(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2018 年 11 月 15 日 (15.11.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/205778 A1

(51) 国际专利分类号:
G05B 13/00 (2006.01) *G05D 1/12* (2006.01)

(21) 国际申请号: PCT/CN2018/082105

(22) 国际申请日: 2018 年 4 月 8 日 (08.04.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710329549.9 2017年5月11日 (11.05.2017) CN

(71) 申请人: 苏州大学张家港工业技术研究院 (ZHANGJIAGANG INSITUTE OF INDUSTRIAL TECHNOLOGIES) [CN/CN]; 中国江苏省苏州市张家港市长泾路 10 号, Jiangsu 215600 (CN)。 苏州大学(SOOCHOW UNIVERSITY) [CN/CN]; 中国江苏省苏州市相城区济学路8号, Jiangsu 215006 (CN)。

(72) 发明人: 章宗长(ZHANG, Zongzhang); 中国江苏省苏州市苏州大学计算机科学与技术学院十梓街1号158信箱, Jiangsu 215000 (CN)。 潘致远(PAN, Zhiyuan); 中国江苏省苏州市苏州大学计算机科学与技术学院十梓街1号158信箱, Jiangsu 215000 (CN)。 王辉(WANG, Hui); 中国江苏省苏州市苏州大学计算机科学与技术学院十梓街1号158信箱, Jiangsu 215000 (CN)。

(74) 代理人: 苏州翔远专利代理事务所 (普通合伙) (SUZHOU FLY PATENT AGENCY (GENERAL PARTERNERSHIP)); 江苏省苏州市工业园区东环路1408号东环时代广场1607室陆金星, Jiangsu 215400 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

WO 2018/205778 A1

(54) **Title:** LARGE-RANGE MONITORING METHOD BASED ON DEEP WEIGHTED DOUBLE-Q LEARNING AND MONITORING ROBOT

(54) 发明名称: 基于深度带权双Q学习的大范围监控方法及监控机器人

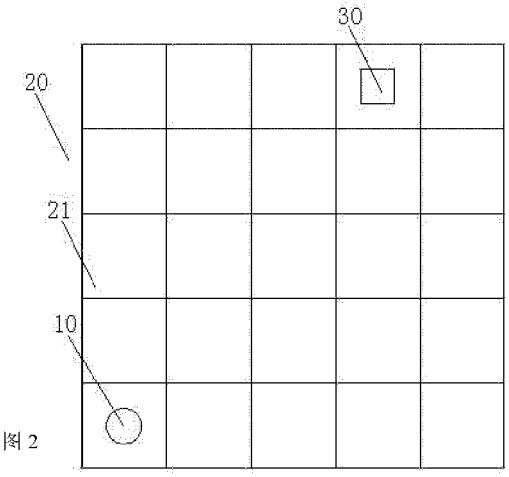


图 2

(57) **Abstract:** A large-range monitoring method based on deep weighted double-Q learning. The method involves: firstly, providing a robot (10) having a Q value table comprising a QA table and a QB table; secondly, an unidentified object (30) entering a large-range space (20) so as to trigger the robot (10); then, the robot (10) sensing a current state s, and determining whether the current state s is a target state; if so, the robot (10) reaching the next state and monitoring the unidentified object (30); if not, the robot (10) reaching the next state, and the robot (10) obtaining a reward value according to the next state; and the robot (10) choosing to update a QA value or a QB value equiprobably, and then updating a Q value, until the optimal monitoring policy is obtained through convergence. Not only does the method solve the problems of a limited monitoring range and a limited camera capacity, the problem of the synchronisation of a plurality of cameras also does not need to be considered, thereby reducing the cost.

