输入待优化策略,以使待优化策略输出结束状态和目标窗口;在目标窗口内使目标状态为起始点,使结束状态为终点,并通过曲线拟合确定目标窗口内的各轨迹点;连接各轨迹点得到运动轨迹,并生成能够控制车辆沿运动轨迹行驶的目标策略动作;将目标状态、目标策略动作和目标策略动作的奖励值构建为优化样本。曲线拟合可采用多项式等拟合方法。

[0148] 其中,通过曲线拟合确定目标窗口内的各轨迹点,包括:在目标窗口内拟合得到位移变化曲线;在目标窗口内拟合得到速度变化曲线;匹配位移变化曲线和速度变化曲线中的各点,以确定目标窗口内的各轨迹点。或在目标窗口内拟合得到位移变化曲线;在目标窗口内拟合速度变化曲线时与位移变化曲线进行匹配,以确定目标窗口内的各轨迹点。或在目标窗口内拟合得到速度变化曲线;在目标窗口内拟合位移变化曲线时与速度变化曲线进行匹配,以确定目标窗口内的各轨迹点。一个轨迹点对应一个控制指令,目标窗口内的各轨迹点对应一系列有排列顺序的控制指令,因此基于目标窗口内各轨迹点形成的运动轨迹,能够对应生成控制车辆沿此运动轨迹行驶的目标策略动作。

[0149] S204、在成本函数的约束下,以最大奖励为求解目标,构建拉格朗日目标函数。

[0150] 其中,拉格朗日目标函数含用于约束成本函数的正则化项,能够约束成本函数的值不至于过高,过高的成本函数值会增大安全约束,会使驾驶策略过于保守。

[0151] S205、利用优化样本迭代求解拉格朗日目标函数,以优化待优化策略,得到目标驾驶策略。

[0152] 在一种示例中,在任意一次迭代中执行如下优化步骤:固定拉格朗日乘子和成本函数,通过最大化拉格朗日目标函数优化优化当前待优化策略;和/或固定当前待优化策略和成本函数,通过最小化拉格朗日目标函数优化拉格朗日乘子;和/或固定当前待优化策略和拉格朗日乘子,通过最大化拉格朗日目标函数优化成本函数。可见,上述三个优化步骤可以在每一次迭代中都执行,也可以在每一次迭代中仅执行其中一个或两个。

[0153] 需要说明的是,在一次优化完成后,可以用本次优化得到的驾驶策略执行步骤S202,以基于局部轨迹再次进行策略学习,而后再次执行S203重新构建优化样本,之后再开启新的一轮优化。由此可实现:策略优化与基于局部轨迹学习策略的联合进行,更有利于保障驾驶策略的可靠性。

[0154] 下面对本发明实施例提供的一种车辆控制装置进行介绍,下文描述的一种车辆控制装置与本文描述的其他实施例可以相互参照。

[0155] 参见图3所示,本发明实施例公开了一种车辆控制装置,包括:

[0156] 获取模块301,用于获取多个驾驶策略;

[0157] 收集模块302,用于分别运行每个驾驶策略,并收集每个驾驶策略运行过程中每一

次的输入状态、策略动作及动作奖励,得到每个驾驶策略的运行三元集;策略动作用于控制车辆沿设定轨迹点行驶预设距离;

[0158] 评估模块303,用于根据运行三元集分别计算每个驾驶策略的运行估计值,并根据运行估计值分别确定每个驾驶策略的置信度;

[0159] 应用模块304,用于在多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶。

[0160] 在一种实施方式中,收集模块具体用于:针对每一驾驶策略,利用当前驾驶策略控制真实车辆进行自动驾驶,并收集真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态、策略动作及动作奖励;在自动驾驶结束后,汇总各次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

[0161] 在一种实施方式中,收集模块具体用于:分别利用每个驾驶策略控制同一真实车辆进行自动驾驶,以在同一真实车辆上分别运行每个驾驶策略。

[0162] 在一种实施方式中,收集模块包括:

[0163] 样本准备单元,用于针对每一驾驶策略,利用当前驾驶策略控制真实车辆进行自动驾驶,并收集真实车辆自动驾驶过程中当前驾驶策略每一次的输入状态及策略动作;在自动驾驶结束后,汇总各次的输入状态及策略动作,得到训练样本;

[0164] 训练单元,用于利用训练样本和当前驾驶策略训练得到虚拟驾驶模型;

[0165] 生成单元,用于利用当前驾驶策略和虚拟驾驶模型生成多次的输入状态、策略动作及动作奖励,得到当前驾驶策略的运行三元集。

[0166] 在一种实施方式中,训练单元具体用于:将训练样本和当前驾驶策略训练预设的高斯神经网络模型,得到虚拟驾驶模型。

[0167] 在一种实施方式中,训练单元具体用于:将训练样本划分为至少两个子样本集;利用每个子样本集和当前驾驶策略分别训练一个子模型,得到至少两个子模型;在至少两个子模型中选择模型评估值最低的子模型作为虚拟驾驶模型。

[0168] 在一种实施方式中,训练单元具体用于:计算每个子模型在训练样本上的模型评估值;选择模型评估值最低的子模型作为虚拟驾驶模型。

[0169] 在一种实施方式中,生成单元具体用于:若当前迭代次数未超出预测总次数,则获取前一次输入状态及前一次策略动作;将前一次输入状态及前一次策略动作输入虚拟驾驶模型,以使虚拟驾驶模型输出当前输入状态;使当前驾驶策略根据当前输入状态输出当前策略动作;使当前驾驶策略对应的奖励函数根据当前策略动作计算当前动作奖励;将当前输入状态、当前策略动作和当前动作奖励构建为三元组,并将三元组作为当前驾驶策略的运行三元集中的一个元素;将当前输入状态作为前一次输入状态,将当前策略动作作为前一次策略动作,并使当前迭代次数递增一,然后判断当前迭代次数是否超出预测总次数。

[0170] 在一种实施方式中,奖励函数为:$r=\lambda_e\times r_e+\lambda_s\times r_s+\lambda_{ot}\times r_{ot}$;$r$为当前动作奖励,$\lambda_e$为当前驾驶策略的第一奖励系数,$\lambda_s$为当前驾驶策略的第二奖励系数,$\lambda_{ot}$为当前驾驶策略的第三奖励系数,$r_e$为当前车辆效率,$r_s$为当前安全奖励,$r_{ot}$为当前超车奖励。

[0171] 在一种实施方式中,多个驾驶策略中的任意驾驶策略i的第一奖励系数、第二奖励系数和第三奖励系数的计算公式包括:$\lambda_{e,i}=\lambda_{e,max}-[(i-1)(\lambda_{e,max}-\lambda_{e,min})]/m$;$\lambda_{s,i}=\lambda_{s,min}-[i(\lambda_{s,max}-\lambda_{s,min})]/m$;$\lambda_{ot,i}=\lambda_{ot,min}-[i(\lambda_{ot,max}-\lambda_{ot,min})]/m$;其中,$\lambda_{e,i}$为驾驶策略$i$的第一奖励系

数，$\lambda_{s,i}$为驾驶策略i的第二奖励系数，$\lambda_{ot,i}$为驾驶策略i的第三奖励系数，$\lambda_{e,max}$为第一奖励系数对应的预设最大值，$\lambda_{e,min}$为第一奖励系数对应的预设最小值，$\lambda_{s,max}$为第二奖励系数对应的预设最大值，$\lambda_{s,min}$为第二奖励系数对应的预设最小值，$\lambda_{ot,max}$为第三奖励系数对应的预设最大值，$\lambda_{ot,min}$为第三奖励系数对应的预设最小值，m为驾驶策略的总个数。

[0172] 在一种实施方式中，还包括：驾驶策略生成模块，该模块用于生成多个驾驶策略中的任意目标驾驶策略；其中，驾驶策略生成模块包括：

[0173] 初始单元，用于设定奖励函数，并构建包括奖励函数的初始策略；

[0174] 训练单元，用于利用强化学习方法训练初始策略，得到待优化策略；

[0175] 样本构建单元，用于利用待优化策略构建优化样本；

[0176] 函数构建单元，用于在成本函数的约束下，以最大奖励为求解目标，构建拉格朗日目标函数；

[0177] 优化单元，用于利用优化样本迭代求解拉格朗日目标函数，以优化待优化策略，得到目标驾驶策略。

[0178] 在一种实施方式中，样本构建单元具体用于：将目标状态输入待优化策略，以使待优化策略输出结束状态和目标窗口；在目标窗口内使目标状态为起始点，使结束状态为终点，并通过曲线拟合确定目标窗口内的各轨迹点；连接各轨迹点得到运动轨迹，并生成能够控制车辆沿运动轨迹行驶的目标策略动作；将目标状态、目标策略动作和目标策略动作的奖励值构建为优化样本。

[0179] 在一种实施方式中，样本构建单元具体用于：在目标窗口内拟合得到位移变化曲线；在目标窗口内拟合得到速度变化曲线；匹配位移变化曲线和速度变化曲线中的各点，以确定目标窗口内的各轨迹点。

[0180] 在一种实施方式中，样本构建单元具体用于：在目标窗口内拟合得到位移变化曲线；在目标窗口内拟合速度变化曲线时与位移变化曲线进行匹配，以确定目标窗口内的各轨迹点。

[0181] 在一种实施方式中，样本构建单元具体用于：在目标窗口内拟合得到速度变化曲线；在目标窗口内拟合位移变化曲线时与速度变化曲线进行匹配，以确定目标窗口内的各轨迹点。

[0182] 在一种实施方式中，拉格朗日目标函数含用于约束成本函数的正则化项。

[0183] 在一种实施方式中，在任意一次迭代中执行如下优化步骤：固定拉格朗日乘子和成本函数，通过最大化拉格朗日目标函数优化优化当前待优化策略；和/或固定当前待优化策略和成本函数，通过最小化拉格朗日目标函数优化拉格朗日乘子；和/或固定当前待优化策略和拉格朗日乘子，通过最大化拉格朗日目标函数优化成本函数。

