(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102723112 B
(45)授权公告日 2015.06.17

(21)申请号 201210188573.2

(22)申请日 2012.06.08

(73)专利权人 西南大学
地址 400716 重庆市北碚区天生路2号

(72)发明人 王丽丹 何朋飞 段书凯 钟宇平

(74)专利代理机构 重庆弘旭专利代理有限责任公司 50209
代理人 周韶红

(51)Int.Cl.
G11C 16/34(2006.01)
G11C 16/24(2006.01)

(56)对比文件
US 2011/0004579 A1,2011.01.06, 全文.
CN 101951258 A,2011.01.19, 全文.
CN 102354128 A,2012.02.15, 全文.
高士咏等.“忆阻细胞神经网络及图像去噪和边缘提取中的应用”.《西南大学学报(自然科学版)》.2011, 第33卷(第11期), 全文.
胡柏林等.“忆阻器Simulink建模和图形用户界面设计”.《西南大学学报(自然科学版)》.2011, 第33卷(第9期), 全文.

审查员 周正

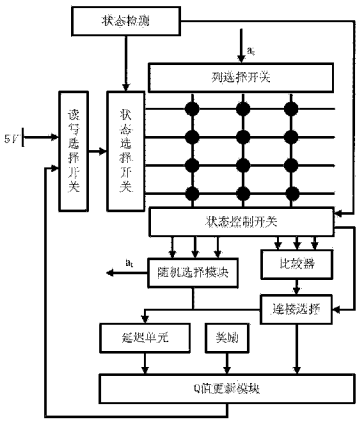
权利要求书1页 说明书4页 附图3页

(54)发明名称

一种基于忆阻交叉阵列的Q学习系统

(57)摘要

本发明公开了一种基于忆阻交叉阵列的Q学习系统,包括忆阻交叉阵列,其特征在于:所述系统还包括读写选择开关:控制忆阻交叉阵列的读写操作,状态选择开关:状态检测模块检测当前环境状态 s_t,通过状态选择开关,选择相应的行线;列选择开关:当需要对Q值,也即对忆阻交叉阵列的某一个忆阻值进行更新时,列选择开关选择动作 a_t所对应的列线。延迟单元:将选择的列线的电压延迟一个时间步长;状态检测模块:检测当前的环境状态,保存上一个环境状态。本发明将新的电路元件——忆阻器成功应用到了强化学习中,解决了强化学习需要大量的存储空间问题,为以后强化学习的研究提供了一种新的思路。



CN 102723112 B

1. 一种基于忆阻交叉阵列的 Q 学习系统，包括忆阻交叉阵列，其特征在于：所述系统还包括

读写选择开关：控制忆阻交叉阵列的读写操作；

状态选择开关：状态检测模块检测当前环境状态 s_t，通过状态选择开关，选择相应的行线；

列选择开关：当需要对 Q 值，也即对忆阻交叉阵列的某一个忆阻值进行更新时，列选择开关选择动作 a_t 所对应的列线；

延迟单元：将选择的列线的电压延迟一个时间步长；

状态检测模块：检测当前的环境状态，保存上一个环境状态，当需要根据状态选择动作时，状态检测模块检测当前环境状态，并将此状态提供给状态选择开关和状态控制开关，执行动作以后，状态选择开关检测此时的环境状态，保存上一个环境状态，并将此时的环境状态提供给状态选择开关和状态控制开关；当对 Q 值进行更新的时候，状态检测模块输出前一个时刻的环境状态，并提供给状态选择开关，选择相应的行线，在忆阻器两端加上写电压

$$V_i = \alpha(r + \gamma \max V(s_{t+1}, a) - V(s_t, a_t))$$

就可以去对 s_t 和 a_t 所对应的忆阻器的阻值进行更新，从而改变该忆阻器的输出电压 $V(s_t, a_t)$，也即 $Q(s_t, a_t)$ 值；此处 $V(s_t, a_t)$ 的值与 $Q(s_t, a_t)$ 值相等；

