(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 112218744 A
(43)申请公布日 2021.01.12

(21)申请号 201980027638.5

(22)申请日 2019.04.22

(30)优先权数据
62/661,055 2018.04.22 US

(85)PCT国际申请进入国家阶段日
2020.10.22

(86)PCT国际申请的申请数据
PCT/US2019/028454 2019.04.22

(87)PCT国际申请的公布数据
WO2019/209681 EN 2019.10.31

(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州

(72)发明人 J.谭 T.张 A.伊森 E.库曼斯 Y.白

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 金玉洁

(51)Int.Cl.
B25J 9/16(2006.01)
B62D 57/02(2006.01)

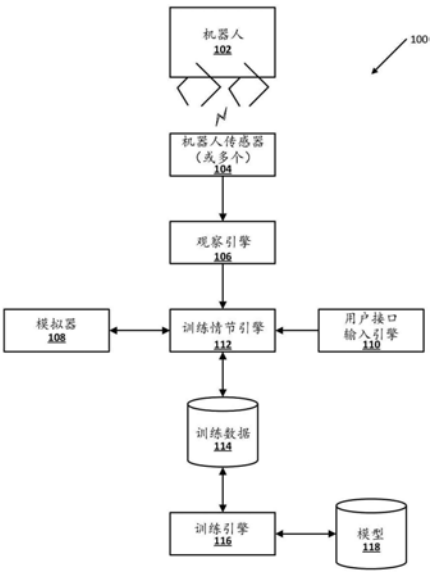
权利要求书2页 说明书11页 附图8页

(54)发明名称

学习多足机器人的敏捷运动的系统和方法

(57)摘要

训练和/或使用机器学习模型用于机器人的运动控制,其中将模型解耦。在许多实现方式中,模型被解耦为开环分量和反馈分量,其中用户可以提供期望的参考轨迹(例如,对称正弦曲线)作为开环分量的输入。在附加的和/或可替代的实现方式中,模型被解耦为模式生成器分量和反馈分量,其中用户可以提供受控参数(或多个)作为模式生成器分量的输入,以生成模式生成器相位数据(例如,非对称正弦曲线)。神经网络模型可用于生成机器人控制参数。



CN 112218744 A

1.一种由一个或多个处理器实现的方法,包括:

接收传感器数据的实例,传感器数据的实例是基于来自机器人的一个或多个传感器的输出生成的,

其中,传感器数据的实例基于在使用由利用神经网络模型生成的机器人控制参数的先前实例对机器人进行控制之后的机器人的状态,神经网络模型表示对机器人运动任务的学习策略,并且事先通过强化学习进行了训练;

接收用于机器人运动任务的参考轨迹,其中,参考轨迹与传感器数据解耦并且受到经由一个或多个用户接口输入设备的用户交互的影响;

根据传感器数据实例生成观察;

将观察和参考轨迹应用于神经网络模型,以生成机器人控制参数的当前实例;以及

基于机器人控制参数的当前实例来控制机器人的一个或多个致动器。

2.根据权利要求1所述的方法,其中,机器人是包括多条腿的行走机器人,并且其中,机器人控制参数的当前实例为多条腿中的每条腿定义腿的期望姿势。

3.根据权利要求1所述的方法,其中,生成的观察指示马达角度、机器人的基座的滚动、机器人的基座的俯仰以及机器人的角速度。

4.根据权利要求3所述的方法,其中,生成的观察仅指示马达角度、滚动、俯仰和机器人的角速度。

5.根据权利要求3所述的方法,其中,生成的观察排除由传感器数据的实例指示的一个或多个可用观察。

6.根据权利要求5所述的方法,其中,排除的一个或多个可用观察包括机器人的基座的偏航。

7.根据权利要求1所述的方法,还包括:

使用模拟机器人并使用强化学习在模拟器中训练神经网络模型。

8.根据权利要求7所述的方法,其中,在模拟中训练神经网络模型包括:对于模拟机器人,对模拟延迟进行建模,模拟延迟在捕获来自机器人的一个或多个传感器的输出的时间与基于机器人控制参数的当前实例控制机器人的一个或多个致动器的时间之间。

9.根据权利要求7所述的方法,其中,使用强化学习在模拟中训练神经网络模型包括:

在强化学习期间利用奖励函数,其中,利用的奖励函数惩罚高机器人能量消耗。

10.根据权利要求9所述的方法,其中,利用的奖励函数还鼓励更快的向前机器人速度。

11.根据权利要求1所述的方法,其中,参考轨迹指示机器人步态和机器人高度。

12.根据权利要求1所述的方法,其中,参考轨迹包括对称正弦函数。

13.根据权利要求1所述的方法,其中,一个或多个传感器是一个或多个马达编码器和一个或多个惯性测量单元。

14.一种由一个或多个处理器实现的方法,包括:

接收传感器数据的实例,传感器数据的实例是基于来自机器人的传感器组件的一个或多个传感器的输出而生成的;

基于传感器数据的实例生成观察,以基于神经网络模型执行机器人动作,神经网络模型表示对机器人的运动任务的学习策略,其中,神经网络模型被解耦为模式生成器分量和神经网络反馈分量;

基于与用户接口输入设备的用户交互来接收受控参数，其中，在训练表示强化学习策略的神经网络模型之后，用户可以在用户接口输入设备处改变受控参数；

将受控参数应用于模式生成器分量，以生成模式生成器相位数据；

将观察、受控参数和模式生成器相位数据应用于神经网络反馈分量，以生成机器人控制参数；以及

基于机器人控制参数来控制机器人的一个或多个致动器。

15.根据权利要求14所述的方法，其中，机器人是四足机器人。

16.根据权利要求15所述的方法，其中，从由步态、运动速度和运动高度组成的组中选择受控参数。

17.根据权利要求15所述的方法，其中，生成模式生成器相位包括生成不对称正弦曲线。

18.根据权利要求15所述的方法，其中，非对称正弦曲线包括摆动相和支撑相，其中，摆动相指示四足机器人的一条或多条腿离开地，并且支撑相指示四足机器人的一条或多条腿在地上。

19.根据权利要求18所述的方法，其中，受控参数改变不对称正弦曲线。

20.根据权利要求14所述的方法，其中，一个或多个传感器是一个或多个马达编码器和一个或多个惯性测量单元。

学习多足机器人的敏捷运动的系统和方法

背景技术

[0001] 许多机器人被编程为执行某些任务。例如,可以对装配线上的机器人进行编程以识别特定对象,并对这些对象执行具体操纵。

[0002] 此外,可以对行走机器人(legged robot)进行编程,以穿越复杂地形。行走机器人(例如,具有两条或更多条腿的多足机器人)可以基于不同的行走表面调整步态、运动速度、脚位和/或离地距离。例如,双足机器人(即,具有两条腿的机器人)可以像人一样直立行走以穿越各种地形。附加地或可替代地,四足机器人(即,具有四条腿的机器人)可以用四个肢体穿越表面,并且在某些情况下可以模仿诸如马、狗和/或灵长类动物的多种动物的运动。然而,训练行走机器人在不同表面上行走的任务可能是非常复杂的任务。物理机器人本身可以被训练走路,但是这可能特别耗时,并且在某些情况下无效。作为替代方案,可以训练物理行走机器人的模拟以穿越地形。然而,将行走机器人的训练后的模拟转换为行走机器人的运动可能会带来一系列挑战。

发明内容

[0003] 本文公开的实现方式利用深度强化学习来训练模型(例如,深度神经网络模型),该模型可以被用于确定诸如四足机器人或其他多足机器人的行走机器人的运动。附加地或可替代地,实现方式可以涉及在控制多足机器人中使用这种模型。在那些实现方式中的一些中,机器人运动(locomotion)可以部分地由运动控制器确定,该运动控制器可以被解耦为开环分量和反馈分量。开环分量可以接收基于用户输入的信息,该信息被用于训练机器人运动。接收的信息可以包括例如直接或间接定义机器人步态、机器人高度和/或其他控制参数(或多个)的信息。在一些实现方式中,提供给开环分量的信息可以是基于用户输入生成的参考轨迹(例如,正弦波)的形式。反馈回路分量可以填充信息的缺失部分,例如关节角度、基本方向、角速度和/或信息的其他缺失部分的控制,使得行走机器人仍可以相对于用户提供到开环分量的输入行走。例如,可以使用强化学习来训练反馈回路分量,以确定对于用户提供的参考轨迹的平衡控制(这对于手动设计而言可能是乏味的)。附加地或可替代地,可以减少观察空间,这可以使训练模拟更容易转换为现实世界。在一些实现方式中,可以通过减少在训练行走机器人进行运动任务中使用的传感器数据量来减少观察空间。

[0004] 在一些附加或可替代的实现方式中,机器人运动可以由控制策略控制器(例如运动策略控制器)生成,其可以被解耦为开环模式生成器和神经网络反馈分量。用户可以通过提供开环模式生成器控制参数(或多个)——诸如腿步态、腿高和/或其他控制参数(或多个)——在训练机器人运动中控制开环模式生成器。神经网络反馈分量可以填充信息的缺失部分(例如,对关节角度、基本方向、角速度的控制),使得行走机器人相对于由开环模式生成器提供的相位信息和用户提供的控制参数仍然可以行走。在一些实现方式中,开环模式生成器可以创建不对称正弦曲线以提供给神经网络反馈分量。非对称正弦曲线可以包括摆动相和支撑相。摆动相通常表示一条或多条机械腿离开地面,而支撑相通常表示一条或多条机械腿位于地面上。在各种实现方式中,即使在已经训练神经网络反馈控制器之后,用