[见续页]

根据细则4.17的声明：

— 关于申请人有权要求在先申请的优先权(细则4.17(iii))

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种基于深度带权双Q学习的大范围监控方法，首先，提供Q值表包括Q_A表和Q_B表的机器人（10），其次，不明物体（30）进入大范围空间（20）以触发机器人（10），再次，机器人（10）感知当前状态s，判断当前状态s是否为目标状态，如是，则机器人（10）到达下一状态并监控不明物体（30），如不是，机器人（10）到下一状态，机器人（10）根据下一状态得到奖赏值，机器人（10）等概率的选择更新Q_A值或Q_B值，然后更新Q值，直到收敛得到一个最优监控策略。本方法不但解决监控范围有限和摄像头容量有限的问题，而且不需考虑多台摄像头同步的问题，降低了成本。

基于深度带权双 Q 学习的大范围监控方法及监控机器人

技术领域

本发明涉及一种大范围监控领域，具体涉及一种基于深度带权双 Q 学习的大范围监控方法及监控机器人。

背景技术

在我们平时的生活中，监控系统已经无处不在，例如：交通路口的红绿灯监控、住宅小区里的安全监控等。监控系统结合多媒体技术、计算机网络、工业控制以及人工智能等多方面的知识，可用于安全防范、信息获取和调度指挥等方面，还可为生产流程，远程教育提供多种服务。但在一些需要完成具体任务的大范围环境中，例如寻找并追踪监控不明物体，目前的监控系统还不能完全布控。原因在于：一方面由于监控摄像头的固有缺陷，例如监控范围，信息容量有限等问题；另一方面需要考虑布控的成本，大范围布控需要多摄像头、多地的同步监控，布控难，成本高。综上所述，如何对大范围的环境进行监控已经成为一项急需解决的问题。

现有的监控系统存在如下缺陷：①有线模拟视频信号的传输距离有限。当传输距大于 1km 时，信号容易产生衰耗、畸变、群延等问题，图像质量将严重下降。②监控系统大多以录像的形式保存信息，但这种方法的容量有限，且需定期上传或替换存储介质，不能实现自动化监控。③多摄像头的监控往往成本过高，同时现有的监控摄像头还未考虑多设备同步的问题。

发明内容

本发明的发明目的是提供一种基于深度带权双 Q 学习的大范围监控方法，不但解决了因监控范围过大导致监控系统无法完全布控的难题，而且解决了摄像头容量有限的问题。同时通过深度估值网络，直接对图像信息进行特征提取，自行判断是否有不明物体出现在监视范围内，其中特征并直接与 Q 值表关联，省去了多台摄像头同步的问题，降低了成本。

为实现上述发明目的，本发明提供以下的技术方案：一种基于深度带权双 Q 学习的大范围监控方法，包括如下步骤：

S1、提供一大范围空间和一机器人，所述机器人在所述大范围空间中，在工作状态下，通过双 Q 学习方法从当前状态到达目标状态，所述机器人的 Q 值

表包括Q_A表和Q_B表，Q值通过深度估值网络参数θ进行计算，其中，

Q_A值的更新公式如下：

$$\beta_A = \frac{\left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|}{c + \left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_A Q_A(s',a^*;\theta) + (1-\beta_A)Q_B(s',a^*;\theta)\right] - Q_A(s,a;\theta);$$

$$Q_A \leftarrow Q_A(s,a;\theta) + \alpha(s,a)\delta;$$

Q_B值的更新公式如下：

$$\beta_B = \frac{\left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|}{c + \left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_B Q_B(s',a^*;\theta) + (1-\beta_B)Q_A(s',a^*;\theta)\right] - Q_B(s,a;\theta);$$

$$Q_B \leftarrow Q_B(s,a;\theta) + \alpha(s,a)\delta;$$

其中，β_A，β_B表示权重；s'表示下一状态；a^*表示下一状态的最优动作；a_L表示下一状态的最差动作；c为自由参数，c≥0；δ表示时间差分；R表示奖赏值；γ表示目标折扣，$0 \le \gamma \le 1$；s表示当前状态，a表示当前动作；α表示学习率，α在区间（0,1）内；θ表示深度估值网络参数；