其中，α 为学习率，r 为奖励函数，γ 为折扣率。

一种基于忆阻交叉阵列的 Q 学习系统

技术领域

[0001] 本发明涉及一种存储矩阵和智能学习算法。

背景技术

[0002] 强化学习是一种高级的智能学习算法，近年来被广泛的应用于智能机器人领域，成为研究的热点。1954 年，Minsky 提出了 SNARCs 的强化学习计算模型。接着，Sutton 在其博士论文中提出了 AHC 算法和 TD 学习算法。后来，Watkins 等人在 TD 学习算法的基础上，提出了目前强化学习算法中的经典算法 -Q 学习算法，Q 学习算法是强化学习发展过程中的一个重要里程碑。Q 学习算法提出后，很多研究者将 Q 学习算法应用于移动机器人的导航，机器人足球系统和智能 I/O 的调度。但是强化学习也有其自身的局限性，当问题较为复杂时，它需要大量的状态 - 动作存储空间。1971 年，Chua 根据电路的完备性理论，提出了第四种电路元件 - 忆阻器 (L.O.Chua.Memristor-the missing circuit element.IEEE Trans. Circuit Theory.1971,18(5) :507-519.)。

[0003] 2008 年，HP 实验室成功制造了第一个物理实现的忆阻器，此后忆阻器引起了广泛的关注。忆阻器具有纳米尺寸、非线性特性，其阻值随着输入激励的变化而变化，并且这种变化是非易失性的，因此忆阻器非常适合用来设计大规模存储器。忆阻器交叉阵列是忆阻器存储器中的一种，它的结构简单，设计方便。胡小方等人利用忆阻器交叉阵列实现了图像的存储 (胡小方，段书凯，王丽丹，等 . 忆阻器交叉阵列及在图像处理中的应用 . 中国科学 F 辑 :信息科学 . 2011,41(4) :500-512.)。由于忆阻器具有纳米尺寸，因此忆阻器交叉阵列能够做成大规模存储器，可以解决强化学习在解决复杂问题时，需要大量的状态 - 动作存储空间的问题，因此，利用忆阻交叉阵列来实现 Q 学习是一种好的选择。

[0004] HP 忆阻器的物理模型如图 1 所示，忆阻器由掺杂区和非掺杂区两部分组成。其中 w 和 D 分别表示忆阻器中掺杂区域的宽度和忆阻器的总宽度。其数学模型如下 :

[0005] $$M(t)=R_{ON}\frac{w(t)}{D}+R_{OFF}(1-\frac{w(t)}{D})$$

[0006] 其中，R_{OFF}和 R_{ON}分别表示 w 等于 0 和 D 时，忆阻器的阻值。

[0007] $$\frac{dw(t)}{dt}=\frac{\mu_V R_{ON}}{D}i(t)$$

[0008] 这里，μ_v表示平均离子的移动，单位为 $cm^2s^{-1}V^{-1}$。

[0009] $$T_w=\frac{\Phi_D}{V_A R_{OFF}^2}[(R(w_0))^2-(R(w))^2]$$

[0010] 其中，

[0011] $$\Phi_D=\frac{(\beta D)^2}{2\mu_v(\beta-1)}$$

[0012] 这里，Tw 是输入忆阻器两端的脉冲电压的脉冲宽度，V_A是脉冲的幅度，$R(w_0)$ 表示忆阻器的初始阻值，R(w) 表示忆阻器可以达到的阻值，$\beta = R_{OFF}/R_{ON}$。

[0013] 当 $R(w_0)$ 小于等于 $R(w)$ 时，可以得到

[0014] $$R(w)=\sqrt{(R(w_0))^2-\frac{V_A T_w R_{OFF}^2}{\Phi_D}}\ ,\ R_{ON}\le R(w)\le R_{OFF}$$

[0015] 因此，当 Tw 一定时，随着 V_A 的变化，忆阻器的阻值会发生变化，并且这种变化是非易失性的。

[0016] 忆阻器存储电路如图 2 和图 3 所示。写入数据的电路如图 2 所示，读出数据的电路如图 3 所示。当写入数据时，给忆阻器加上一个正的电压脉冲，$R(w)$ 会减小，因此忆阻器会记忆所加电压脉冲。当读出数据时，忆阻器的阻值不同，得到的 V_{out} 也不同，V_{out} 与忆阻器的阻值之间形成了一个对应关系，因此能够正确反映忆阻器的阻值大小，也即忆阻器存储值的大小。

[0017] 忆阻器的阻值会随着输入激励的变化而变化，而且这种变化是非易失性；因此，忆阻器具有非常好的存储特性。并且，忆阻器具有纳米尺寸，非常适合用在大规模存储器中。而忆阻交叉阵列就是一个忆阻器作存储器的例子。