户也可以提供机器人参数以动态地改变机器人的运动行为。例如，用户可以通过改变用户提供的控制参数来在训练之后动态改变机器人的速度或步态。

[0005] 在一些实现方式中，提供一种由一个或多个处理器实现的方法，所述方法包括：接收传感器数据的实例，传感器数据的实例是基于来自机器人的一个或多个传感器的输出生成的，其中，传感器数据的实例基于在使用由神经网络模型生成的机器人控制参数的先前实例对机器人进行控制之后的机器人的状态，神经网络模型表示对机器人运动任务的学习策略并且事先通过强化学习进行了训练。所述方法还包括接收用于机器人运动任务的参考轨迹，其中，参考轨迹与传感器数据解耦并且受到经由一个或多个用户接口输入设备的用户交互的影响。所述方法还包括：根据传感器数据的实例生成观察；以及将观察和参考轨迹应用于神经网络模型，以生成机器人控制参数的当前实例。所述方法还包括基于机器人控制参数的当前实例来控制机器人的一个或多个致动器。

[0006] 这些和其他实现方式可以包括以下特征中的一个或多个。

[0007] 在一些实现方式中，机器人是包括多条腿的行走机器人，并且其中，机器人控制参数的当前实例为多条腿中的每条腿定义了腿的期望姿势。

[0008] 在一些实现方式中，生成的观察指示马达角度、机器人的基座的滚动、机器人的基座的俯仰以及机器人的角速度。

[0009] 在一些实现方式中，生成的观察仅指示机器人的马达角度、滚动、俯仰和角速度。

[0010] 在一些实现方式中，生成的观察排除由传感器数据的实例指示的一个或多个可用观察。

[0011] 在一些实现方式中，排除的一个或多个可用观察包括机器人的基座的偏航。

[0012] 在一些实现方式中，所述方法还包括：使用模拟机器人并使用强化学习在模拟器中训练神经网络模型。

[0013] 在一些实现方式中，在模拟中训练神经网络模型包括：对于模拟机器人，对模拟延迟进行建模，所述模拟延迟在捕获来自机器人的一个或多个传感器的输出的时间与基于机器人控制参数的当前实例控制机器人的一个或多个致动器的时间之间。

[0014] 在一些实现方式中，使用强化学习在模拟中训练神经网络模型包括：在强化学习期间利用奖励函数，其中，利用的奖励函数惩罚高机器人能量消耗。

[0015] 在一些实现方式中，其中，利用的奖励函数还鼓励更快的向前机器人速度。

[0016] 在一些实现方式中，机器人是四足机器人。

[0017] 在一些实现方式中，参考轨迹指示机器人步态和机器人高度。

[0018] 在一些实现方式中，参考轨迹包括对称正弦函数。

[0019] 在一些实现方式中，运动任务是慢跑。

[0020] 在一些实现方式中，运动任务是飞驰。

[0021] 在一些实现方式中，一个或多个传感器是一个或多个马达编码器和一个或多个惯性测量单元。

[0022] 在一些实现方式中，提供一种由一个或多个处理器实现的方法，并且所述方法包括：接收传感器数据的实例，传感器数据的实例是基于来自机器人的传感器组件的一个或多个传感器的输出而生成的。所述方法还包括基于传感器数据的实例生成观察，以基于神经网络模型执行机器人动作，神经网络模型表示对机器人的运动任务的学习策略，其中，神

经网络模型被解耦为模式生成器分量和神经网络反馈分量。所述方法还包括基于与用户接口输入设备的用户交互来接收受控参数，其中，在训练了表示强化学习策略的神经网络模型之后，用户可以在用户接口输入设备上改变受控参数。所述方法还包括将受控参数应用于模式生成器分量，以生成模式生成器相位数据。所述方法还包括将观察、控制参数和模式生成器相位数据应用于神经网络反馈分量，以生成机器人控制参数。所述方法还包括基于机器人控制参数来控制机器人的一个或多个致动器。

[0023] 在一些实现方式中，机器人是四足机器人。

[0024] 在一些实现方式中，从由步态、运动速度和运动高度组成的组中选择受控参数。

[0025] 在一些实现方式中，生成模式生成器相位包括生成参数化不对称正弦曲线。

[0026] 在一些实现方式中，非对称正弦曲线包括摆动相和支撑相，其中，摆动相指示四足机器人的一条或多条腿离开地，并且支撑相指示四足机器人的一条或多条腿在地上。

[0027] 在一些实现方式中，受控参数改变不对称正弦曲线。

[0028] 在一些实现方式中，运动任务是慢跑。

[0029] 在一些实现方式中，运动任务是飞驰。

[0030] 在一些实现方式中，一个或多个传感器是一个或多个马达编码器和一个或多个惯性测量单元。

[0031] 其他实现方式可以包括非暂时性计算机可读存储介质，其存储可由一个或多个处理器(例如，一个或多个中央处理单元)执行的指令，以执行诸如上述和/或本文其他地方中的一个或多个的方法。另一实现方式可以包括一个或多个计算机和/或一个或多个机器人的系统，该系统包括一个或多个处理器，可操作以执行存储的指令以执行诸如上述和/或本文其他地方所述的一个或多个方法的方法。

[0032] 应当理解，本文中更详细描述的前述概念和附加概念的所有组合被认为是本文公开的主题的一部分。例如，出现在本公开的结尾处的要求保护的主题的所有组合被认为是本文公开的主题的一部分。

附图说明

[0033] 图1示出可以实现各种实现方式的示例环境。

[0034] 图2示出根据本文公开的实现方式的示例神经网络模型。

[0035] 图3是示出根据本文公开的实现方式的基于使用神经网络模型生成的机器人控制参数来控制机器人的致动器的示例处理的流程图。

[0036] 图4是示出根据本文公开的实现方式的训练神经网络模型的示例处理的另一流程图。

[0037] 图5示出根据本文公开的实现方式的神经网络模型的另一示例。

[0038] 图6是示出根据本文公开的实现方式的基于使用神经网络模型生成的机器人控制参数来控制机器人的致动器的另一示例处理的另一流程图。

[0039] 图7是示出根据本文公开的实现方式的训练神经网络模型的另一示例处理的另一流程图。

[0040] 图8示意性地描绘了机器人的示例架构。

[0041] 图9示意性地描绘了计算机系统的示例架构。

具体实施方式

[0042] 下面公开的各种实现方式涉及在行走机器人(诸如四足机器人)的运动中训练和/或利用机器学习模型(例如,神经网络模型)。在本文公开的一些实现方式中,利用强化学习来训练机器学习模型,并且在训练时,机器学习模型表示用于生成可以用于驱动行走机器人的致动器以控制行走机器人的运动的控制参数的策略。在那些实现方式的一些版本中,利用这样的模型在给定时间生成的控制参数可以基于对运动控制器的开环分量的用户输入。附加地或可替代地,控制参数可以基于用户提供给控制策略控制器的模式生成器分量的受控参数。

[0043] 可以通过强化学习在许多实现方式中学习机器人运动任务。强化学习的目的是控制试图最大化奖励函数的代理,在机器人技能(在本文中也称为任务)的背景下,该奖励函数表示用户提供的机器人应尝试完成的内容的定义。在时间t的状态x_t,代理根据其策略$\pi(x_t)$选择并执行操作u_t,根据机器人$p(x_t,u_t,x_{t+1})$的动态转换到新状态x_{t+1},并接收奖励$r(x_t,u_t)$。强化学习的目标是找到最佳策略π^*,该策略使初始状态分布的预期奖励之和最大化。奖励是基于奖励函数确定的,如上所述,该奖励函数取决于要完成的机器人任务。因此,机器人情境中的强化学习试图学习用于执行给定机器人任务(例如,机器人运动)的最佳策略。

[0044] 转到附图,图1示出可以实现本文描述的实现方式的示例环境100。图1包括示例机器人102、观察引擎106、机器人模拟器108、用户接口输入引擎120、训练情节引擎112和训练引擎116。还包括机器人传感器(或多个)104、训练数据114和一个或多个机器学习模型118。

[0045] 机器人102是具有多个自由度的行走机器人,多个自由度通过控制机器人102的腿的致动器(或多个)使能机器人运动。例如,机器人102可以是四足机器人(即,四腿机器人),其中每条腿由两个致动器控制,允许腿在矢状面(sagittal plane)中移动。例如,对应腿的第一致动器可以在机器人102的腿与身体之间的附接点处,并且对应腿的第二致动器可以在该附接点与对应腿的远端之间(例如,在腿的“膝盖”处)。可以使用功率宽度调制(PWM)信号通过位置控制来致动马达。在一些实现方式中,可以使用其他数量的马达和/或其他马达控制方法(除PWM外)。

[0046] 在各种实现方式中,机器人102配备有各种机器人传感器104,诸如测量马达角度的马达编码器、测量机器人基座的方向和角速度的惯性测量单元(IMU)和/或测量机器人的位置的附加传感器。尽管在图1中示出了具体机器人102,但是可以利用附加和/或可替代的机器人,包括具有更多腿的机器人(例如,五腿机器人、六腿机器人、八腿机器人和/或具有附加腿的机器人)、具有更少腿的机器人(例如,三腿机器人、两腿机器人)、具有机械臂的机器人、具有人形的机器人、具有动物形的机器人、除机器人腿外还包括一个或多个轮子的机器人等。