[0184] 在一种实施方式中，初始单元具体用于：确定目标驾驶策略在多个驾驶策略中的标识信息；根据标识信息计算第一奖励系数、第二奖励系数和第三奖励系数，并构建奖励函数。

[0185] 在一种实施方式中，应用模块具体用于：使置信度最高的驾驶策略针对车辆当前状态输出可信策略动作；按照可信策略动作确定由多个控制指令构成的指令序列；按照指令序列控制车辆沿设定轨迹点自动行驶预设距离。

[0186] 其中，关于本实施例中各个模块、单元更加具体的工作过程可以参考前述实施例

中公开的相应内容,在此不再进行赘述。

[0187] 可见,本实施例提供了一种车辆控制装置,能够基于驾驶策略的置信度衡量驾驶策略的风险程度,并选择置信度最高的驾驶策略来确保车辆驾驶期间始终执行最优驾驶策略,保障长尾情况下的驾驶性能的稳定性。

[0188] 下面对本发明实施例提供的一种电子设备进行介绍,下文描述的一种电子设备与本文描述的其他实施例可以相互参照。

[0189] 参见图4所示,本发明实施例公开了一种电子设备,包括:

[0190] 存储器401,用于保存计算机程序;

[0191] 处理器402,用于执行所述计算机程序,以实现上述任意实施例公开的方法。

[0192] 进一步的,本发明实施例还提供了一种电子设备。其中,上述电子设备既可以是如图5所示的服务器,也可以是如图6所示的终端。图5和图6均是根据一示例性实施例示出的电子设备结构图,图中的内容不能被认为是对本发明的使用范围的任何限制。

[0193] 图5为本发明实施例提供的一种服务器的结构示意图。该服务器具体可以包括:至少一个处理器、至少一个存储器、电源、通信接口、输入输出接口和通信总线。其中,所述存储器用于存储计算机程序,所述计算机程序由所述处理器加载并执行,以实现前述任一实施例公开的车辆控制中的相关步骤。

[0194] 本实施例中,电源用于为服务器上的各硬件设备提供工作电压;通信接口能够为服务器创建与外界设备之间的数据传输通道,其所遵循的通信协议是能够适用于本发明技术方案的任意通信协议,在此不对其进行具体限定;输入输出接口,用于获取外界输入数据或向外界输出数据,其具体的接口类型可以根据具体应用需要进行选取,在此不进行具体限定。

[0195] 另外,存储器作为资源存储的载体,可以是只读存储器、随机存储器、磁盘或者光盘等,其上所存储的资源包括操作系统、计算机程序及数据等,存储方式可以是短暂存储或者永久存储。

[0196] 其中,操作系统用于管理与控制服务器上的各硬件设备以及计算机程序,以实现处理器对存储器中数据的运算与处理,其可以是Windows Server、Netware、Unix、Linux等。计算机程序除了包括能够用于完成前述任一实施例公开的车辆控制方法的计算机程序之外,还可以进一步包括能够用于完成其他特定工作的计算机程序。数据除了可以包括应用程序的更新信息等数据外,还可以包括应用程序的开发商信息等数据。

[0197] 图6为本发明实施例提供的一种终端的结构示意图,该终端具体可以包括但不限于智能手机、平板电脑、笔记本电脑或台式电脑等。

[0198] 通常,本实施例中的终端包括有:处理器和存储器。

[0199] 其中,处理器可以包括一个或多个处理核心,比如4核心处理器、8核心处理器等。处理器可以采用DSP(Digital Signal Processing,数字信号处理)、FPGA(Field—Programmable Gate Array,现场可编程门阵列)、PLA(Programmable Logic Array,可编程逻辑阵列)中的至少一种硬件形式来实现。处理器也可以包括主处理器和协处理器,主处理器是用于对在唤醒状态下的数据进行处理的处理器,也称CPU(Central Processing Unit,中央处理器);协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中,处理器可以在集成有GPU(Graphics Processing Unit,图像处理器),GPU用于负责

显示屏所需要显示的内容的渲染和绘制。一些实施例中,处理器还可以包括AI(Artificial Intelligence,人工智能)处理器,该AI处理器用于处理有关机器学习的计算操作。

[0200] 存储器可以包括一个或多个计算机可读存储介质,该计算机可读存储介质可以是非暂态的。存储器还可包括高速随机存取存储器,以及非易失性存储器,比如一个或多个磁盘存储设备、闪存存储设备。本实施例中,存储器至少用于存储以下计算机程序,其中,该计算机程序被处理器加载并执行之后,能够实现前述任一实施例公开的由终端侧执行的车辆控制方法中的相关步骤。另外,存储器所存储的资源还可以包括操作系统和数据等,存储方式可以是短暂存储或者永久存储。其中,操作系统可以包括Windows、Unix、Linux等。数据可以包括但不限于应用程序的更新信息。

[0201] 在一些实施例中,终端还可包括有显示屏、输入输出接口、通信接口、传感器、电源以及通信总线。

[0202] 本领域技术人员可以理解,图6中示出的结构并不构成对终端的限定,可以包括比图示更多或更少的组件。

[0203] 下面对本发明实施例提供的一种可读存储介质进行介绍,下文描述的一种可读存储介质与本文描述的其他实施例可以相互参照。

[0204] 本发明实施例提供了一种可读存储介质,用于保存计算机程序,其中,所述计算机程序被处理器执行时实现前述实施例公开的车辆控制方法。其中,可读存储介质为计算机可读存储介质,其作为资源存储的载体,可以是只读存储器、随机存储器、磁盘或者光盘等,其上所存储的资源包括操作系统、计算机程序及数据等,存储方式可以是短暂存储或者永久存储。

[0205] 下面进一步介绍驾驶策略的生成过程,下文描述的内容与本文其他实施例可以相互参照。

[0206] 当前自动驾驶策略用于产生车辆级别的控制命令,如:每个时刻的车辆转向、加速指令等,这种单步控制的自动驾驶策略难以实现复杂度更高的高级驾驶行为。而本实施例能够学习得到基于局部运动规划的驾驶策略,驾驶策略的输出为策略动作。

[0207] 具体的,为了表示包括自我车辆、周围目标的时空信息、道路几何和导航信息等在内的驾驶环境,本实施例通过鸟瞰图(Birds-eye view,BEV)来表征车辆驾驶环境,将BEV图像作为驾驶策略的输入,也就是说:驾驶策略的输入状态不仅包括车辆的位置、速度等车辆信息,还包括车辆周围目标的时空信息、道路几何和导航信息等。驾驶策略的输出为控制车辆行驶一段距离的策略动作,根据驾驶策略的输入及输出的策略动作,并借助曲线拟合,可以生成与该策略动作对应的一条曲率连续的行驶轨迹。

[0208] 给定规划窗口和起止边界条件,利用强化学习方法使驾驶策略学习起始边界条件到结束边界条件的映射。其中,起始边界条件包括车辆起始位置(x_s, y_s)、航向角φ_s、速度v_s、加速度$\dot{v}_s$;结束边界条件,即T时刻后的车辆行驶状态,包括车辆结束位置(x_e, y_e)、航向角φ_e、速度v_e、加速度$\dot{v}_e$。

[0209] 在驾驶策略输出策略动作后,可相应生成一条局部轨迹曲线,将曲线按照设定执行步长进行离散化处理,即可得到一系列轨迹点及其对应的控制指令序列。

[0210] 对于一条待规划路径,当已知路径的起止点位置信息,可以采用多项式曲线拟合

方法产生连接起点与终点的运动轨迹。其中,路径结束点处的位姿由纵向位置y_e、横向位置x_e和航向角φ_e三个参数表征。

[0211] 其中,结束点处的横向位置取值范围通常根据可行驶车道的数量及宽度设置为连续区间,易导致车辆出现长时间压线行驶的行为。也就是:以道路中心线为基准设置一个最大横向偏移距离x_{max},以限制x_e的取值范围$x_e \in (-x_{max}, x_{max})$,确保车辆处于可行驶区域内,然而这将导致车辆一味准求高行驶效率,而出现长时间跨车道线行驶的行为,不符合实际驾驶要求。为鼓励车辆尽可能处于车道中心线,同时可实现变道、超车等任务,这里将取值范围设定为离散位置点集;也就是:将x_e的取值设定在车道中心点处,考虑车辆变道、超车等驾驶行为,x_e可设定在当前车道、左侧车道或右侧车道,因此x_e应当具有离散的取值空间$x_e \in \{x_{left}, x_{center}, x_{right}\}$,其中$x_{left}$、$x_{center}$、$x_{right}$分别为左车道、当前车道和右车道的中心点。

[0212] 其中,结束点处的纵向位置通常会被设定为车辆在固定规划时间窗口内可以达到的最远距离,也就是:将路径结束点纵向位置设定为车辆在固定规划时间窗口T内可以达到的最远距离:$y_e = y_s + v_{max} \times T$(1)。其中$v_{max}$为车辆最大速度。该方式虽然能够确保可行的路径-速度匹配投影,但会使得每次规划得到的轨迹执行步数固定,无法根据驾驶情况自适应调整,当出现突发紧急事件,车辆将难以做出应急避让行为,缺乏风险响应能力。为满足动态不确定场景中的车辆多样化驾驶需求,本发明将y_e和规划窗口T均作为待学习参数,从而使驾驶策略可根据车辆所处环境的风险程度自适应调整路径长度,以实现长度可变的自适应路径规划,有效增强所生成候选路径的灵活性与风险响应能力。其中,$y_e \in (0, y_{max})$,y_{max}为车辆最远感知距离,$\mathrm{T} \in (0, \mathrm{T}_{max})$,$\mathrm{T}_{max}$为最大规划窗口。也就是说:强化学习过程中,每次所生成的运动轨迹的长度不同。