[0018] 忆阻交叉阵列的结构如图 4 所示，每一个圆形区域代表的电路如图 5 所示。在图 5 中，读 \ 写开关是写入数据和读出数据的控制开关。当给某一个忆阻器写入数据时，开关接左边的点，此时，对应的行线输入写数据电压 V_{in}；当读出某一个忆阻器的数据时，开关接右边的点，此时，对应的行线输入读数据电压 V_{in}，对应的列线输出电压 V_{out}。

发明内容

[0019] 本发明的目的是提供一种实现 Q 学习算法的基于忆阻交叉阵列的 Q 学习系统。

[0020] 为了实现上述目的，采用以下技术方案：一种基于忆阻交叉阵列的 Q 学习系统，包括忆阻交叉阵列，其特征在于：所述系统还包括

[0021] 读写选择开关：控制忆阻交叉阵列的读写操作；

[0022] 状态选择开关：状态检测模块检测当前环境状态 s_t，通过状态选择开关，选择相应的行线；

[0023] 列选择开关：当需要对 Q 值，也即对忆阻交叉阵列的某一个忆阻值进行更新时，列选择开关选择动作 a_t 所对应的列线。

[0024] 延迟单元：将选择的列线的电压延迟一个时间步长；

[0025] 状态检测模块：检测当前的环境状态，并且保存上一个环境状态。当需要根据状态选择动作时，状态检测模块检测当前环境状态，并将此状态提供给状态选择开关和状态控制开关。当执行动作以后，状态选择开关检测此时的环境状态，并且保存上一个环境状态，并将此时的环境状态提供给状态选择开关和状态控制开关。当对 Q 值进行更新的时候，状态检测模块输出前一个时刻的环境状态，并提供给状态选择开关，选择相应的行线。

[0026] 本发明将新的电路元件 - 忆阻器成功应用到了强化学习中，解决了强化学习需要大量的存储空间问题，为以后强化学习的研究提供了一种新的思路。

附图说明

[0027] 图 1 为 HP 忆阻器的物理模型结构图；

[0028] 图 2 为忆阻器写数据时的电路图；

[0029] 图 3 为忆阻器读数据时的电路图；

[0030] 图 4 为忆阻交叉阵列的结构示意图；

[0031] 图 5 为忆阻交叉阵列中单个忆阻电路图；

[0032] 图 6 为本发明的结构示意图；

[0033] 图 7 为本发明实施例中机器人和障碍物的结构示意图；

[0034] 图 8 为本实施例的仿真结果。

具体实施例

[0035] 下面结合附图和具体实施例对本发明做进一步描述。

[0036] Q 学习算法是强化学习算法中的一个经典算法，Q 学习中最简单的一种形式为单步 Q 学习，其 Q 值的更新公式为

[0037] $Q(s_t, a_t) = Q(s_t, a_t) + \alpha (r_{t+1} + \gamma \max Q(s_{t+1}, a) - Q(s_t, a_t))$

[0038] 其中，α 为学习率，γ 为折扣率。r_{t+1}表示在状态 s_t执行动作 a_t所获得环境的奖励。$Q(s_t, a_t)$ 表示动作状态对值函数，即在状态 s_t，执行动作 a_t，所得到的值的大小。

[0039] 强化学习的局限在于需要大量的存储空间，而新的电路元件 - 忆阻器，具有纳米尺寸和存储特性，基于忆阻器的交叉阵列具有大量的存储空间和并行处理能力，非常适合用来解决这个问题。

[0040] 在 Q 学习算法中，每执行一个动作，会得到环境的奖励值，并选择当前状态 - 动作对中的最大 Q 值和获得的奖励去更新前一个状态和选择的动作对的 Q 值。而用忆阻交叉阵列去实现 Q 学习时，每一个忆阻器的输出电压代表所对应的状态 - 动作对的 Q 值。根据忆阻器的存储原理，可以知道掉电之后阻值不会改变，因此只需在忆阻器两端加上写电压

[0041] $V_i = \alpha (r + \gamma \max V(s_{t+1}, a) - V(s_t, a_t))$

[0042] 就可以去对 s_t和 a_t所对应的忆阻器的阻值进行更新，从而改变该忆阻器的输出电压 $V(s_t, a_t)$，也即 $Q(s_t, a_t)$ 值。