[0047] 基于来自现实世界物理机器人的数据训练机器学习模型可能很耗时(例如,实际穿越大量路径需要大量时间),可能会消耗大量资源(例如,操作机器人所需的电力)和/或可能导致所使用的机器人磨损。鉴于这些和/或其他考虑,可以将机器人模拟器(诸如机器人模拟器108)用于生成可以在机器学习模型(诸如模型118)的训练中使用的模拟训练数据。然而,在真实机器人和真实环境与由机器人模拟器模拟的模拟机器人和/或模拟环境之间通常存在有意义的“现实差距”。

[0048] 在许多实现方式中，可以通过适配模拟器(例如，模拟器108)和/或使用模拟器生成的数据(例如，模拟机器人、模拟环境和/或使用模拟器生成的附加数据)来减小现实差距。例如，可以将在模拟器108中使用的致动器模型设计为更准确地模拟机器人致动器。在各种实现方式中，模拟致动器(或多个)的这种提高的精度可以减小现实差距。

[0049] 例如，当使用传统方法模拟致动器(或多个)时，会发现较大的现实差距。例如，为每个马达制定一个约束$e_{n+1}=0$，其中，e_{n+1}是当前时间步长结束时的误差。误差可以被定义为：

[0050] $$e_{n+1} = k_p\left(\underline{q} - q_{n+1}\right) + k_d(\underline{\dot{q}} - \dot{q}_{n+1}) \quad (1)$$

[0051] 其中，$\underline{q}$和$\underline{\dot{q}}$是期望的马达角度和速度，q_{n+1}和$\dot{q}_{n+1}$是当前时间步长结束时的马达角度和速度，k_p是比例增益，且k_d是微分增益。公式(1)确保未来(即在时间步长结束时)的马达角度和速度满足误差约束e_{n+1}。如果使用大增益，则提高模拟中的马达稳定性，但实际上马达可能会振荡。

[0052] 为了消除致动器的这种差异，许多实现方式使用根据理想DC马达的动力学特性的致动器模型。给定PWM信号，马达的扭矩可以被表示为：

[0053] $\tau = K_t I$ (2)

[0054] $$I = \frac{V*PWM - V_{emf}}{R} \quad (3)$$

[0055] $$V_{emf} = K_t\dot{q} \quad (4)$$

[0056] 其中，I是电枢电流，K_t是扭矩常数或反电动势(EMF)常数，V是提供的电压，V_{emf}是反EMF电压，且R是电枢电阻。参数K_t和R可以由特定致动器确定。训练控制器时利用公式(2)-(4)表示的马达模型，机器人经常下沉，无法抬起脚，而同一个控制器在模拟中效果很好，因为线性扭矩-电流关系仅适用于理想马达。实际上，扭矩随着电流的增加而饱和。可以利用逐段线性函数来表征该非线性扭矩-电流关系。在模拟中，一旦从PWM计算出电流(公式(3)和(4))，就可以利用逐段函数来查找对应的扭矩。

[0057] 在位置控制模式下通过经典PD伺服器控制PWM。

[0058] $$PWM = k_p\left(\underline{q} - q_n\right) + k_d(\underline{\dot{q}} - \dot{q}_n) \quad (5)$$

[0059] 附加地或可替代地，目标速度可以被设置为零(即，$\underline{\dot{q}} = 0$)。使用该致动器模型来致动具有期望正弦曲线轨迹的马达符合正确标注(ground truth)。

[0060] 在许多实现方式中，使用模拟器108模拟的延迟可以提供现实差距的附加和/或替代的减少。延迟是导致反馈控制不稳定的原因，可以包括：发送导致机器人的状态改变的马达命令与机器人接收到马达命令之间的时间延迟；机器人接收到马达命令与机器人的状态改变之间的时间延迟；在机器人处捕获到状态改变的传感器测量与报告回控制器之间的时间延迟；和/或其他延迟。马达命令立即生效并且传感器立即报告回状态的机器人模拟器使得反馈控制器在模拟中的稳定性区域远大于其在硬件上的实现。这可能导致在模拟中学习的反馈策略开始振荡、发散，并最终在现实世界中失败。因此，当用于控制真实机器人时，本文公开的延迟模拟技术被用于减轻这些和/或其他缺点，从而导致现实差距的减轻和模型的性能的改善，该模型至少部分地在模拟数据上训练。

[0061] 为了根据模拟器108对延迟进行建模，可以保留观察及其测量时间{(t_i，O_i)$_{i=0,1,\cdots,n-1}$}的历史，其中，$t_i = i\Delta t$，Δt是时间步长。在当前步n，当控制器需要观察时，模拟器108可以在历史中搜索以找到两个相邻的观察O_i和O_{i+1}，其中，$t_i \leq n\Delta t - t_{latency} \leq t_{i+1}$并对其进行线性插值。为了测量物理系统上的延迟，可以发送PWM信号尖峰(例如，对于仅一个时间步长，PWM＝1)，这会导致马达角度突然改变。可以测量发送尖峰与报告结果马达运动之间的时间延迟。

[0062] 观察引擎106可以利用由机器人传感器(或多个)104测量的数据来确定各种观察。观察可以包括机器人基座的滚动、机器人基座的俯仰、机器人基座沿一个或多个轴的角速度(诸如机器人基座沿与滚动相对应的轴的角速度和/或机器人基座沿与俯仰相对应的轴的角速度)、与机器人腿的马达(或多个)对应的马达角度和/或其他观察中的一个或多个。在许多实现方式中，观察空间可以被限制为排除不可靠的测量，包括：具有高噪声水平的测量，诸如马达速度；可能快速漂移的测量，诸如机器人基座的偏航；和/或其他不可靠的测量。保持观察空间紧凑有助于将在模拟中训练的策略传递给真实机器人。

[0063] 用户接口输入引擎110可以捕获各种用户输入，以用于训练机器学习模型118以及控制真实机器人的运动。例如，用户可以提供参考轨迹(如图2所示)、受控参数(如图5所示)和/或其他用户接口输入，以供训练机器学习模型118使用。

[0064] 根据各种实现方式，训练情节引擎112可以被用于生成强化学习训练情节，诸如训练数据114。例如，训练情节引擎112可以使用由模拟器108和用户接口输入引擎110生成的数据来创建训练情节。附加地或可替换地，训练情节引擎112可以利用使用观察引擎106生成的观察来生成训练情节。训练引擎116可以利用由训练情节引擎112生成的训练数据114来训练机器学习模型118。在各种实现方式中，机器学习模型118是神经网络模型，其是解耦网络，并且可以包括卷积神经网络模型、循环网络模型和/或附加类型的神经网络模型。训练引擎116使用强化学习来训练各种实现方式中的机器学习模型118。在图2中示出了示例机器学习模型118。在图5中示出了附加或替代的机器学习模型118。

[0065] 转到图2，框图200示出根据本文描述的实现方式的解耦的机器学习模型(诸如图1的机器学习模型118)。在许多实现方式中，机器学习模型被解耦为开环分量206和反馈分量208。解耦的机器学习模型允许用户对运动策略的训练进行更多控制。开环分量206允许用户提供参考轨迹202(例如，用户提供的对称正弦曲线)以表达例如机器人的期望步态。机器学习模型的反馈分量208基于指示机器人当前状态的观察204(例如使用模拟器108确定的模拟观察和/或使用观察引擎106确定的真实机器人的观察)在参考轨迹202上调整腿部姿势。

[0066] 网络的策略可以表示为：

[0067] $$a(t,o) = \underline{a}(t) + \pi(o) \qquad (6)$$

[0068] 其中，a(t,o)是对于参考轨迹202和观察204的机器学习模型，$\underline{a}(t)$是开环分量206，π(o)是反馈分量208，t是时间并且o是观察204。这代表提供全方位的机器人可控性的混合策略。从完全由用户指定到全部从头开始学习，确定机器人的运动可以有所不同。例如，可以通过将π(o)(即反馈分量)的下限和上限都设置为零来使用用户指定的策略。另外地或可替代地，可以通过设置$\underline{a}(t)=0$(即，将开环分量设置为等于零)并且给反馈分量π(o)

宽的输出范围，策略可以是从头开始学习。在许多实现方式中，可以通过改变开环信号和/或反馈分量的输出界限来确定应用于系统的用户控制量。

[0069] 图3是示出使用解耦的神经机器学习模型生成机器人控制参数以控制机器人运动的示例处理300的流程图。为了方便，参考执行操作的系统来描述处理300的操作。该系统可以包括各种计算机系统的各种组件，诸如图8和/或图9所示的一个或多个组件。此外，尽管以特定顺序示出处理300的操作，但这并不意味着是限制性的。可以重新排序、省略和/或添加一个或多个操作。

[0070] 在块302，系统接收传感器数据的实例。传感器数据可以由各种传感器(例如，马达编码器(或多个)、IMU和/或其他传感器(或多个))捕获，并且基于机器人的状态(例如，马达角度(或多个)、机器人方向、机器人的速度等)。在许多实现方式中，传感器数据基于使用机器人控制参数的先前实例对机器人进行控制之后的机器人状态。

[0071] 在块304，系统使用传感器数据生成观察。例如，系统可以限制观察空间以排除快速漂移和/或通常包含大量噪声的测量。例如，观察空间可以被限制为基座的滚动、基座的俯仰、基座沿滚动轴和俯仰轴的角速度以及机器人腿马达的马达角度(例如，8个马达角度，其中，四足机器人的每条腿包含两个马达)。