[0213] 对于结束点处的航向角φ_e,考虑车辆自身动力学限制,其取值范围表示为$\varphi_e \in (-\varphi_{max}, \varphi_{max})$,其中$\varphi_{max}$为车辆最大航向角。

[0214] 其中,路径结束点处的运动状态由速度v_e、加速度$\dot{v}_e$两个参数表征,二者取值范围均受车辆自身动力学限制,分别表示为$v_e \in (v_s - \dot{v}_{max}\mathrm{T}_{max}, v_s + \dot{v}_{max}\mathrm{T}_{max})$,$\dot{v}_e \in (-\dot{v}_{max}, \dot{v}_{max})$,其中$\dot{v}_{max}$为车辆最大加速度。

[0215] 综上所述,对于任意一个局部运动轨迹,其待学习参数包括横向位置x_e、纵向位置y_e、航向角φ_e、规划窗口T、速度v_e以及加速度$\dot{v}_e$,因此本发明将驾驶策略的输出动作设计为$a = (x_e, y_e, \varphi_e, \mathrm{T}, v_e, \dot{v}_e)$,各动作分量的取值范围表示为:

$$\begin{cases} x_e \in \{x_{left}, x_{center}, x_{right}\} \\ y_e \in (0, y_{max}) \\ \mathrm{T} \in (0, \mathrm{T}_{max}) \\ \varphi_e \in (-\varphi_{max}, \varphi_{max}) \\ v_e \in (v_s - \dot{v}_{max}\mathrm{T}_{max}, v_s + \dot{v}_{max}\mathrm{T}_{max}) \\ \dot{v}_e \in (-\dot{v}_{max}, \dot{v}_{max}) \end{cases} \quad (2)。$$

[0216] 在实际应用中,车辆行驶路径可行性通常会受到车辆自身动力学约束,包括转向角、安全距离等。因此路径上的任意一点的曲率必须小于目标的最大曲率限制(或最小转弯半径约束),同时所产生的可行路径还应该是曲率连续的。为满足路径可行性约束,同时降低求解空间维度,这里基于五次多项式曲线分别生成位移变化曲线和速度变化曲线,将车辆运动规划问题转化为曲线参数的搜索寻优。还可以采用其他方式进行曲线拟合,如三次多项式曲线拟合等。

[0217] 位移变化曲线的生成:假设一次规划期中车辆起始状态为s,驾驶策略根据状态s获取到策略动作,此时已知路径起始状态为$\mathrm{p}_s=(x_s,y_s,\varphi_s,v_s,\dot{v}_s)$,结束状态为$\mathrm{p}_e=(x_e,y_e,\varphi_e,v_e,\dot{v}_e)$,采用5次多项式描述横向位置$x$与纵向位置$y$的变化关系:$x=f_p(y)=\alpha_0+\alpha_1(y-y_s)+\alpha_2(y-y_s)^2+\alpha_3(y-y_s)^3+\alpha_4(y-y_s)^4+\alpha_5(y-y_s)^5$ (3);其中$\alpha_0\sim\alpha_5$均为路径曲线系数。

[0218] 基于路径起止状态构造端点约束条件:$$\begin{cases}f_p(y_s)=x_s\\ f_p'(y_s)=\tan\varphi_s\\ f_p''(y_s)=0\\ f_p(y_e)=x_e\\ f_p'(y_e)=\tan\varphi_e\\ f_p''(y_e)=0\end{cases}$$ (4)。

[0219] 根据端点约束条件可以通过下式求解各系数:

[0220] $$\begin{cases}\alpha_0=x_s\\ \alpha_1=\tan\varphi_s\\ \alpha_2=0\\ \alpha_3=[10(x_e-x_s)-(4\tan\varphi_e+6\tan\varphi_s)(y_e-y_s)]/(y_e-y_s)^3\\ \alpha_4=[-15(x_e-x_s)+(7\tan\varphi_e+8\tan\varphi_s)(y_e-y_s)]/(y_e-y_s)^4\\ \alpha_5=[6(x_e-x_s)-(3\tan\varphi_e+3\tan\varphi_s)(y_e-y_s)]/(y_e-y_s)^5\end{cases}$$ (5)。

[0221] 对车辆纵向位置在$[y_s,y_e]$内等间隔采样取点,结合公式(3)与(5),可得到一系列离散轨迹点,其中采样间隔的设置与规划窗口T相关,表示为:$\Delta y=0.01(y_e-y_s)/\mathrm{T}$ (6)。

[0222] 将轨迹点序列表示为$\{(x_s,y_s),\cdots,(x_i,y_i),\cdots,(x_e,y_e)\}_{1\le i\le 100\mathrm{T}-1}$,该序列包含$100\mathrm{T}+1$个轨迹点,其中$(x_i,y_i)$为第$i$个中间轨迹点,该点处的车辆航向角通过下式计算:$\varphi_i=\left[\tan^{-1}\left((x_i-x_{i-1})/(y_i-y_{i-1})\right)+\tan^{-1}\left((x_{i+1}-x_i)/(y_{i+1}-y_i)\right)\right]/2$ (7)。

[0223] 综合航向角和各轨迹点处位置信息,即可得到该规划周期内的局部候选路径$\{(x_i,y_i,\varphi_i)\}_{0\le i\le 100\mathrm{T}}$,其中,起始轨迹点$(x_0,y_0,\varphi_0)=(x_s,y_s,\varphi_s)$,结束轨迹点$(x_{100\mathrm{T}},y_{100\mathrm{T}},\varphi_{100\mathrm{T}})=(x_e,y_e,\varphi_e)$。

[0224] 速度变化曲线的生成:由于位移变化曲线不包含速度、加速度等车辆运动信息,为

确保车辆可沿该路径行驶,还需相应地进行速度规划。已知路径起止点处的车辆运动信息,采用5次多项式描述速度v与时间步t的变化关系:

$v = f_v(t) = \beta_0 + \beta_1(t-t_s) + \beta_2(t-t_s)^2 + \beta_3(t-t_s)^3 + \beta_4(t-t_s)^4 + \beta_5(t-t_s)^5$(8);其中$t_s$为局部路径的起始时间步,$\beta_0 \sim \beta_5$为速度曲线系数。为便于处理,这里令起始时间步$t_s = 0$,结束时间步等同于规划窗口大小,即$t_e = \mathrm{T}$。

[0225] 根据起止点处的车辆运动状态构造端点约束条件:$\begin{cases} f_v(0) = v_s \\ f_v'(0) = \dot{v}_s \\ f_v''(0) = 0 \\ f_v(\mathrm{T}) = v_e \\ f_v'(\mathrm{T}) = \dot{v}_e \\ f_v''(\mathrm{T}) = 0 \end{cases}$(9)。

[0226] 根据端点约束条件可以通过下式求解各系数:

[0227] $\begin{cases} \beta_0 = v_s \\ \beta_1 = \dot{v}_s \\ \beta_2 = 0 \\ \beta_3 = [10(v_e - v_s) - (4\dot{v}_e + 6\dot{v}_s)\mathrm{T}]/\mathrm{T}^3 \\ \beta_4 = [-15(v_e - v_s) + (7\dot{v}_e + 8\dot{v}_s)\mathrm{T}]/\mathrm{T}^4 \\ \alpha_5 = [6(v_e - v_s) - (3\dot{v}_e + 3\dot{v}_s)\mathrm{T}]/\mathrm{T}^5 \end{cases}$(10)。

[0228] 同样在规划窗口内对时间步等间隔采样取点,结合公式(8)与(10),可得到一系列离散速度值。为了减小各轨迹点处位置和速度的匹配误差,这里将速度采样点数设置为路径采样点数的1/10,因此运动曲线的采样间隔设置为$\Delta t = 0.1$,产生$10\mathrm{T}+1$个运动点$\{(v_j, \dot{v}_j)\}_{0 \le j \le 10\mathrm{T}}$,其中,起始运动点$(v_0, v_0) = (v_s, \dot{v}_s)$,结束运动点$(v_{10\mathrm{T}}, a_{10\mathrm{T}}) = (v_e, a_e)$,$\dot{v}_j$为$j\Delta t$时刻的加速度值,通过下式估算:$\dot{v}_j = 0.5(v_{j+1} - v_{j-1})/\Delta\mathrm{t}$(11)。

[0229] 将位移变化曲线和速度变化曲线进行匹配:将位移变化曲线中的各点和速度变化曲线中的各点进行匹配,可构成车辆的行驶轨迹。

[0230] 首先对运动曲线(即速度变化曲线)进行积分,计算第j个离散速度值处的车辆已行驶距离:$dist(j) = \int_0^{j\Delta t} f_v(t)dt \approx \beta_0(j\Delta t) + \frac{1}{2}\beta_1(j\Delta t)^2 + \frac{1}{3}\beta_2(j\Delta t)^3 + \frac{1}{4}\beta_3(j\Delta t)^4 + \frac{1}{5}\beta_4(j\Delta t)^5 + \frac{1}{6}\beta_5(j\Delta t)^6$(12)。根据上式可得到与运动序列相对应的第一距离序列$\{dist(j)\}_{0 \le j \le 10\mathrm{T}}$,其中车辆行驶距离在$dist(j)$处的速度和加速度为$(v_j, \dot{v}_j)$。

[0231] 然后,基于路径序列$\{(x_i, y_i, \varphi_i)\}_{0 \le i \le 100\mathrm{T}}$计算行驶距离:

$\widehat{dist}(i) = \sum_{k=1}^{i} \sqrt{(x_k - x_{k-1})^2 + (y_k - y_{k-1})^2}$(13)。根据上式可得到与路径序列相对应的第二距离序列$\{\widehat{dist}(i)\}_{0 \le i \le 100\mathrm{T}}$,其中车辆行驶距离在$\widehat{dist}(i)$处的位置和转向角为

(x_i, y_i, φ_i)。

[0232] 由于该示例中路径曲线(即位移变化曲线)和运动曲线的生成过程相对独立,无法直接按照规划步长或序列索引将其等间隔对应起来,此处设计了一种基于距离信息的路径-运动状态匹配方法,通过对比路径序列和运动序列的累计行驶距离,实现运动点与路径点间的一一对应。即:通过对比上述第一距离序列和第二距离序列,实现速度运动点与位移点间的耦合。对于任意速度运动点$(v_j, \dot{v}_j)$,已知该点处的行驶距离应当为$dist(j)$,按照距离值大小选取与之最接近的轨迹点索引:$ind = \min_{0\le i\le 100T}\left(\widehat{dist}(i) - dist(j) \ge 0\right)$(14)。

[0233] 为确保路径与运动状态匹配准确度,此处不直接采用路径序列中的采样点与运动点进行硬匹配,而是通过计算距离接近比,进一步根据距离接近程度计算与该运动点匹配的轨迹点位姿:$\begin{cases} \hat{x} = x_{idx-1} + \epsilon(x_{idx} - x_{idx-1}) \\ \hat{y} = y_{idx-1} + \epsilon(y_{idx} - y_{idx-1}) \\ \hat{\varphi} = \varphi_{idx-1} + \epsilon(\varphi_{idx} - \varphi_{idx-1}) \end{cases}$(15)。其中,$\epsilon$为距离权重,通过下式计算:

$\epsilon = \left(dist(j) - \widehat{dist}(ind-1)\right) / \left(\widehat{dist}(ind) - \widehat{dist}(ind-1)\right)$(16)。

[0234] 最后,综合运动与位姿信息,即可得到所需运动轨迹$\tau = \{\mathrm{p}_j = (\hat{x}_j, \hat{y}_j, \hat{\varphi}_j, v_j, \dot{v}_j)\}_{0\le j\le 10T}$。

[0235] 参照本示例中的前述内容,并采用强化学习框架可训练基于运动规划的驾驶策略。强化学习是一项强大的自学习技术,该框架中,策略将通过与环境交互不断进行探索与试错,可以在线产生学习样本。以这些样本为基础,可将预期累计回报最大化作为优化目标进行策略的优化,这里采用π表示驾驶策略,π^*表示最优策略,其优化过程可以表示为:$\pi^* = arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{T_i} \gamma^t r(s_t, a_t = \pi(s_t))\right]$(17)。其中$T_i$为一次交互过程总步长,$r$为奖励函数,通常设计为行驶效率、安全性、平稳性等量化指标的线性组合,$\gamma \in (0,1)$为奖励衰减因子。期间,策略性能通过价值函数进行评估:$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{T_i} \gamma^t r(s_t, a_t) | s_0 = s\right]$(18)。

[0236] 因此,驾驶策略的优化目标可以表示为:$\pi^* = arg\max_\pi \mathbb{E}_s[V^\pi(s)]$(19)。

[0237] 为了提升驾驶安全性,同时避免策略过保守对交通效率产生负面影响,本发明在策略训练过程中引入基于安全间距的成本约束项,以实现驾驶安全和效率之间的良好平衡,则有:$\max_\pi \mathbb{E}_s[V^\pi(s)]\ \mathrm{s.t.}\ \ \phi(s') - \max\{\phi(s) - \eta_d, 0\} < 0$(20)。一般研究为了确保安全性会对驾驶策略施加严格的安全约束,不考虑安全约束对于效率的影响,可能会导致策略过于保守。本发明设计了一种与安全指数相关的成本函数,对驾驶策略目标函数进行约束,可以在确保安全性的同时,避免策略过保守。

[0238] 公式(20)中,η_d是控制安全指数下降率的松弛变量,$\phi(s)$是为了避免碰撞所设置的安全成本函数,定义为:$\phi(s) = (\sigma + d_{min})^n - d^n - k\dot{d}$(21)。$d$是车辆与要避开的运动目标之间的距离,$d_{min}$是最小安全距离,$\dot{d}$是距离相对于时间的导数,$\xi = [\sigma, k, n]$是待优化的可调参数。成本值越高意味着安全指数越低,这时安全约束也将变得更保守。采用一种可学习的成本函数,避免人为设计的成本和误差,该成本函数可以在训练期间和驾驶策略进行联合优化。

[0239] 为求解式(20)中带约束的策略优化问题,构建拉格朗日函数作为目标函数:

$L(\pi,\lambda,\phi)=\mathbb{E}_s[V^\pi(s)+\lambda(s)(\phi(s')-\max\{\phi(s)-\eta_d,0\})]$(22)；其中，$\lambda(s)$为拉格朗日乘子网络，用于处理状态约束。基于该目标函数，可以实现驾驶策略π和成本函数ϕ的联合学习。

[0240] 为了避免成本函数值过高影响效率，这里在目标函数中添加一个与成本值相关的正则化项，则有：$\hat{L}(\pi,\lambda,\phi)=L(\pi,\lambda)+\alpha(\sigma+d_{min})^n+bk$(23)。通过引入与成本值大小相关的正则化项，防止策略过保守，实现效率上的提升。其中，α和b均为超参数，取值在$[0,1]$范围内。基于上述优化目标，在每个迭代周期中，策略π、拉格朗日乘子λ以及成本函数ϕ将交替完成更新。

[0241] 在每一次迭代中，可以执行如下优化步骤：固定拉格朗日乘子λ和成本函数ϕ，通过最大化目标函数学习策略π，则有：$\pi^*\sim\max_\pi\hat{L}(\pi,\lambda,\phi)$(24)。固定策略$\pi$和成本函数$\phi$，通过最小化目标函数学习拉格朗日乘子$\lambda$，则有：$\lambda^*\sim\min_\lambda\hat{L}(\pi,\lambda,\phi)$(25)。固定策略$\pi$和拉格朗日乘子$\lambda$，通过最大化目标函数学习成本函数$\phi$，则有：$\phi^*\sim\max_\phi\hat{L}(\pi,\lambda,\phi)$(26)。

[0242] 请参见图7，一个驾驶策略的训练过程包括以下步骤：基于策略π执行环境交互，收集驾驶数据后，执行如下步骤：

[0243] ①获取车辆驾驶状态s，采用策略π选择规划参数$a=(x_e,y_e,\varphi_e,\mathrm{T},v_e,\dot{v}_e)$，并将车辆当前位姿与运动信息作为起始状态$\mathrm{p}_s=(x_s,y_s,\varphi_s,v_s,\dot{v}_s)$；

[0244] ②基于起始位姿(x_s,y_s,φ_s)、结束位姿(x_e,y_e,φ_e)以及规划窗口T，生成路径曲线$\{(x_i,y_i,\varphi_i)\}_{0\le i\le 100\mathrm{T}}$；

[0245] ③基于起始运动状态$(v_s,\dot{v}_s)$、结束位姿$(v_e,\dot{v}_e)$以及规划窗口T，生成运动曲线$\{(v_j,\dot{v}_j)\}_{0\le j\le 10\mathrm{T}}$；

[0246] ④结合路径曲线和运动曲线进行信息匹配，生成运动轨迹$\tau=\{\mathrm{p}_j=(\hat{x}_j,\hat{y}_j,\hat{\varphi}_j,v_j,\dot{v}_j)\}_{0\le j\le 10\mathrm{T}}$；

[0247] ⑤车辆执行轨迹τ，得到环境反馈r，将驾驶数据(s,a,r)放入数据集$\mathcal{D}$；

[0248] ⑥判断是否达到最大交互步数T_i，若是则结束交互，进入策略优化阶段，反之转至步骤①。

[0249] 策略优化阶段：利用历史驾驶数据进行策略优化：

[0250] ①从数据集$\mathcal{D}$中随机采样小批次数据$\{(s,a,r)\}\sim\mathcal{D}$；

[0251] ②基于式(24)更新策略π；

[0252] ③基于式(25)更新拉格朗日乘子网络；

[0253] ④基于式(26)成本函数ϕ；

[0254] ⑤判断结束条件：若迭代数达到上限，则结束更新；否则，迭代数+1，转至步骤①。

[0255] 可见，本实施例采用鸟瞰图作为驾驶策略状态输入表示形式，并设计基于运动规

划参数的动作表示;将车辆运动规划问题转化为曲线参数的搜索寻优,设计基于驾驶策略动作输出的车辆轨迹生成方法;结合自适应安全成本函数学习驾驶策略,并设计了带有可学习安全成本函数的强化学习框架训练基于运动规划的驾驶策略,实现训练期间安全约束的自适应调节。

[0256] 由于强化学习训练过程的不稳定和神经网络固有的内在不确定性,自动驾驶策略难以确保良好的安全性和稳定的驾驶性能,无法满足车辆平稳驾驶需求。本发明提出一种自适应安全约束的驾驶策略学习方法,通过将运动规划参数作为策略输出,设计基于运动规划的驾驶策略学习方法,以实现规划窗口自适应可调的车辆运动规划,确保车辆行驶轨迹的平稳性和复杂驾驶行为的实现能力,提升驾驶策略的可行性;在策略训练期间,将可学习的安全成本函数与强化学习的优化目标相结合进行自适应安全性约束,在确保驾驶安全性的同时避免过保守行为,增强策略学习稳定性,实现安全性和效率双提升。

[0257] 本发明提出的一种自适应安全约束的驾驶策略学习方法,以运动规划参数作为待学习动作空间,设计面向运动规划的自动驾驶策略,实现规划窗口自适应可调的车辆运动规划,同时将可学习的安全成本函数与强化学习优化目标相结合,实现驾驶策略的自适应安全约束训练。