工作状态时，所述大范围空间中有不明物体；

目标状态时，所述不明物体在所述机器人的监控范围内；

S2、所述机器人设置其初始状态为当前状态s；

S3、所述机器人检测并判断当前状态s是否为工作状态，如否，进入S4，如是，进入S5；

S4、所述机器人待机后到达下一状态s'，进入S11；

S5、所述机器人通过深度估值网络检测并判断当前状态s是否为目标状态，如否，进入S6，如是，进入S7；

S6、所述机器人选择并执行当前动作a后到达下一状态s'，进入S8；

S7、所述机器人选择并执行当前动作a后到达下一状态s'并监控所述不明物体，进入S8；

S8、所述机器人根据下一状态s'得到奖赏值R，进入S9；

S9、所述机器人等概率的选择更新Q_A值或Q_B值并进行更新，进入S10

S10、所述机器人判断其Q值表是否收敛，如否，进入S11，如是，进入S12；

S11、所述机器人重置下一状态s'为当前状态s，回到S3；

S12、所述机器人制定最优监控策略，进入S13；

S13、所述机器人重置下一状态s'为当前状态s，进入S14；

S14、所述机器人检测并判断当前状态s是否为工作状态，如否，进入S15，如是，进入S16；

S15、所述机器人待机后到达下一状态s'，返回S13；

S16、所述机器人检测并判断当前状态s是否为目标状态，如否，进入S17，如是，进入S18；

S17、所述机器人根据所述最优监控策略到达下一状态s'，回到S13；

S18、所述机器人选择并执行当前动作a后到达下一状态s'并监控所述不明物体，回到S13。

其中，在同一大范围空间中，所述机器人仅初次选择并执行当前动作a前初始化其Q值、学习率α、目标折扣γ、深度估值网络的结构和参数θ、动作选择方式以及权重β。

深度带权双Q方法利用深度估值网络，输入视频图像等状态信息和环境给予的反馈信号（如奖赏等），先通过可提取空间结构信息的卷积层神经网络来抽取图像中重要目标的特征信息，再通过全连接层的非线性变换来做分类或回归，最终在输出层产生每个动作的Q值。该网络通过网络参数θ,将奖赏值和误差项缩小至有限范围内，保证了Q值处于合理的范围内，并提高了该方法的稳定性。该网络对传统的Q学习算法进行了多处改进，其中包括：（1）在训练过程中利用经验回放机制，将得到的状态转移样本存放至记忆单元中，训练时从样本池中随机选择一小批样本对网络参数θ进行更新，增加了可利用的样本量，降低了样本间的关联性，提高了算法的稳定性。（2）除了使用深度网络表示当前Q值外，还另外使用一个网络来产生目标Q值，使用当前Q值和目标Q值

的线性组合来选择动作，并通过最小化当前 Q 值和目标 Q 值之间的均方误差来更新网络参数θ。引入该网络后，在一段时间内目标 Q 值保持不变，可以减少值波动对训练过程的影响，提升算法的稳定性。

上述技术方案中，所述大范围空间划分为若干子空间，所述机器人选择并执行当前动作a后，静止于当前子空间或移动至与当前子空间相邻的子空间，每个所述子空间不大于所述机器人的监控范围。

上述技术方案中，所述机器人检测当前状态s时，通过其传感器获知所述不明物体的大概位置loc_i和所述机器人的精确位置loc_a，记作$s=\langle loc_i, loc_a\rangle$。

上述技术方案中，所述机器人通过其摄像头获取图像信息，通过深度估值网络进行特征提取和分类，自行判断是否有不明物体在监控范围内，若有则通过其报警器进行报警。

上述技术方案中，所述机器人通过ò−$greedy$方式选择当前动作a。

上述技术方案中，所述机器人选择当前动作a时，有较大概率选择最大 Q 值所代表的动作，较小概率选择其他任意动作。

上述技术方案中，所述奖赏值R的设置如下：

$$R(s,a)=\begin{cases}10 & loc_i=loc_a\\ -0.1 & loc_i\neq loc_a\end{cases}$$

其中，loc_a为机器人的精确位置，loc_i为不明物体的大概位置，即当不明物体在机器人的监控范围内时给予正奖赏，而机器人未观测到不明物体时给予负奖赏。

上述技术方案中，所述机器人不断更新其学习率α，$\alpha(s,a)=\frac{1}{n(s,a)^{0.8}}$，其中，所述机器人执行当前动作$a$时不明物体也在移动构成双移状态，n 为双移状态下执行 a 动作的次数。