[0072] 在块306，系统经由用户接口输入设备(或多个)接收与传感器数据解耦的参考轨迹。参考轨迹可以定义用户指定的期望步态。在许多实现方式中，参考轨迹是对称正弦曲线。

[0073] 在块308，系统通过将观察和参考轨迹应用于训练后的机器学习模型来生成机器人控制参数。机器人控制参数可以指示机器人在下一个状态的期望姿势。在各种实现方式中，机器人控制参数可以指示从当前状态到下一个期望状态的期望变化。

[0074] 在块310，系统基于机器人控制参数来控制机器人的致动器(或多个)。

[0075] 图4是示出训练用于机器人运动的解耦的机器学习模型的示例处理400的流程图。为了方便，参考执行操作的系统来描述处理400的操作。该系统可以包括各种计算机系统的各种组件，诸如图8和/或图9所示的一个或多个组件。此外，尽管以特定顺序示出处理400的操作，但这并不意味着是限制性的。可以重新排序，省略和/或添加一个或多个操作。

[0076] 在块402，系统通过将观察和参考轨迹作为输入应用到机器学习模型的来生成机器人控制参数的实例。在一些实现方式中，使用机器人模拟器(诸如图1的模拟器108)来生成机器人控制参数的实例。

[0077] 在块404，系统基于机器人控制参数的实例来控制机器人的致动器(或多个)。例如，系统可以控制机器人腿的一个或多个马达的马达角度。

[0078] 在块406，系统确定更新后的观察。在各种实现方式中，更新后的观察基于：在块404中系统控制机器人的致动器(或多个)之后的机器人的位置、IMU读数、马达角度(或多个)和/或机器人的其他传感器测量。

[0079] 在块408，系统基于观察、更新后的观察和参考轨迹来确定奖励信号。在各种实现方式中，可以使用鼓励更快的向前奔跑速度和/或惩罚高能量消耗的奖励函数来确定奖励信号。例如，奖励函数可以包括：

[0080] $$r = (p_n - p_{n-1}) \cdot d - \omega \Delta t |\tau_n \cdot \dot{q}_n| \quad (7)$$

[0081] 其中，p_n是当前时间步长的机器人基座的位置，p_{n-1}是先前时间步长的机器人基座

的位置，d是期望的奔跑方向，Δt是时间步长，τ是马达扭矩，且$\dot{q}$是马达速度。第一项测量朝向期望方向的奔跑距离，且第二项测量能量消耗。ω是平衡这两项的权重。

[0082] 在块410，系统使用奖励信号来更新机器学习模型的一个或多个参数。在许多实现方式中，在每个情节处累积奖励。在一些实现方式中，训练情节在满足特定机器人条件之后终止，例如：机器人已进行期望数量的步(例如，训练情节在机器人已进行1000步之后终止)和/或机器人失去平衡(例如，机器人基座相对于地面倾斜超过0.5弧度)。

[0083] 转到图5，框图500示出附加和/或可替代的机器学习模型(诸如图1的机器学习模型118)。在许多实现方式中，机器学习模型被解耦为模式生成器分量504和反馈分量510。在许多实现方式中，用户可以提供受控参数502，诸如期望的运动速度、行走高度和/或其他用户提供的参数，以生成模式生成器相位数据506。换句话说，改变一个或多个受控参数502将改变使用模式生成器分量504生成的模式生成器相位数据506。在许多实现方式中，模式生成器相位数据506提供机器人运动的整体行为的参考(诸如腿的轨迹)，并且可以由不对称正弦曲线表示。

[0084] 可以提供指示机器人当前状态的一个或多个观察508(诸如，使用模拟器108确定的模拟观察和/或使用图1的观察引擎106从真实机器人捕获的观察)作为到解耦的神经网络的反馈分量510的输入。在许多实现方式中，受控参数502和/或模式生成器相位数据506可以附加地或替代地使用反馈分量510作为输入来处理。反馈分量510生成的输出可以与模式生成器相位数据506组合以确定一个或多个机器人控制参数512。

[0085] 图6是示出根据各种实现方式的使用解耦的机器学习模型来生成机器人控制参数的示例处理600的流程图。为了方便，参考执行操作的系统来描述处理600的操作。该系统可以包括各种计算机系统的各种组件，诸如图8和/或图9所示的一个或多个组件。此外，尽管以特定顺序示出处理600的操作，但这并不意味着是限制性的。可以重新排序，省略和/或添加一个或多个操作。

[0086] 在块602，系统接收传感器数据的实例。如上所述，传感器数据可以由各种传感器(例如，马达编码器(或多个)、IMU和/或其他传感器(或多个))捕获，并且基于机器人的状态。在许多实现方式中，传感器数据基于在使用机器人控制参数的先前实例对机器人进行控制之后的机器人状态。

[0087] 在块604，系统基于传感器数据的实例生成观察。例如，系统可以限制观察空间以排除快速漂移和/或通常包含大量噪声的测量。在许多实现方式中，观察空间限于基座的滚动、基座的俯仰、沿着滚动轴和俯仰轴的角速度以及机器人腿马达的马达角度。

[0088] 在块606，系统基于与用户接口输入设备的用户交互来接收受控参数。受控参数可以包括定义机器人的期望步态的一个或多个参数，并且可以包括运动速度、脚位(foot placement)、地面放置和/或附加参数(或多个)。

[0089] 在块608，系统通过将受控参数作为输入应用到训练后的机器学习模型的模式生成器分量来生成模式生成器相位数据。在许多实现方式中，模式生成器相位数据是不对称正弦曲线，表示机器人腿的摆动相和支撑相。

[0090] 在块610，系统通过将(1)观察、(2)受控参数以及(3)模式生成器相位数据作为输入应用到机器学习模型的反馈分量来生成机器人控制参数。在许多实现方式中，机器学习模型的反馈分量与机器学习模型的模式生成器分量解耦。

[0091] 在块612，系统基于机器人控制参数来控制机器人的致动器(或多个)。

[0092] 图7是示出根据多种实现方式的训练用于机器人运动的机器学习模型的示例处理700的流程图。为了方便，参考执行操作的系统来描述处理700的操作。该系统可以包括各种计算机系统的各种组件，诸如图8和/或图9所示的一个或多个组件。此外，尽管以特定顺序示出处理700的操作，但这并不意味着是限制性的。可以重新排序，省略和/或添加一个或多个操作。

[0093] 在块702，系统通过将受控参数作为输入应用于机器学习模型的模式生成器分量来生成模式生成器相位数据。在各种实现方式中，受控参数由用户提供给系统。

[0094] 在块704，系统通过将(1)观察、(2)受控参数以及(3)模式生成器相位数据作为输入应用到机器学习模型的反馈分量来生成机器人控制参数的实例。在许多实现方式中，机器学习模型的反馈分量与模式生成器分量解耦。

[0095] 在块706，系统基于机器人控制参数的实例控制机器人的致动器(或多个)。例如，系统可以通过控制机器人的致动器(或多个)来移动机器人的一条或多条腿。

[0096] 在块708，系统确定更新后的观察。在许多实现方式中，使用由机器人的一个或多个传感器捕获的反馈数据来确定更新后的观察。

[0097] 在块710，系统基于观察、更新后的观察和受控参数来确定奖励信号。在一些实现方式中，奖励信号优化能量有效的运动。在一些实现方式中，奖励信号类似于在图4的块408处确定的奖励信号。

[0098] 在块712，系统使用奖励信号来更新机器学习模型的一个或多个参数。例如，可以更新机器学习模型的反馈分量的一个或多个权重。在许多实现方式中，在每个情节处累积奖励。在一些实现方式中，训练情节在满足特定机器人条件之后终止，例如：机器人已进行期望数量的步(例如，训练情节在机器人已进行1000步之后终止)和/或机器人失去平衡(例如，机器人基座相对于地面倾斜超过0.5弧度)。

[0099] 图8示意性地描绘了机器人825的示例架构。机器人825包括机器人控制系统860、一个或多个操作组件825a-825n以及一个或多个传感器842a-842m。传感器842a-842m可以包括例如视觉传感器、光传感器、压力传感器、压力波传感器(例如，麦克风)、接近传感器、加速计、陀螺仪、温度计、气压计等。尽管传感器842a-m被描述为与机器人825集成在一起，但这并不意味着是限制性的。在一些实现方式中，传感器842a-m可以位于机器人825的外部，例如作为独立单元。

[0100] 操作组件840a-840n可以包括例如一个或多个末端致动器和/或一个或多个伺服马达或其他致动器，以实现机器人的一个或多个组件的运动。例如，机器人825可以具有多个自由度，并且每个致动器可以响应于控制命令而在一个或多个自由度内控制机器人825的致动。如本文中所使用的，除了可以与致动器相关联并且将接收的控制命令转换成一个或多个信号以驱动致动器的任何驱动器(或多个)之外，术语致动器包括用于产生运动的机械或电气设备(例如，马达)。因此，向致动器提供控制命令可以包括向驱动器提供控制命令，该驱动器将控制命令转换为用于驱动电气或机械设备以产生期望运动的适当信号。

[0101] 机器人控制系统860可以在一个或多个处理器中实现，诸如机器人825的CPU、GPU和/或其他控制器(或多个)。在一些实现方式中，机器人825可以包括“脑箱(brain box)”，其可以包括控制系统860的全部或各个方面。例如，脑箱可以向操作组件840a-n提供数据的