[0258] 需要说明的是,可通过调整驾驶策略中奖励函数的奖励系数取值来调整驾驶策略的驾驶偏好,由此可确定具有不同驾驶偏好的候选策略$\Pi = G_d(s)$,其中s表示车辆当前驾驶状态,$G_d(\cdot)$表示策略训练过程。策略集Π中包含m个候选策略:$\Pi = (\pi_1, \pi_2, \cdots, \pi_m)$(27)。这些策略可以具有相同的网络结构,但驾驶偏好或保守或激进。

[0259] 为便于调整奖励函数的奖励系数,采用如下奖励函数形式:$r = \lambda_e r_e + \lambda_s r_s + \lambda_{ot} r_{ot}$(28);其中,$r_e$、$r_s$和$r_{ot}$分别为车辆效率、安全和超车奖励,各奖惩项具体设置为:$r_e = -\lambda_j R_{jerk}(s,a) - \lambda_v \left| \dot{x}_e - \dot{x}_{target} \right|$,$\lambda_j$和$\lambda_v$是用于调节奖励尺度的系数,$R_{jerk}$是根据$s$和动作$a$生成局部轨迹$\mathbf{X}$的加加速度平方间积分:$R_{jerk} = \int \dddot{x}\, d\mathbf{X}$(29),$\dddot{x}$表示轨迹$\mathbf{X}$任一点处的加速度。$r_s = \mathbb{I}_{collision} r_c$,$\mathbb{I}_{collision}$是碰撞指示标志,若发生碰撞,$\mathbb{I}_{collision} = 1$,否则为0,$r_c$是碰撞惩罚值。$r_{ot}$=$\mathbb{I}_{ot} r_{pass}$,$\mathbb{I}_{ot}$是超车指示标志,若实现超车行为,$\mathbb{I}_{ot} = 1$,否则为0,$r_{pass}$是超车奖励值。

[0260] 通过调节各奖励项的权重可实现差异化奖励设计,这里不同的奖励权重对应于不同的驾驶保守水平。式(28)中,λ_e、λ_s和λ_{ot}对应各奖励项系数,取值范围表示为$\left[\lambda_{x,min} - \lambda_{x,max}\right]_{x \in \{e,s,ot\}}$。假设$\pi_1$到$\pi_m$策略的激进性逐渐增强,对于其中任意策略$\pi_i$,其奖励系数设置为:

[0261]
$$\begin{cases} \lambda_{e,i} = \lambda_{e,max} - \frac{i-1}{m}\left(\lambda_{e,max} - \lambda_{e,min}\right) \\ \lambda_{s,i} = \lambda_{s,min} + \frac{i}{m}\left(\lambda_{s,max} - \lambda_{s,min}\right) \\ \lambda_{ot,i} = \lambda_{ot,min} + \frac{i}{m}\left(\lambda_{ot,max} - \lambda_{ot,min}\right) \end{cases} \quad (30)。$$

[0262] 基于候选策略集Π,这里定义一个包含所有候选动作的离散动作空间:

$A=(a_1,a_2,\cdots,a_m)$(31);其中,任意动作a_m将由相应的候选策略π_m生成,$a_m=\pi_m(s)$(32)。

[0263] 进一步,针对每个候选动作,可以按照式$\mathbf{X}=f_s(a)=[s,\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_T]$(33)生成相应的局部运动轨迹,以进行策略评估。$a=\{x_e,\varphi_e,\dot{x}_e,\ddot{x}_e\},s=\{x_s,y_s,\varphi_s,\dot{x}_s,\ddot{x}_s\}$。在车辆行驶期间,驾驶策略将根据车辆每一时刻的驾驶状态s选择轨迹参数a,采用π表示驾驶策略,则驾驶状态到动作间的映射关系可表示为$a=\pi(s)$,进一步采用运动规划方法生成局部运动轨迹。$\mathrm{x}_t=\{x_t,y_t,\varphi_t,\dot{x}_t,\ddot{x}_t\}$表示轨迹$\mathbf{X}$中的第$t$个轨迹点。由于每次规划期内都将生成T个时间步的运动轨迹,为了确保车辆驾驶过程中应对突发情况的机动响应能力,在获取到局部轨迹$\mathbf{X}$后,车辆仅执行第一个轨迹点x_1完成单步状态转移,同时考虑整个局部轨迹的累计奖励。

[0264] 通过设置一组具有不同驾驶偏好的候选策略,后续可通过动态置信估计实时调整将要执行的策略,确保车辆驾驶性能始终是最优的,增强驾驶安全性和稳定性。也即:同一车辆上部署多个具有不同驾驶偏好的驾驶策略,并实时估计每一驾驶策略的置信度,从而实时选择置信度最高的驾驶策略来控制车辆进行自动驾驶。

[0265] 置信值(即置信度)定义:一般情况下,基于强化学习的驾驶策略可以通过估计价值函数和动作-价值函数来进行性能评估,给定驾驶策略π,其价值函数V_π和动作-价值函数Q_π分别为:$V_\pi(s)=\mathbb{E}_\pi[\sum_{t=0}^{H}\gamma^t r_t|s_0=s]$(34),$Q_\pi(s,a)=\mathbb{E}_\pi[\sum_{t=0}^{H}\gamma^t r_t|s_0=s,a_0=a]$(35);其中,$s$为驾驶状态,$a$为状态$s$下采用策略$\pi$选取的驾驶动作,$r=R(s,a)$为执行动作$a$后环境反馈的奖励信号,$R(\cdot)$为预定义奖励函数,$\gamma\in(0,1)$为奖励衰减因子,$t$表示驾驶策略与环境交互的时间步,$H$表示一次交互总步长,期间交互产生的行驶数据将作为驾驶策略训练样本,表示为:$\tau_\pi(s)=\{s,a,r_0,s_1,a_1,r_1,\cdots,s_H,a_H,r_H\}$(36)。

[0266] 给定策略空间Π,强化学习的优化目标是找到一个最优策略$\pi^*\in\Pi$,可以实现预期累计奖励最大化:$\pi^*=arg\max_{\pi\in\Pi}\mathbb{E}_\pi[V_\pi(s)]$(37)。然而如果训练期间驾驶策略探索不充分,无法产生足够的交互数据以供训练,将难以实现准确的策略评估。作为替代,这里定义一个策略置信值,用于判断不同情况下驾驶策略π的可靠程度。

[0267] 对于任意策略$\pi\in\Pi$,其状态-动作值应满足以下定义:$P\left(\tilde{Q}_\pi(s,a)\geq Q_{\pi L}(s,a,\mathcal{D})|\mathcal{D}\right)\geq 1-\delta$(38);其中,$Q_{\pi L}(s,a,\mathcal{D})$定义了在状态$s$下采取策略$\pi$的置信值,$\mathcal{D}$为历史收集数据,$\tilde{Q}_\pi(s,a)$为策略$\pi$的真实状态-动作值,$\delta\in(0,1)$为一个概率值,用于约束$\tilde{Q}_\pi(s,a)$大于$Q_{\pi L}(s,a,\mathcal{D})$的概率。在上式中,$Q_{\pi L}(s,a,\mathcal{D})$可以理解为驾驶策略的性能下界,$Q_{\pi L}$值越大,表示策略有信心实现良好的性能,即不太可能导致意外风险,相反,$Q_{\pi L}$越小则意味着无法根据历史驾驶数据学习到一个足够可靠的驾驶策略来应对驾驶状态,表明存在潜在的驾驶风险情况。

[0268] 相应地,可以定义一个与策略置信值相反的驾驶风险概率值来判断不同驾驶情况

的风险程度，则有：$L_\pi(s,\mathcal{D}) = Q_{\pi L}(s,a,\mathcal{D})$(39)，当$Q_{\pi L}$值越大，意味着风险存在概率较低，因此将对一个较小的L_π值，反之亦然。

[0269] 给定驾驶策略π与环境转移模型(即虚拟驾驶模型)T，车辆任意t时刻的状态满足$s_{t+1} = T\left(s_t, \pi(s_t)\right)$，假设车辆当前状态为$s$，其一次交互过程中产生的行驶轨迹为$\tau_\pi(s)$，该轨迹的累积驾驶奖励可写为：$G\left(\tau_\pi(s)\right) = \sum_{t=0}^{H} \gamma^t r_t, r_t \in \tau_\pi(s), s_0 = s$(40)。

[0270] 重复交互过程收集行驶数据，基于k条交互轨迹可以构建数据集，对于任意第i条交互轨迹，定义数据单元$d_i(s,a) \coloneqq \left(s,a,G\left(\tau_\pi^i(s)\right)\right)$，则数据集表示为：

$\mathcal{D}_k(s,a) = \{d_i(s,a)\}_{i=1,\cdots,k}$(41)。

[0271] 然后通过以下方式估计平均状态-动作值：

$Q_\pi(s,a,T) \leftarrow \bar{G}(s,a,T) = \frac{1}{k}\sum_{G\left(\tau_\pi^i(s)\right)\in\mathcal{D}_k(s,a)} G\left(\tau_\pi^i(s)\right)$(42)。其中，$Q_\pi(s,a,T)$为真实策略性能评估值，$\bar{G}(s,a,T)$是$Q_\pi(s,a,T)$的点估计。通过采集大量样本，$\bar{G}(s,a,T)$的估计值将逐渐收敛到$Q_\pi(s,a,T)$。