本发明还提供另外一个技术方案：一种基于深度带权双 Q 学习的大范围监控机器人，所述机器人的 Q 值表包括Q_A表和Q_B表，Q 值通过深度估值网络参数θ进行计算，其中，

Q_A值的更新公式如下：

$$\beta_A = \frac{\left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|}{c + \left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_A Q_A(s',a^*;\theta) + (1-\beta_A)Q_B(s',a^*;\theta)\right] - Q_A(s,a;\theta);$$

$$Q_A \leftarrow Q_A(s,a;\theta) + \alpha(s,a)\delta;$$

Q_B值的更新公式如下：

$$\beta_B = \frac{\left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|}{c + \left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_B Q_B(s',a^*;\theta) + (1-\beta_B)Q_A(s',a^*;\theta)\right] - Q_B(s,a;\theta);$$

$$Q_B \leftarrow Q_B(s,a;\theta) + \alpha(s,a)\delta;$$

其中，β_A，β_B表示权重；s'表示下一状态；a^*表示下一状态的最优动作；a_L表示下一状态的最差动作；c 为自由参数，c≥0；δ表示时间差分；R表示奖赏值；γ表示目标折扣，$0 \le \gamma \le 1$；s表示当前状态，a表示当前动作；α表示学习率，α在区间（0,1）内；θ表示深度估值网络参数；

所述机器人还设有实时检测其自身精确位置和不明物体大概位置的传感器和监控所述不明物体的摄像头，所述传感器和所述摄像头分别电连接所述机器人的主控芯片。

上述技术方案中，所述机器人还设有报警器，所述报警器电连接所述机器人的主控芯片。

由于上述技术方案运用，本发明与现有技术相比具有以下优点：

（1）本发明公开的基于深度带权双Q学习的大范围监控方法，能够很好地应对机器人在大范围空间中运动方式不受控的目标（不明物体）的移动跟踪问题，深度带权双Q学习方法通过对感知到的状态的判断，计算该状态下的风险Q值，随后根据该风险Q值来指导机器人选择动作进行移动，实现在无法知道目标（不明物体）的意图，难以对目标（不明物体）的行为进行判断的情况下监控大范围区域的目的。深度带权双Q学习方法是一种在Q学习方法和双Q学习方法之间的折衷方法，来指导 Agent 的下一步行动。原始的 Q 学习方法可用于大范围空间问题，却不能应对目标运动方式不受控的情况，而双 Q 学习方法虽然有效地解

决了运动方式不受控的目标监控问题，但未能够在大范围情景中应用，而深度带权双 Q 学习方法能够很好地应对大范围环境中的实时监控难题。在使用深度带权双Q学习方法实时监控时，需要对Q值表进行更新，Q值表分为Q_A表和Q_B表，每次行动后，各有50%的概率对其中一个表的值进行更新。若需更新Q_A 表中的值，则从Q_A 中选择所有可能的下一状态中拥有最大Q值的动作，再利用该动作在Q_A表和Q_B表中的值进行更新。Q_A表和Q_B表利用的比例由权值β决定。β的计算考虑了不明物体的状态、机器人的状态和当前动作，在传统的Q学习方法和双Q学习方法之间做了折衷。方法通过不断减少目标估计值与当前估计值之差δ，最终会收敛得到一个能监视并追踪不明物体的最优监控策略。