实时突发，其中对于一个或多个操作组件840a-n中的每一个，每个实时突发包括一个或多个控制命令的集合，这些命令尤其指示运动参数(如果有的话)。在一些实现方式中，机器人控制系统860可以执行本文所述的处理300、400、500和/或700的一个或多个方面。如本文所述，在一些实现方式中，由控制系统860生成的控制命令的全部或各个方面可以定位机器人825的肢体以进行机器人运动任务。尽管在图8中示出控制系统860作为机器人825的组成部分，但是在一些实现方式中，控制系统860的所有或各个方面可以在与机器人825分离但与之通信的组件中实现。例如，控制系统860的全部或各个方面可以在与机器人825进行有线和/或无线通信的一个或多个计算设备上实现，诸如计算设备910。

[0102] 图9是示例计算设备910的框图，其可以可选地用于执行本文描述的技术的一个或多个方面。例如，在一些实现方式中，机器人825和/或其他机器人可以利用计算设备910来提供期望的运动。计算设备910通常包括至少一个处理器914，其经由总线子系统912与多个外围设备进行通信。这些外围设备可以包括存储子系统924(包括例如内存子系统925和文件存储子系统926)、用户接口输出设备920、用户接口输入设备922和网络接口子系统916。输入和输出设备允许用户与计算设备910交互。网络接口子系统916提供到外部网络的接口，并且耦合到其他计算设备中的对应接口设备。

[0103] 用户接口输入设备922可以包括键盘、指向设备(诸如鼠标、轨迹球、触摸板或图形输入板)、扫描仪、结合到显示器中的触摸屏、(音频输入设备诸如语音识别系统、麦克风和/或其他类型输入设备)。通常，术语“输入设备”的使用旨在包括将信息输入到计算设备910或通信网络上的所有可能类型的设备和方式。

[0104] 用户接口输出设备920可以包括显示子系统、打印机、传真机或非可视显示器(诸如音频输出设备)。显示子系统可以包括阴极射线管(CRT)、诸如液晶显示器(LCD)的平板设备、投影设备或其他用于创建可见图像的机制。显示子系统还可以例如经由音频输出设备来提供非视觉显示。通常，术语“输出设备”的使用旨在包括将信息从计算设备910输出到用户或另一机器或计算设备的所有可能类型的设备和方式。

[0105] 存储子系统924存储提供本文所述的部分或全部模块的功能的程序和数据构造。例如，存储子系统924可以包括执行图3、图4、图5和/或图7的处理的所选方面的逻辑。

[0106] 这些软件模块通常由处理器914单独或与其他处理器结合执行。存储子系统924中使用的内存925可以包括多个存储器，包括用于在程序执行期间存储指令和数据的主随机存取存储器(RAM)9S0以及存储固定指令的只读存储器(ROM)932。文件存储子系统926可以为程序和数据文件提供持久存储，并且可以包括硬盘驱动器、软盘驱动器以及相关联的可移动介质、CD-ROM驱动器、光盘驱动器或可移动介质盒。实现某些实现方式的功能的模块可以由文件存储子系统926存储在存储子系统924中，或者存储在处理器914可访问的其他机器中。

[0107] 总线子系统912提供了一种用于使计算设备910的各个组件和子系统按预期彼此通信的机制。尽管总线子系统912被示意性地示出为单个总线，但是总线子系统的替代实现方式可以使用多个总线。

[0108] 计算设备910可以具有各种类型，包括工作站、服务器、计算集群、刀片服务器、服务器场或任何其他数据处理系统或计算设备。由于计算机和网络的不断变化的性质，出于说明一些实现方式的目的，图9中描绘的对计算设备910的描述仅旨在作为特定示例。计算

设备910的许多其他配置可能具有比图9所示的计算设备更多或更少的组件。

[0109] 尽管本文已经描述和说明了若干实现方式，但是可以利用用于执行功能和/或获得结果的多种其他手段和/或结构和/或本文所述的一个或多个优点，并且这些变体和/或修改被认为在本文描述的实现方式的范围内。更一般地，本文描述的所有参数、尺寸、材料和配置均是示例性的，并且实际参数、尺寸、材料和/或配置将取决于使用的一个或多个教导的具体一个或多个应用。仅使用常规实验，本领域技术人员将认识到或能够确定本文所述的具体实现方式的许多等同形式。因此，应当理解，前述实现方式仅以示例的方式呈现，并且在所附权利要求及其等同物的范围内，可以以不同于具体描述和要求保护的方式来实施实现方式。本公开的实现方式针对本文所述的每个单独的特征、系统、物品、材料、套件和/或方法。另外，如果这样的特征、系统、物品、材料、套件和/或方法不是相互矛盾的，则包括两个或多个这样的特征、系统、物品、材料、套件和/或方法的任意组合包括在本公开的范围内。

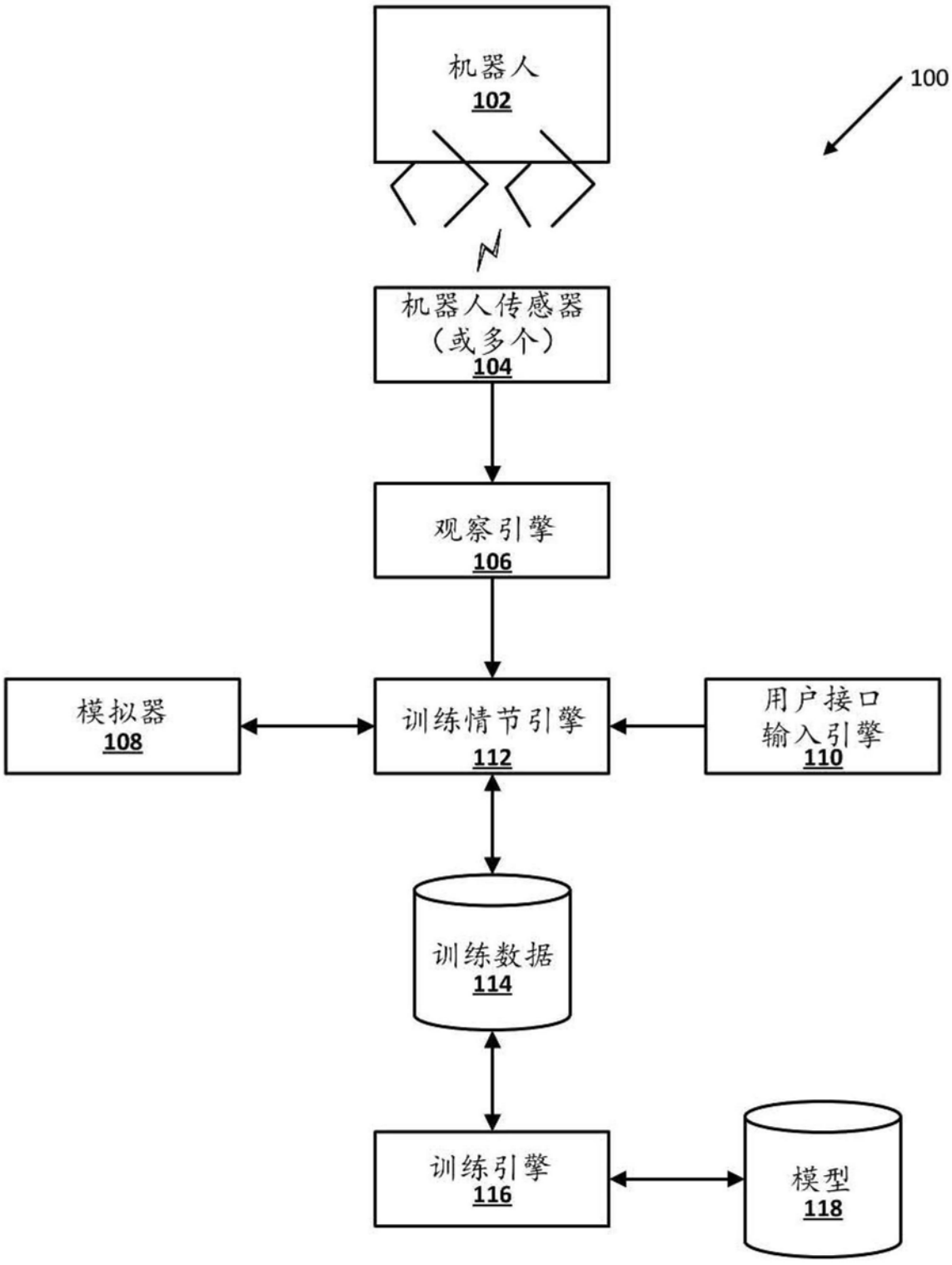

机器人
102
100
机器人传感器
（或多个）
104
观察引擎
106
模拟器
108
训练情节引擎
112
用户接口
输入引擎
110
训练数据
114
训练引擎
116
模型
118