[0043] 忆阻交叉阵列实现 Q 学习的过程如图 6 所示。忆阻交叉阵列中，每一条行线对应一个状态 s，每一条列些对应一个动作 a，其具体实现过程如下所示：

[0044] (1) 读写选择开关选择读有效，机器人中的状态检测模块检测当前环境状态 s_t，通过状态选择开关，选择相应的行线；

[0045] (2) 列选择开关选择所有列，通过状态控制开关将列线连接到随机选择模块，随机选择模块根据每个列线电压的大小随机的选择，电压越大的列线被选择的几率越大，最后随机选择出一个列线，根据选择的列线，得到执行的动作 a_t，机器人执行动作 a_t。也可以在设定的某些状态时，通过状态控制开关将列线连接到比较器模块，选择出电压最大的列线，再通过连接选择开关将该列线连接到延迟单元。通过状态选择开关、随机选择模块、比较器、连接选择模块就可以实现强化学习中的 ε -greedy 策略。

[0046] (3) 将选择的列线连接到延迟单元，延迟单元对列线的电压延迟一个时间步长；

[0047] (4) 状态检测模块检测当前环境状态，机器人进入状态 s_{t+1}，此时状态控制开关将列线连接到比较器，通过比较器，选择电压最大的列线，通过连接选择模块将该列线连接到 Q 值更新模块，Q 值更新模块将该电压与延迟单元的输出电压以及获得环境的奖励按照式 (7) 进行计算，得到写电压 V_i。

[0048] (5) 读写选择开关选择写有效，将写电压 V_i 加在忆阻器的两端，时间为 T_w。

[0049] (6) 重复上面的过程，直到达到设定的次数。

[0050] 机器人避障实验是要让机器人在有障碍的环境中实现无碰撞的行走。本实验采用基于忆阻交叉阵列的 Q 学习来实现机器人的学习，并最终实现无障碍的行走，本实验使用 mobotsim 软件。

[0051] 在图 7 中，圆形区域表示机器人，机器人上有三个传感器，数字 0-2 分别对应 3 个传感器，每一个传感器能够检测的最大距离是 1.5 米，黑色区域表示障碍物。

[0052] 在本实验中，把每一个传感器检测到的与障碍物的距离划分为 3 段，如下所示：

[0053]

$$s0=\begin{cases}0 & ;0\le dist0<0.45\text{以及发生碰撞}\\ 1 & ;0.45\le dist0<0.75\\ 2 & ;\text{其他情况}\end{cases}$$

[0054]

$$s1=\begin{cases}0 & ;0\le dist0<0.45\text{以及发生碰撞}\\ 1 & ;0.45\le dist0<0.75\\ 2 & ;\text{其他情况}\end{cases}$$

[0055]

$$s2=\begin{cases}0 & ;0\le dist0<0.45\text{以及发生碰撞}\\ 1 & ;0.45\le dist0<0.75\\ 2 & ;\text{其他情况}\end{cases}$$

[0056] 其中，dist0-dist2 分表表示每一个传感器检测到的到障碍物的距离，将 s0-s2 进行组合，会得到 27 种情况，将这 27 种情况作为机器人所处的环境中的 27 种状态，用一个三维数组 state[s0,s1,s2] 存储该 27 种状态。由于在本实验平台中，当机器人与障碍物碰撞或者传感器不能检测到障碍物时，传感器返回的值都是 -1，因此，将机器人与障碍物碰撞时的状态，归为状态 0，也即 s0-s2 都为 0 时的情况。

[0057] 奖赏函数 r 定义为：

[0058]

$$r=\begin{cases}-1 & ;\text{发生碰撞}\\ -0.5 & ;0.2\le dist<0.4\\ 0 & ;0.4\le dist<0.6\\ 1 & ;\text{其他情况}\end{cases}$$

[0059] 在本实验中，机器人将执行三种动作：前进，左转和右转。如果机器人所处的状态为 state[2,2,2] 时，动作的执行按照 Q 值的比重随机执行；其他状态时，执行 Q 值最大的动作。