[0272] 然而，由于真实数据收集难度大，为了确保足够的训练数据量，这里将利用采集到的行驶数据对环境转移模型T进行参数化学习，得到一个虚拟环境模型$\hat{T}$，然后基于$\hat{T}$产生大量虚拟数据，能够解决因真实数据不足引起的估计误差过大问题，从而降低Q_π值估计误差。这里采用高斯神经网络来构建环境转移模型：$T(s_{t+1}|s_t,a_t) = \mathcal{N}\left(\mu(s_t,a_t),\sigma(s_t,a_t)\right)$(43)；其中，$\mu$和$\sigma$分别表示高斯分布$\mathcal{N}$的方差和均值。因此，虚拟环境模型$\hat{T}$可以通过从真实轨迹数据$\tau_\pi \in \mathcal{D}_k$中抽取数据单元$\{s_t,a_t,s_{t+1}\}$进行监督训练，表示为：$\hat{T} \leftarrow f(\mathcal{D}_k)$(44)，$f(\cdot)$表示高斯神经网络的监督训练过程。基于真实驾驶数据训练一个虚拟环境转移模型，然后生成虚拟行驶数据，进行驾驶策略性能估计，能够解决因真实数据不足引起的估计误差过大问题。

[0273] 得到虚拟环境模型$\hat{T}$后，可以根据策略π生成虚拟行驶数据：$\hat{t}_\pi(s) = \{s,a_0,\hat{r}_0,\hat{s}_1,\hat{a}_1,\hat{r}_1,\cdots,\hat{s}_H,\hat{a}_H,\hat{r}_H\}$(45)；其中，$\hat{s}$、$\hat{a}$和$\hat{r}$分别表示虚拟生成的输入状态、策略动作和动作奖励。进一步可基于收集到的虚拟数据集$\hat{\mathcal{D}}_k(s,a)$按照式(42)估计策略性能评估值$\hat{Q}_\pi\left(s,a,\hat{T}\right)$。

[0274] 由于基于有限真实驾驶数据学习得到的虚拟环境模型与真实环境模型之间仍存在偏差，这可能导致策略性能估计不准确。为了量化性能估计的置信度，定义分布$P\left(\tilde{Q}_\pi(s,a)|\mathcal{D}_k\right)$来描述给定当前真实数据集$\mathcal{D}_k$下策略$\pi$的真实值概率，即值估计的置信度。直观地说，当数据量足够大，可以进行高置信度的值估计时，分布将集中在真值$\tilde{Q}_\pi(s,a)$附近。反之，在性能估计没有足够置信度的数据稀疏情况下，分布则会更分散。由此可结合概率约束对驾驶性能估计结果的置信度进行量化，降低估计偏差的影响。

[0275] 考虑到当虚拟环境模型$\hat{T}$接近真实的环境模型时,所估计的策略性能也将接近真实性能,因此估计驾驶性能置信度的任务可以转换为估计环境模型的置信度:$P\left(\tilde{Q}_{\pi}(s,a)|\mathcal{D}_k\right)\sim P\left(\hat{T}(s,a)|\mathcal{D}_k\right)$(46)。

[0276] 根据公式(38)对于置信值的定义,置信值的估计可以通过采用概率约束δ对估计分布$P\left(\tilde{Q}_{\pi}(s,a)|\mathcal{D}\right)$进行截断来实现,结合公式(39)中的任务等价,可以考虑基于环境模型概率分布$P(T(s,a)|\mathcal{D})$来实现。假设存在一组环境转移模型$\mathcal{T}$,其中包含真实状态转移$\tilde{T}$的概率大于$1-\delta$,则有:$P\left(\tilde{T}\in\mathcal{T}|\mathcal{D}\right)\geq 1-\delta$(47),那么则置信值可通过以下方式计算:$Q_{\pi L}(s,a,\mathcal{D}_k)=\min_{T\in\mathcal{T}}\hat{Q}_{\pi}(s,a,T)$(48)。从而确定每一策略的安全下边界。其中,$\hat{Q}_{\pi}(s,a,T)$为按照式(42)计算的策略性能估计值(即驾驶策略的置信值)。

[0277] 请参见图8,驾驶策略的置信值的计算过程包括以下步骤:给定驾驶策略π,基于策略π执行环境交互,收集历史驾驶数据$\mathcal{D}_k$;初始化多组参数化环境转移模型$\mathcal{T}=\{\hat{T}_1,\hat{T}_2,\cdots,\hat{T}_n\}$;从数据集$\mathcal{D}_k$中随机采样子数据集$\{\mathcal{D}_{k,1},\mathcal{D}_{k,2},\cdots,\mathcal{D}_{k,n}\}$;分别基于每个子数据集独立训练环境转移模型:$\hat{T}_i\leftarrow f\left(\mathcal{D}_{k,i}\right)$(49);基于任意环境转移模型$\hat{T}_i$,生成虚拟轨迹数据$\hat{\mathcal{D}}_{k,i}$,根据式(42)估计驾驶策略性能值$\hat{Q}_{\pi,i}$;根据式(48)计算驾驶策略的置信值$Q_{\pi L}$。

[0278] 基于策略置信值选取驾驶风险最小的策略生成车辆执行动作,确保驾驶性能始终是最优的。针对具有不同驾驶偏好的候选策略,通过采集驾驶数据来估计策略置信值,然后选择其中风险程度最小的策略作为车辆将要执行的驾驶策略,确保车辆在面对长尾场景时依然能够进行可靠决策。可信动作$a_{op}=arg\min_{\pi\in\Pi}L_{\pi}(s,\mathcal{D})=arg\,max_{\pi\in\Pi}Q_{\pi L}(s,a,\mathcal{D})=arg\max_{\pi\in\Pi}\min_{T\in\mathcal{T}}\hat{Q}_{\pi}(s,a,T)$(50)。因此,驾驶期间车辆动作生成过程可表示为如下步骤:获取车辆驾驶状态s;对于任意候选策略$\pi\in\Pi$,计算置信值$Q_{\pi L}(s,a,\mathcal{D})$;生成动态可信动作$a_{op}$;车辆执行动作$a_{op}$,转移至下一驾驶状态$s'$。

[0279] 本发明为了确保驾驶过程的可靠性,基于差异化奖励权项构建多个具有不同驾驶偏好的候选策略,采用基于动态置信估计的性能评估方法衡量驾驶策略的风险程度,通过选择风险程度最小的策略生成可信驾驶动作,确保车辆驾驶期间始终执行最优驾驶策略,提升长尾情况下的驾驶性能稳定性。

[0280] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其它实施例的不同之处,各个实施例之间相同或相似部分互相参见即可。

[0281] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块,或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的可读存储介质中。

[0282] 本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据

本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

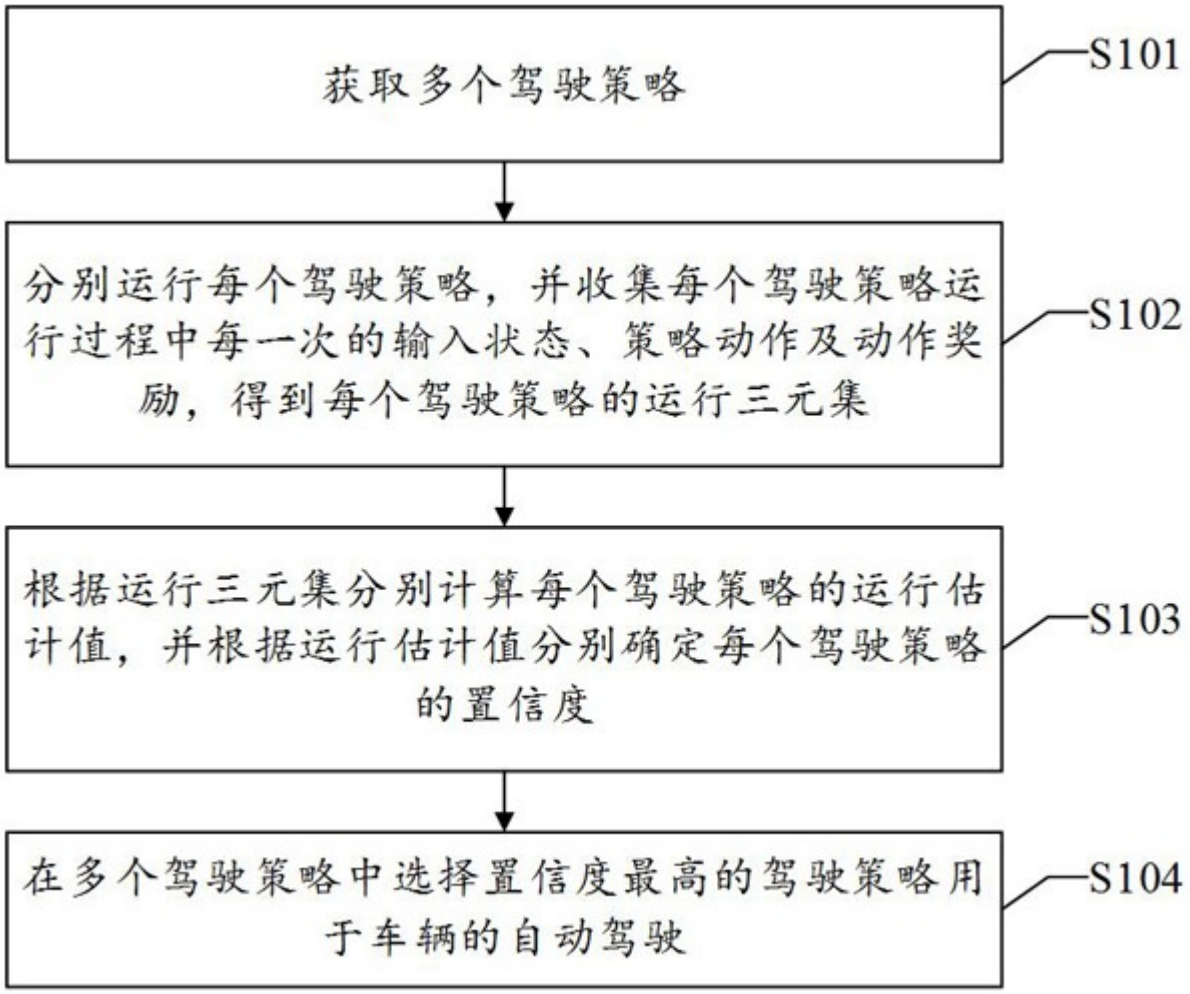

获取多个驾驶策略
S101
分别运行每个驾驶策略，并收集每个驾驶策略运行过程中每一次的输入状态、策略动作及动作奖励，得到每个驾驶策略的运行三元集
S102
根据运行三元集分别计算每个驾驶策略的运行估计值，并根据运行估计值分别确定每个驾驶策略的置信度
S103
在多个驾驶策略中选择置信度最高的驾驶策略用于车辆的自动驾驶
S104