图1

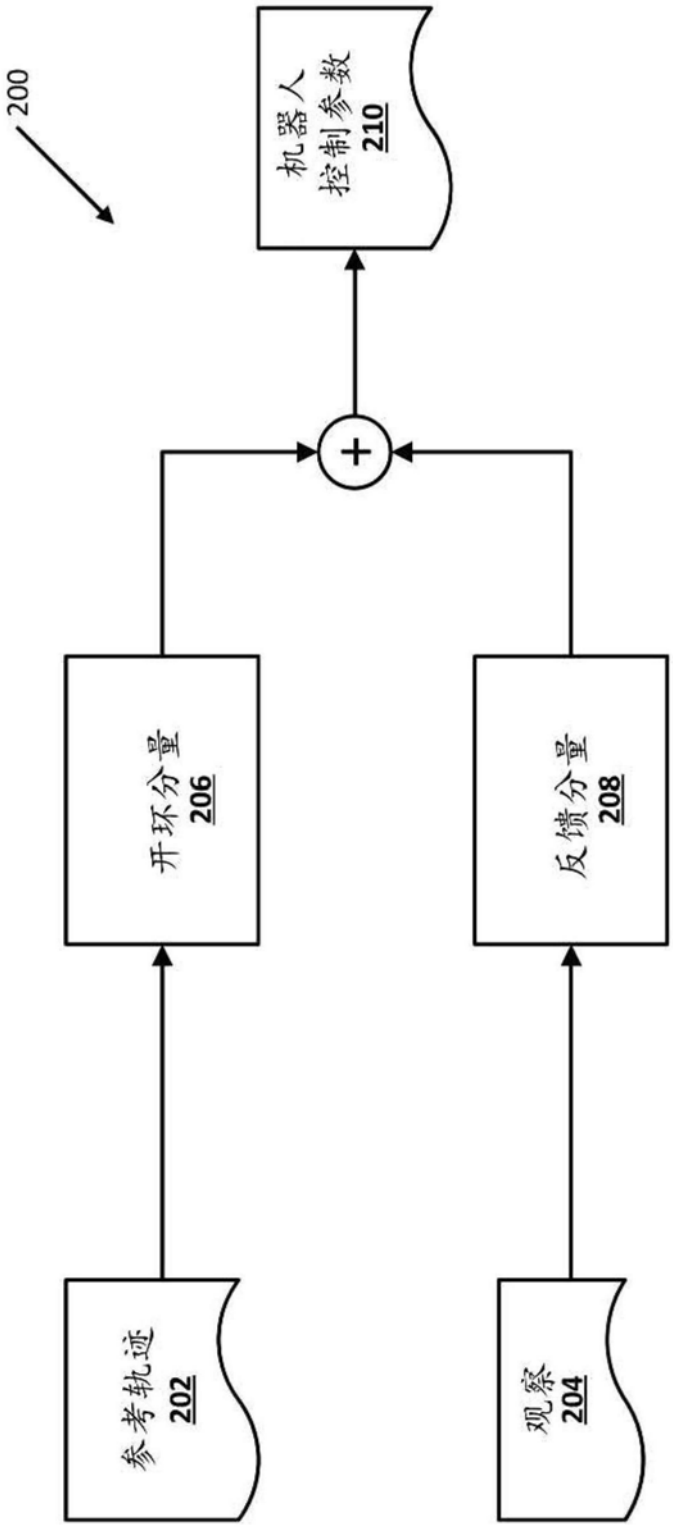


图2

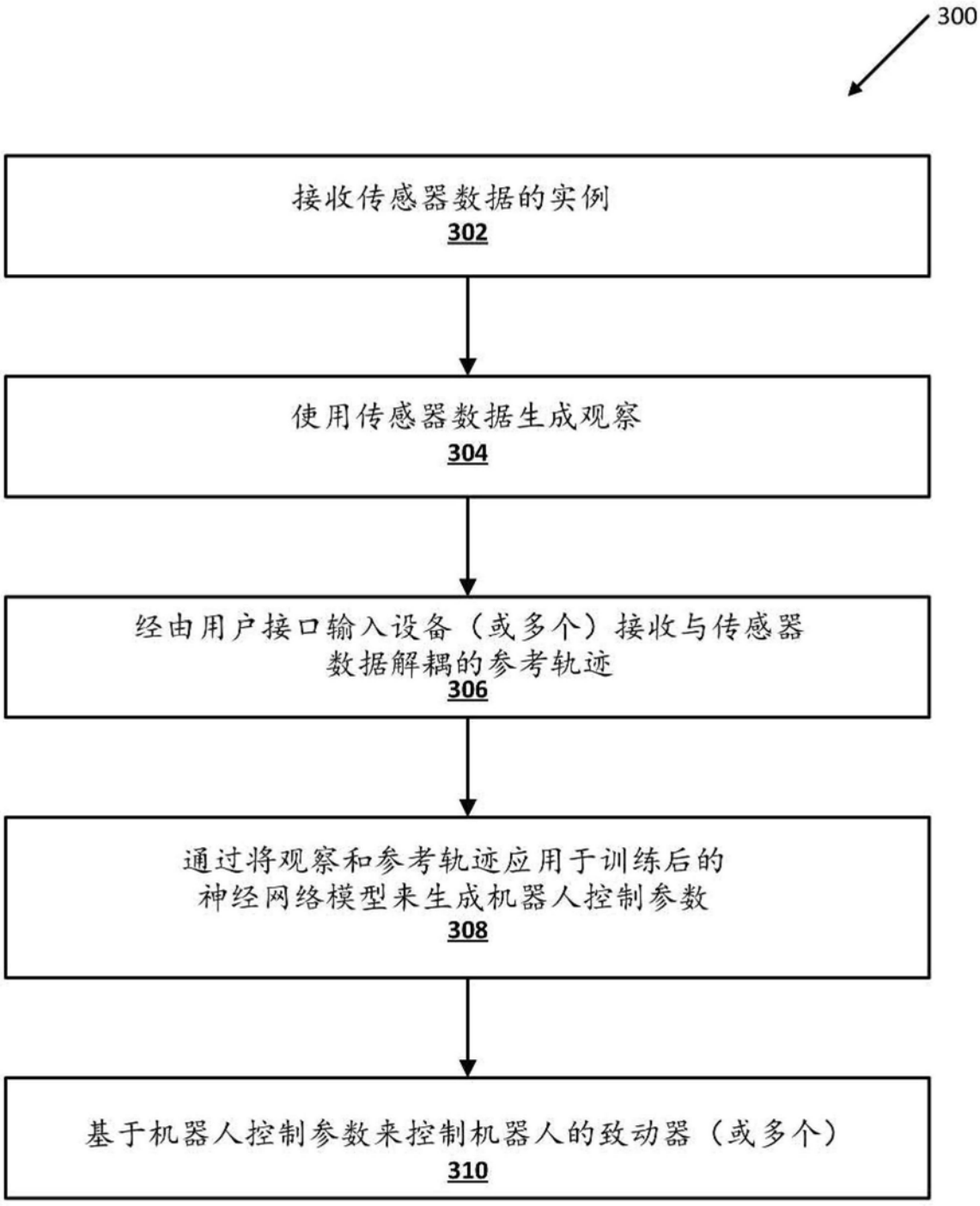

300
接收传感器数据的实例
302
使用传感器数据生成观察
304
经由用户接口输入设备（或多个）接收与传感器数据解耦的参考轨迹
306
通过将观察和参考轨迹应用于训练后的神经网络模型来生成机器人控制参数
308
基于机器人控制参数来控制机器人的致动器（或多个）
310

图3

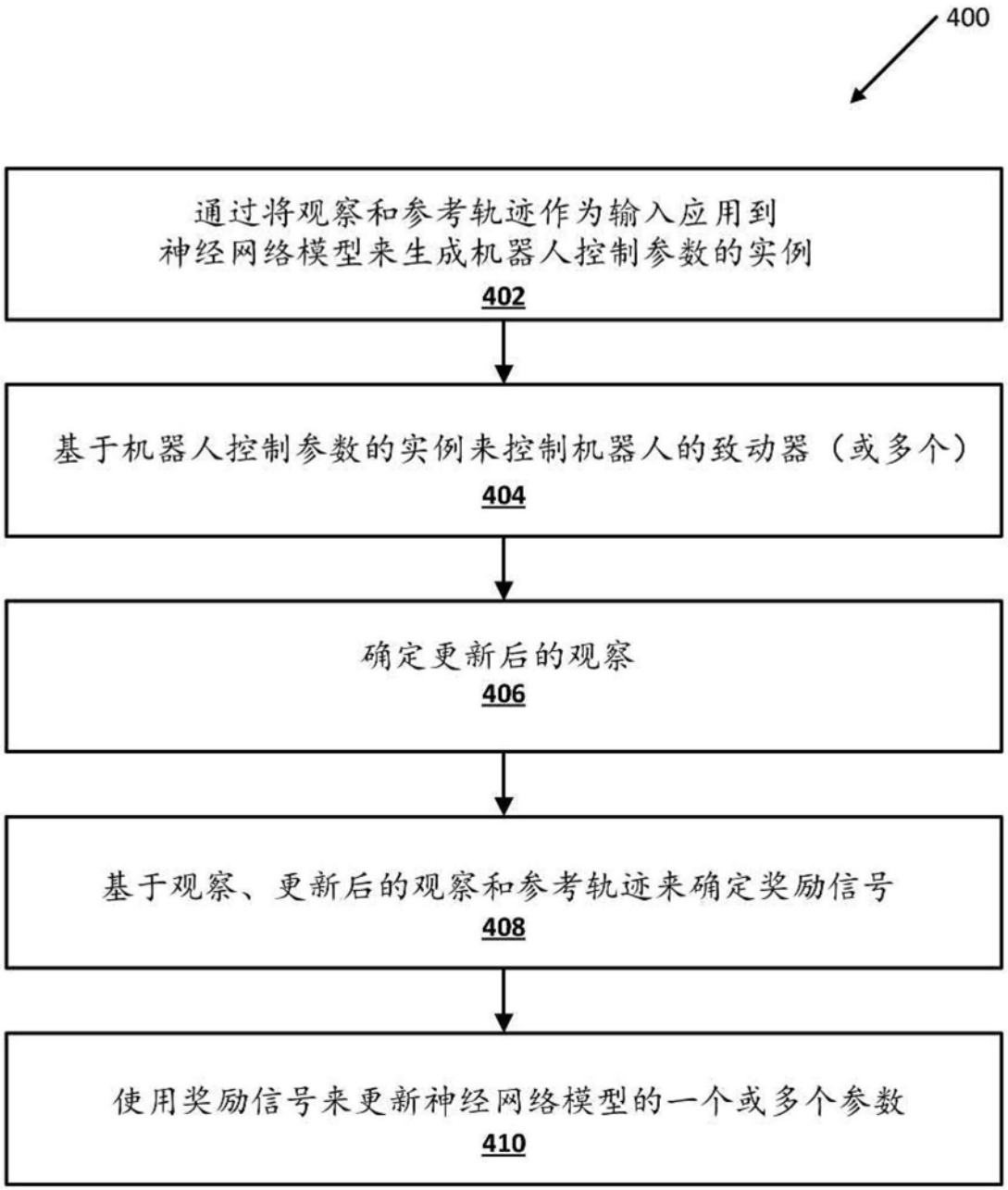

400
通过将观察和参考轨迹作为输入应用到
神经网络模型来生成机器人控制参数的实例
402
基于机器人控制参数的实例来控制机器人的致动器（或多个）
404
确定更新后的观察
406
基于观察、更新后的观察和参考轨迹来确定奖励信号
408
使用奖励信号来更新神经网络模型的一个或多个参数
410