[0060] 取 $\alpha = 0.8$，$\gamma = 0.98$，仿真次数设为 500 次，每次仿真 2000 步，实验仿真结果如图 8 所示。

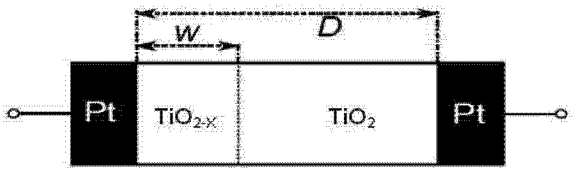


图 1

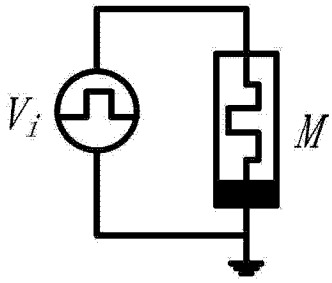


图 2

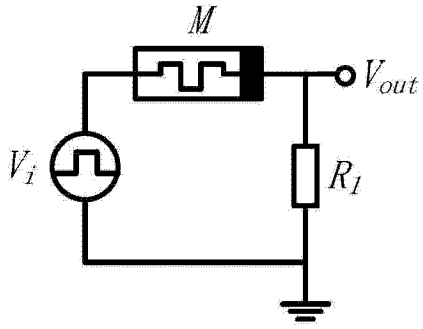


图 3

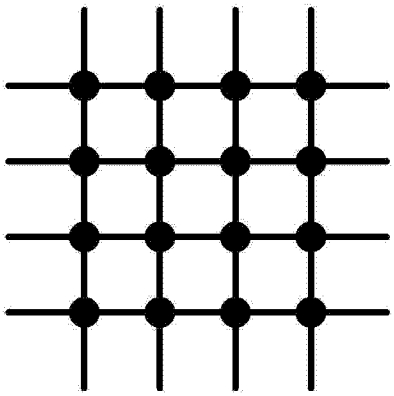


图 4