图 1

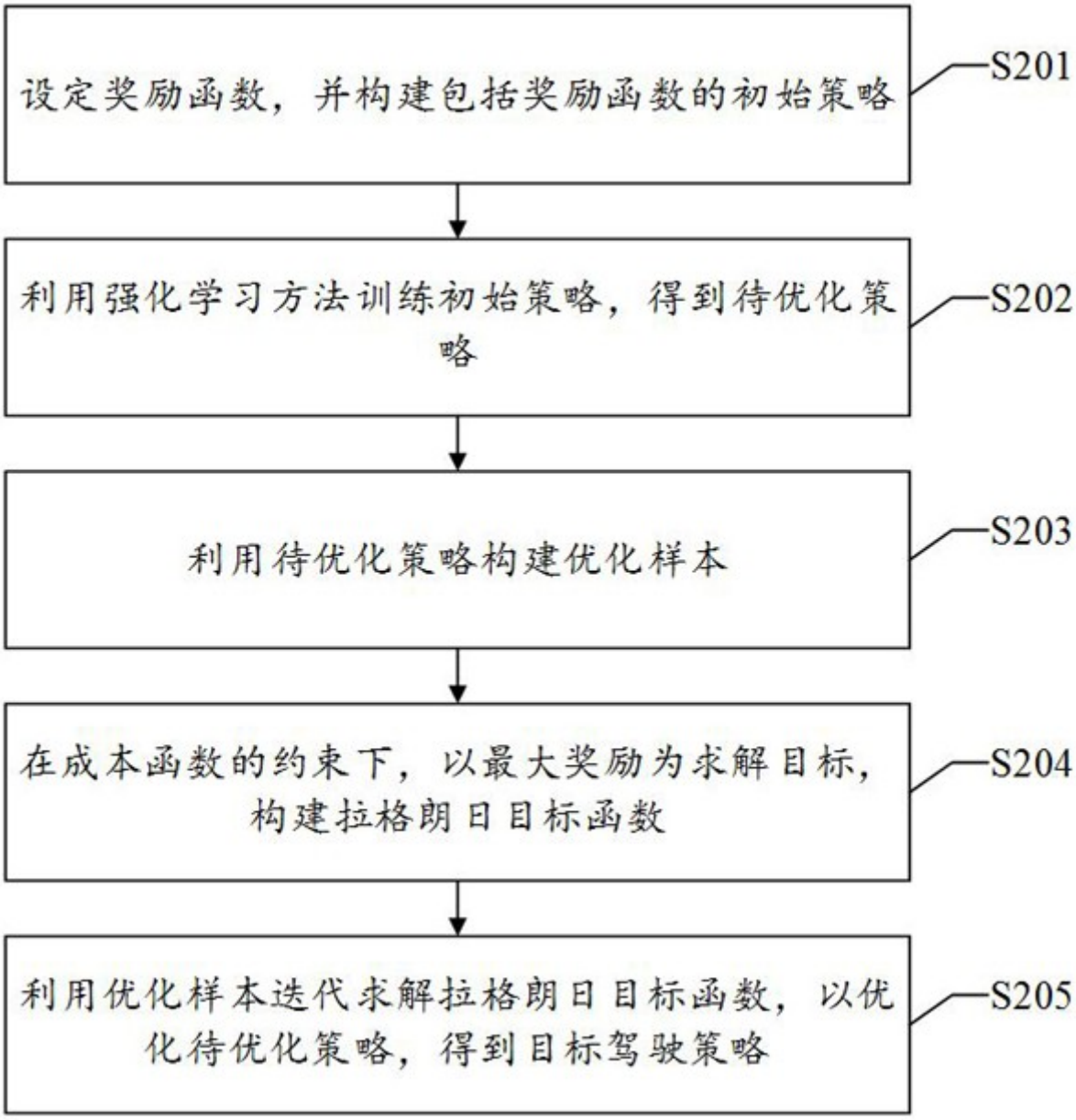

设定奖励函数，并构建包括奖励函数的初始策略
S201
利用强化学习方法训练初始策略，得到待优化策略
S202
利用待优化策略构建优化样本
S203
在成本函数的约束下，以最大奖励为求解目标，构建拉格朗日目标函数
S204
利用优化样本迭代求解拉格朗日目标函数，以优化待优化策略，得到目标驾驶策略
S205