图4

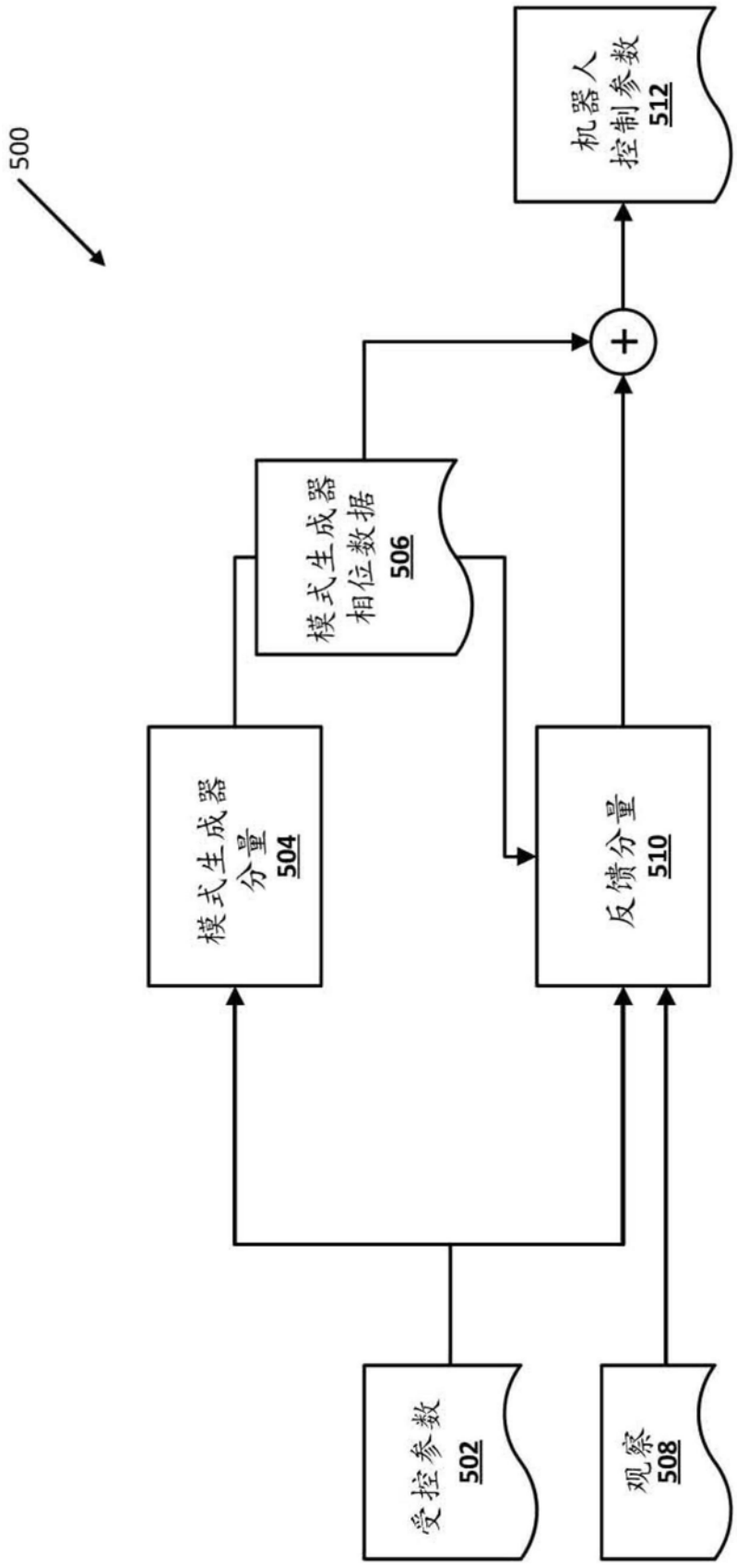

500
机器人
控制参数
512
模式生成器
相位数据
506
模式生成器
分量
504
反馈分量
510
受控参数
502
观察
508

图5

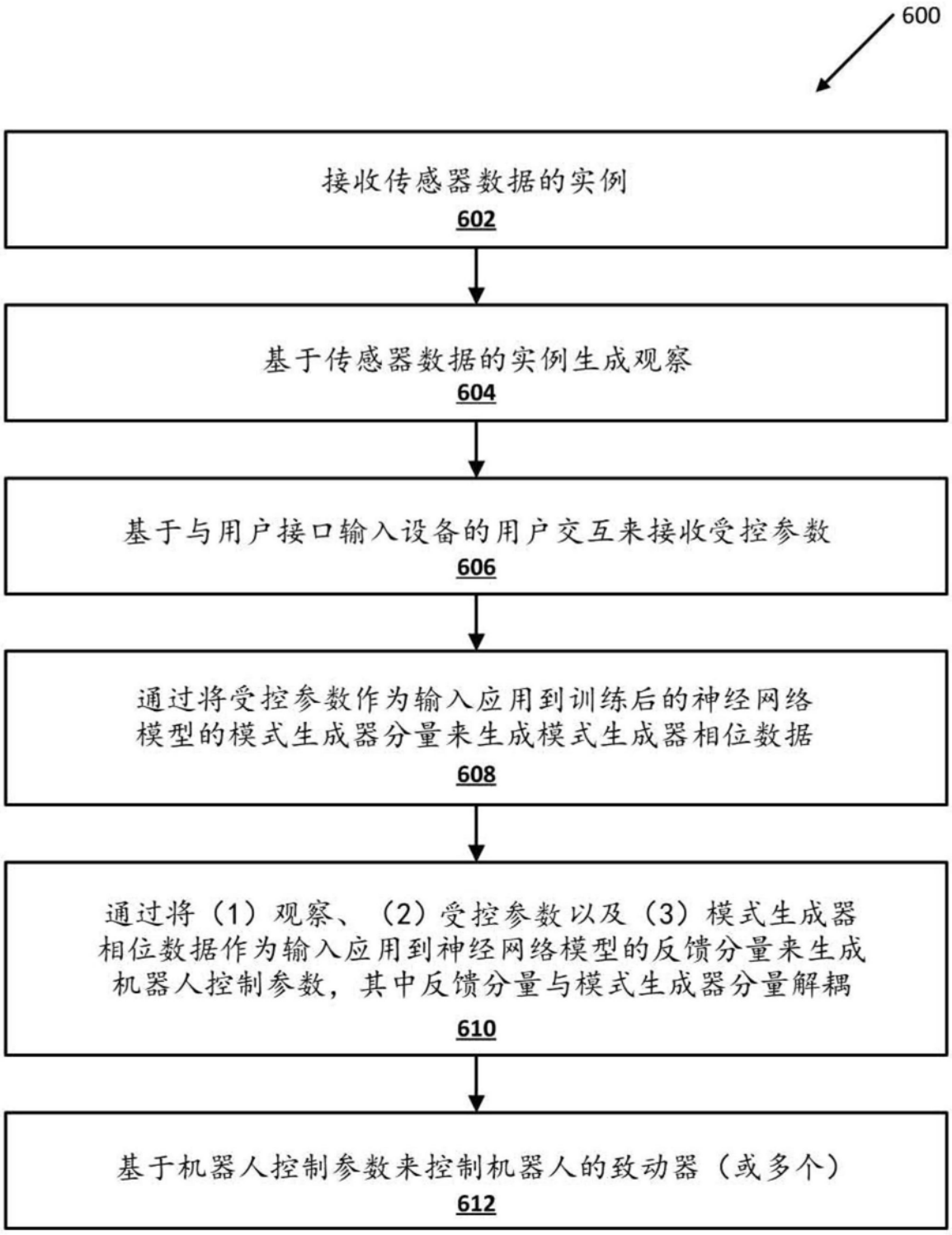

600
接收传感器数据的实例
602
基于传感器数据的实例生成观察
604
基于与用户接口输入设备的用户交互来接收受控参数
606
通过将受控参数作为输入应用到训练后的神经网络模型的模式生成器分量来生成模式生成器相位数据
608
通过将（1）观察、（2）受控参数以及（3）模式生成器相位数据作为输入应用到神经网络模型的反馈分量来生成机器人控制参数，其中反馈分量与模式生成器分量解耦
610
基于机器人控制参数来控制机器人的致动器（或多个）
612