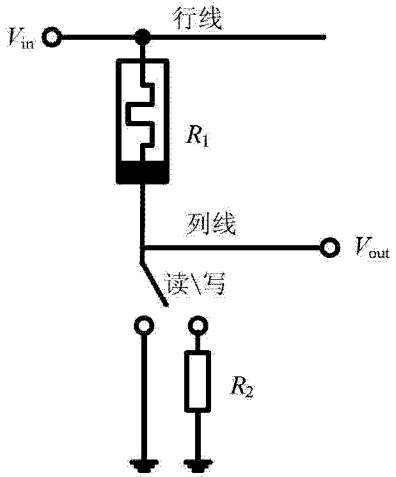


图 5

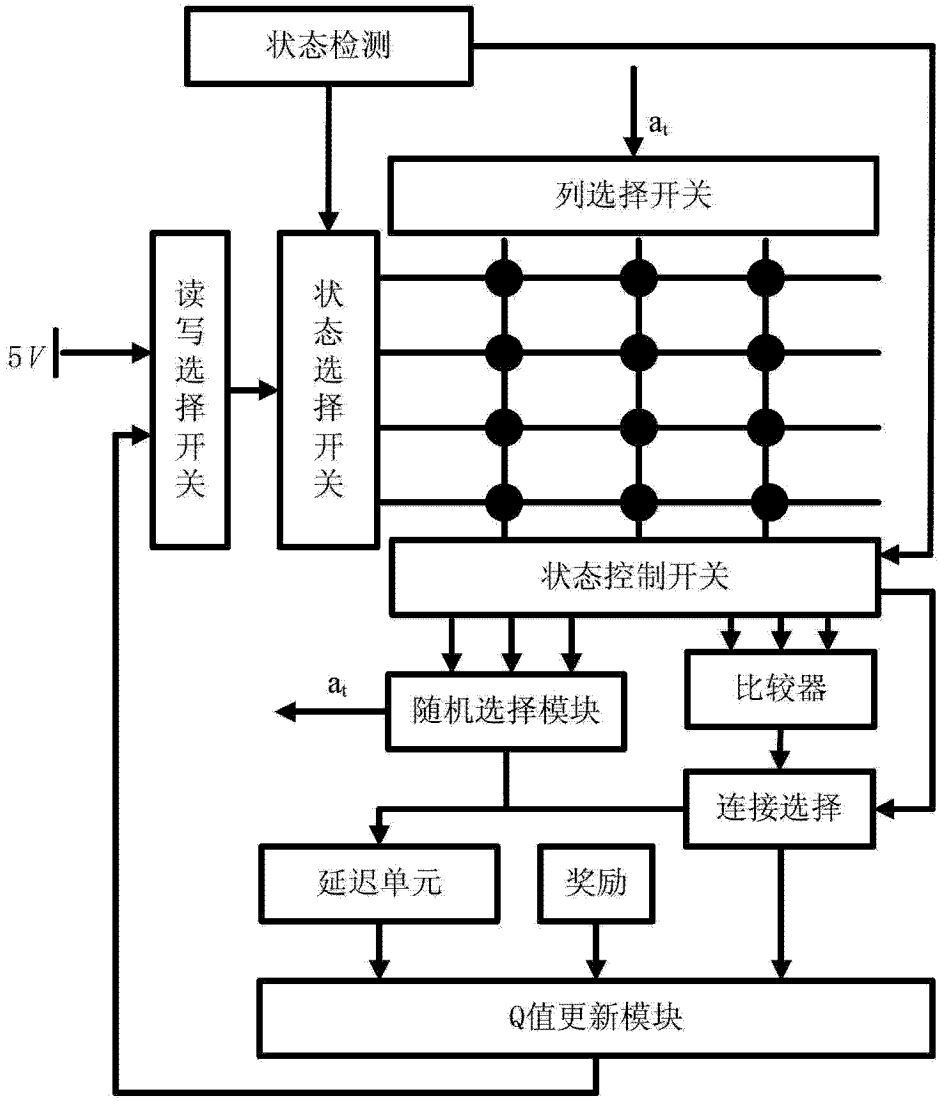

状态检测
a_t
列选择开关
读写选择开关
状态选择开关
5V
状态控制开关
a_t
随机选择模块
比较器
连接选择
延迟单元
奖励
Q值更新模块

图 6

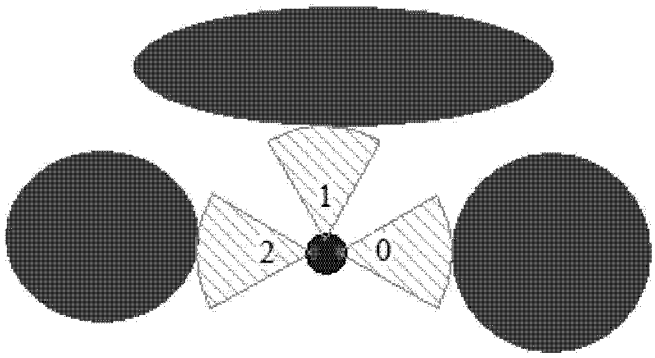

1
2
0

图 7

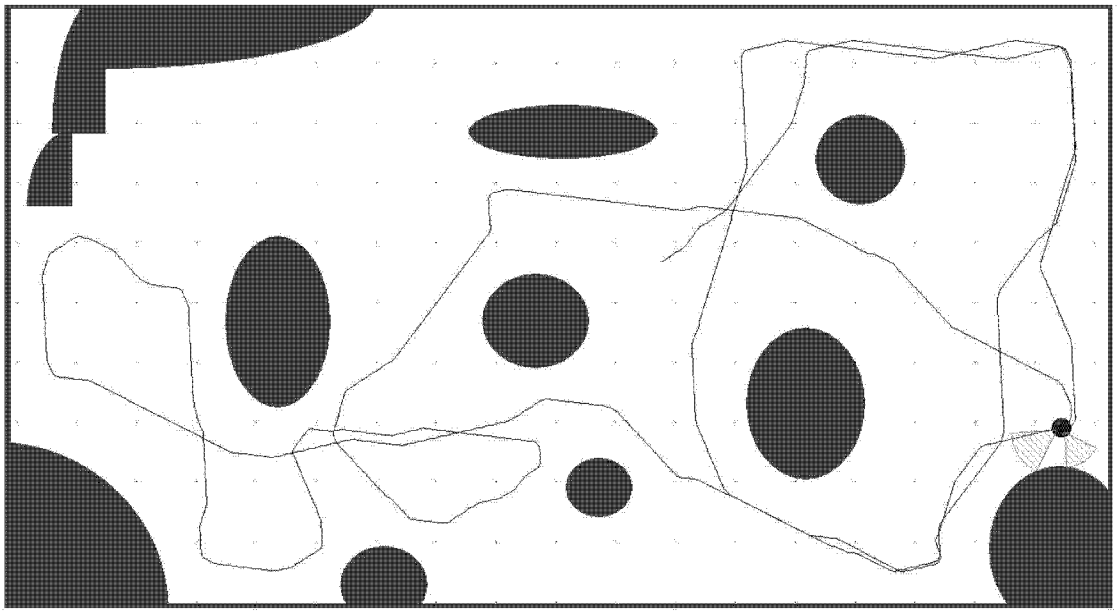


图 8