图 2

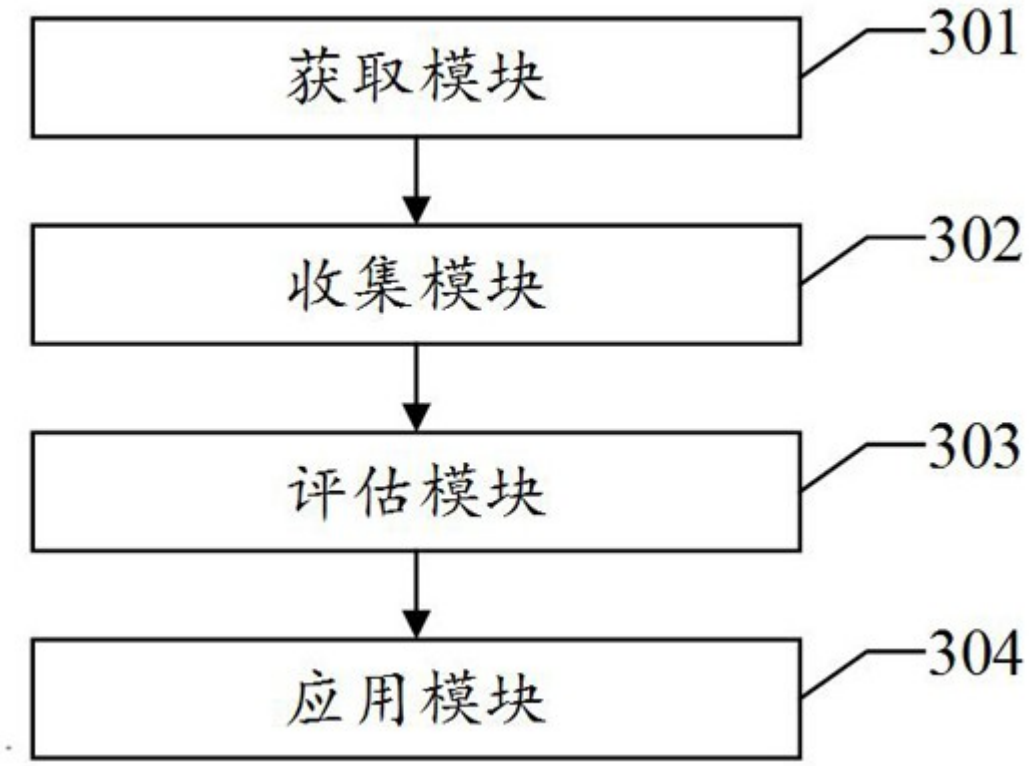


图 3

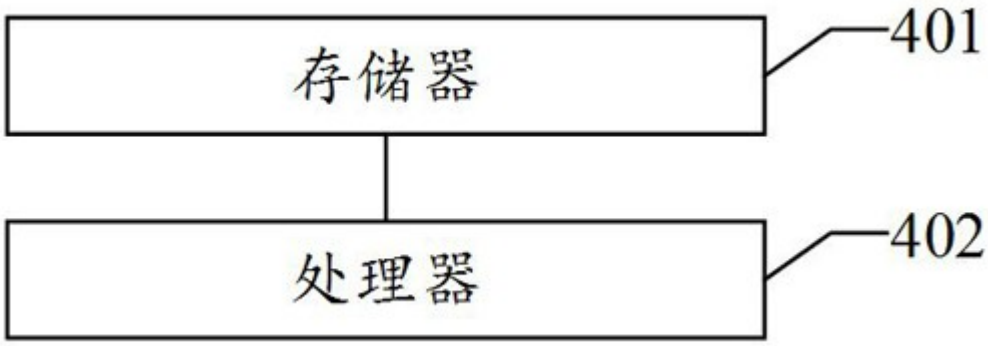


图 4

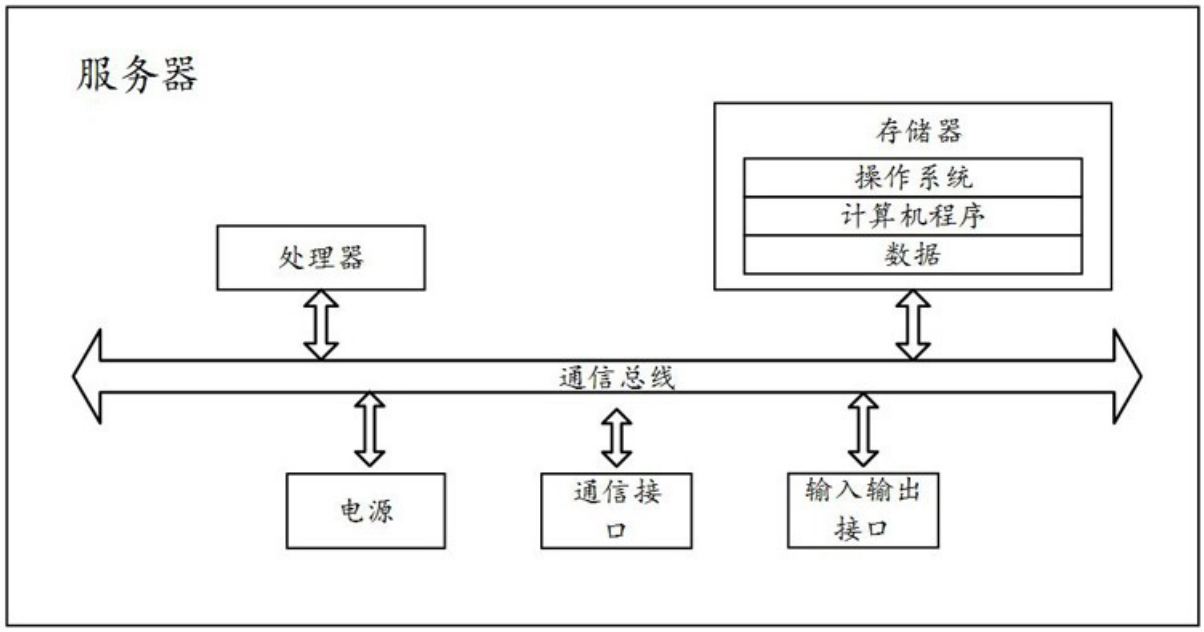

服务器
存储器
操作系统
计算机程序
数据
处理器
通信总线
电源
通信接口
输入输出接口

图 5

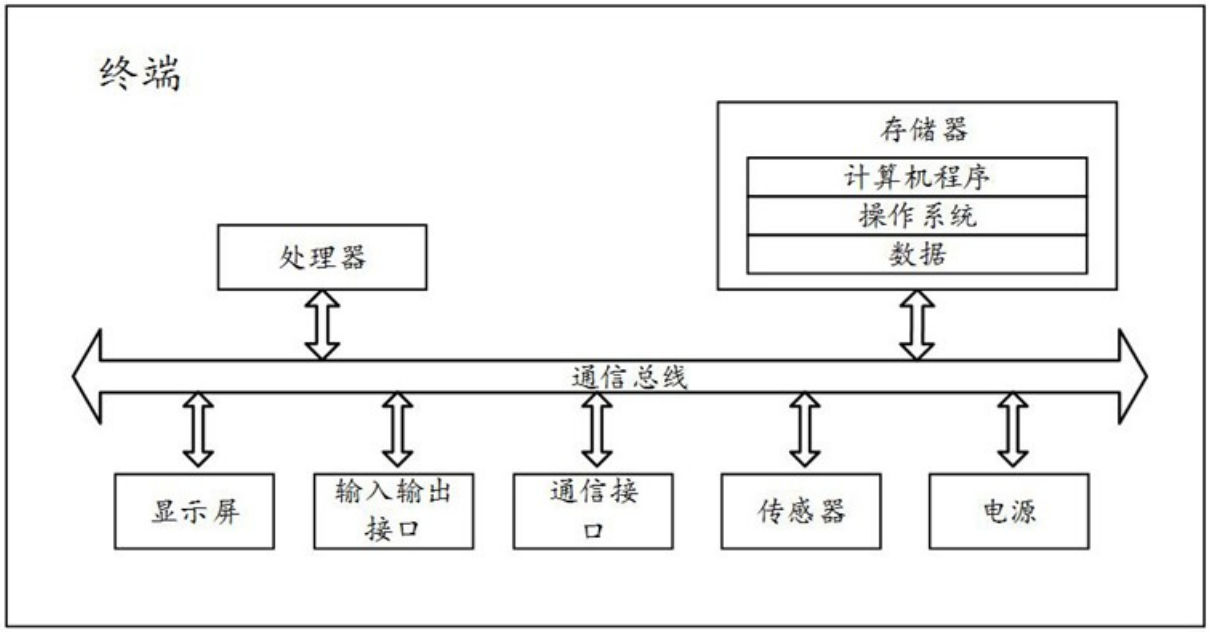

终端
存储器
计算机程序
操作系统
数据
处理器
通信总线
显示屏
输入输出接口
通信接口
传感器
电源

图 6

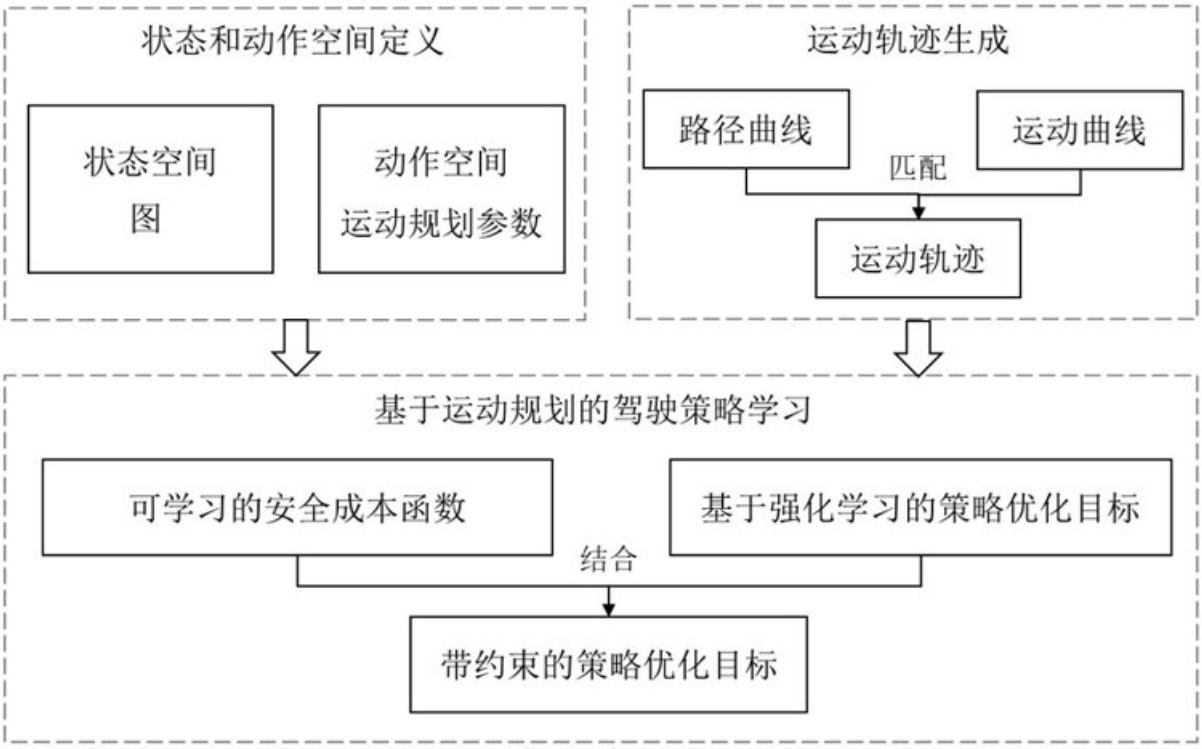

状态和动作空间定义
状态空间图
动作空间运动规划参数
运动轨迹生成
路径曲线
运动曲线
匹配
运动轨迹
基于运动规划的驾驶策略学习
可学习的安全成本函数
基于强化学习的策略优化目标
结合
带约束的策略优化目标

图 7

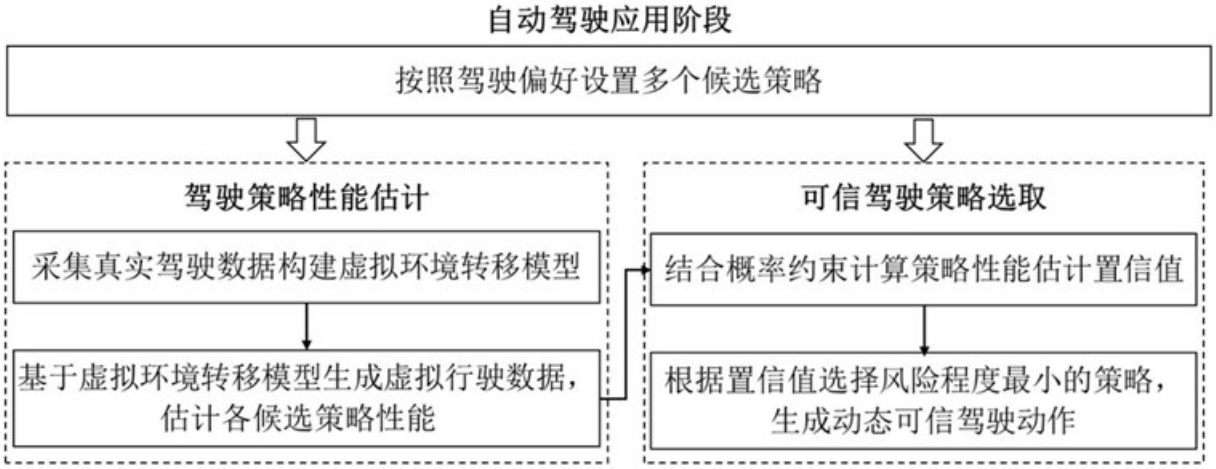

自动驾驶应用阶段
按照驾驶偏好设置多个候选策略
驾驶策略性能估计
采集真实驾驶数据构建虚拟环境转移模型
基于虚拟环境转移模型生成虚拟行驶数据，估计各候选策略性能
可信驾驶策略选取
结合概率约束计算策略性能估计置信值
根据置信值选择风险程度最小的策略，生成动态可信驾驶动作

图 8