图6

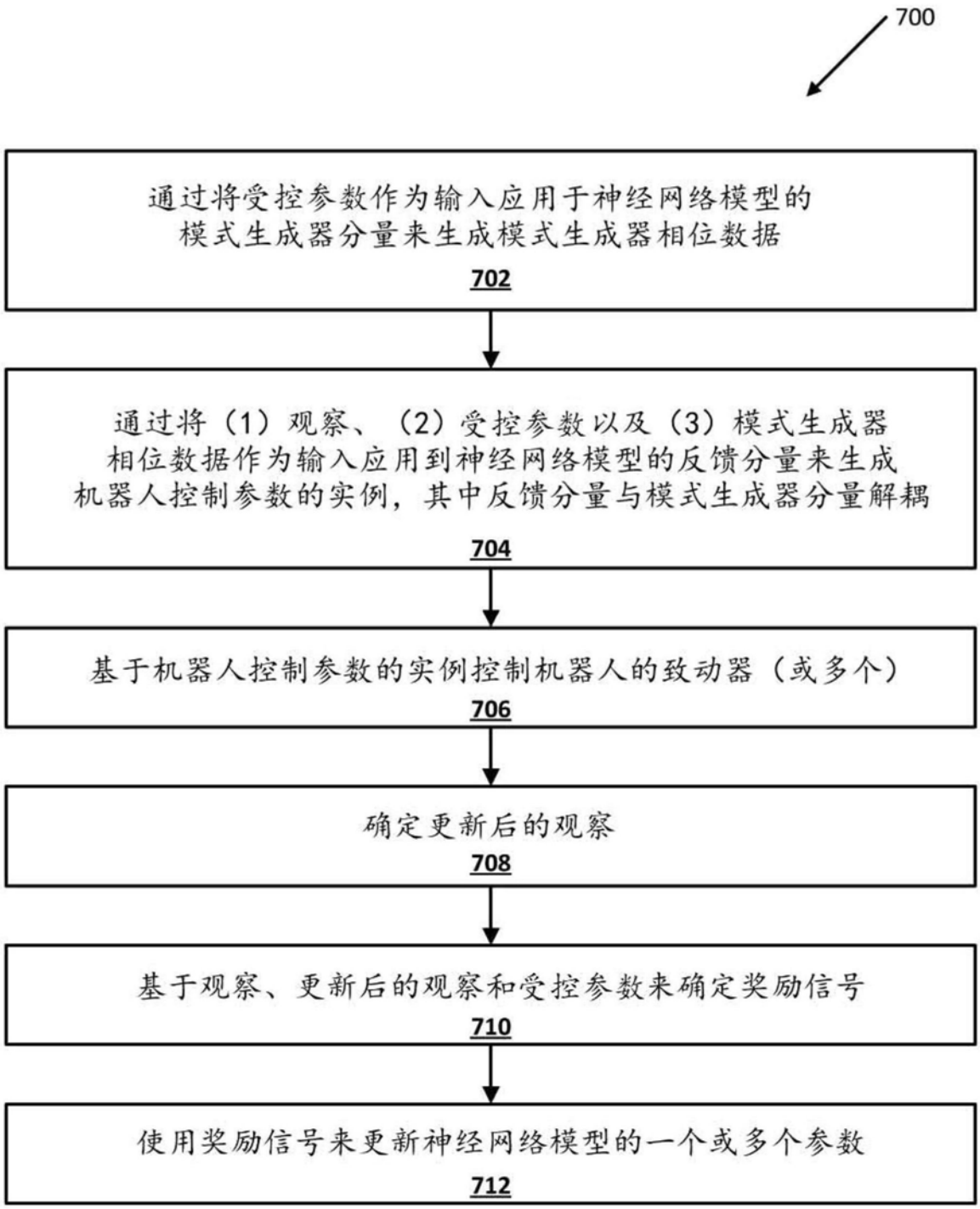

700
通过将受控参数作为输入应用于神经网络模型的模式生成器分量来生成模式生成器相位数据
702
通过将（1）观察、（2）受控参数以及（3）模式生成器相位数据作为输入应用到神经网络模型的反馈分量来生成机器人控制参数的实例，其中反馈分量与模式生成器分量解耦
704
基于机器人控制参数的实例控制机器人的致动器（或多个）
706
确定更新后的观察
708
基于观察、更新后的观察和受控参数来确定奖励信号
710
使用奖励信号来更新神经网络模型的一个或多个参数
712

图7

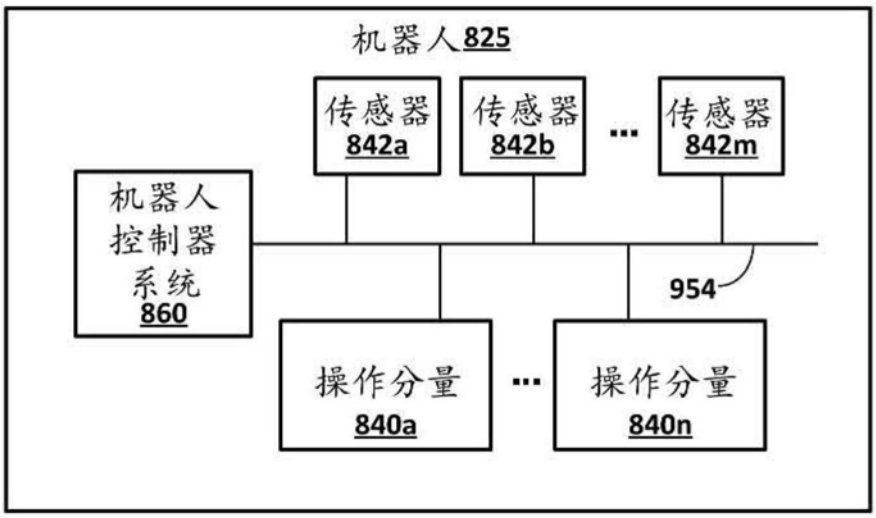

机器人825
传感器
842a
传感器
842b
...
传感器
842m
机器人
控制器
系统
860
954
操作分量
840a
...
操作分量
840n

图8

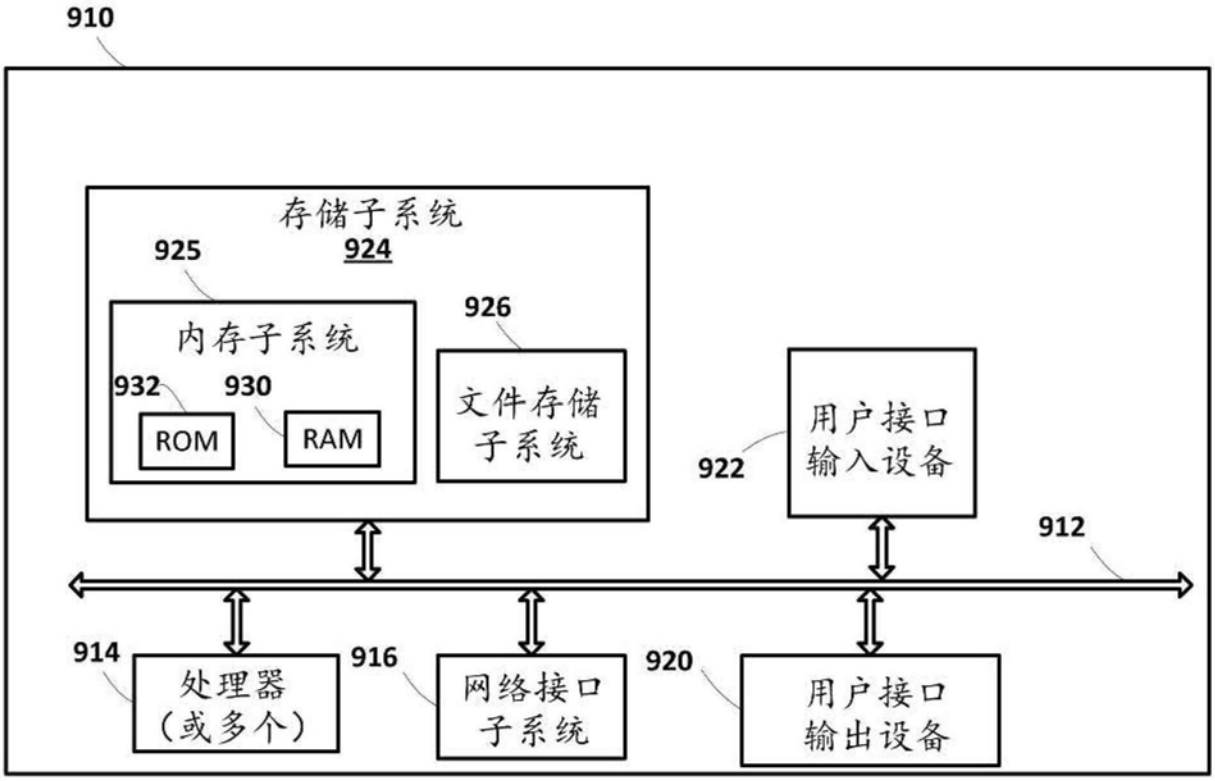

910
存储子系统
924
925
内存子系统
932
930
ROM
RAM
926
文件存储
子系统
922
用户接口
输入设备
912
914
处理器
(或多个)
916
网络接口
子系统
920
用户接口
输出设备

图9