（2）机器人感知大范围空间中自身位置以及不明物体的大致位置，通过双Q学习达到目标状态进行监控，获取不明物体的图像信息并根据图像信息进行报警。

（3）机器人能够导航追踪不明物体，因此，仅需1台机器人即可完成整个危险区域的监控，无需考虑多台监视器同步的问题，而且节约了成本。

（4）机器人感知的状态为具有马尔科夫性质的状态，具有马尔科夫性质的状态的未来状态只与当前状态有关，与之前的状态没有关系，因此摄像头无需保存过去的信息，只需专注于当前自身和不明物体的位置，既解决了摄像头信息容量有限的问题，而且节约了成本。

（5）机器人在导航追踪不明物体时，通过设置奖赏值，根据该奖赏值，可以有效区分状态好的情况与不好的情况，提高机器人学习的速度，并且在一定程度上帮助机器人追踪不明物体。

(6)机器人根据其自身和目标（不明物体）的位置，选择是否需要回到充电点进行自动充电。

附图说明

图1是本发明公开的基于深度带权双Q学习的大范围监控机器人的组成示意图。

图2是本发明公开的基于深度带权双Q学习的大范围空间、机器人以及不明物体的位置示意图。

图3是本发明公开的基于深度带权双Q学习的大范围监控方法的流程图。

其中，10,、机器人；11、传感器；12、摄像头；13、主控芯片；20、大范围空间；21、子空间；30、不明物体。

具体实施方式

下面结合本发明的原理、附图以及实施例对本发明进一步描述

实施例一

参见图 1，如其中的图例所示，为一种基于深度带权双 Q 学习的大范围监控机器人 10，该机器人 10 的 Q 值表包括 Q_A 表和 Q_B 表，Q 值通过深度估值网络参数 θ 进行计算，其中，

Q_A 值的更新公式如下：

$$\beta_A = \frac{\left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|}{c + \left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_A Q_A(s',a^*;\theta) + (1-\beta_A)Q_B(s',a^*;\theta)\right] - Q_A(s,a;\theta);$$

$$Q_A \leftarrow Q_A(s,a;\theta) + \alpha(s,a)\delta;$$

Q_B 值的更新公式如下：

$$\beta_B = \frac{\left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|}{c + \left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_B Q_B(s',a^*;\theta) + (1-\beta_B)Q_A(s',a^*;\theta)\right] - Q_B(s,a;\theta);$$

$$Q_B \leftarrow Q_B(s,a;\theta) + \alpha(s,a)\delta;$$

其中，β_A，β_B 表示权重；s' 表示下一状态；a^* 表示下一状态的最优动作；a_L 表示下一状态的最差动作；c 为自由参数，c≥0；δ 表示时间差分；R 表示奖赏值；γ 表示目标折扣，$0 \le \gamma \le 1$；s 表示当前状态，a 表示当前动作；α 表示学习率，α 在区间（0,1）内；θ 表示深度估值网络参数；

该机器人 10 还设有实时检测其自身精确位置和不明物体大概位置的传感器 11 和监控不明物体的摄像头 12，传感器 11 和摄像头 12 分别电连接机器人 10 的主控芯片 13。

该机器人 10 获取图像后，图像将作为深度估值网络的输入。深度估值网络

由 8 层网络构成，所有网络节点均为修正线性单元 ReLU。第 1 层为输入层，状态是扁平化长为 84*84*3=21168 的向量，奖赏信号为数值型标量，第 2-5 层为卷积层。第 2 层的卷积核尺寸为 8*8，步长为 4*4，输出通道数为 32，这一层的输出维度为 20*20*32。第 3 层的卷积核尺寸为 4*4，步长为 2*2，输出通道数为 64，这一层的输出维度为 9*9*64。第 4 层的卷积核尺寸为 3*3，步长为 1*1，输出通道数为 64，这一层输出维度为 7*7*64。第 5 层的卷积核尺寸为 7*7，步长为 1*1，输出通道数为 1*1*512。第 7 层是全连接层，输出通道数为 512。第 8 层也是全连接层，输出通道数为行动数，即输出值为每个状态-行动对对应的 Q 值。在经验重放机制中，每批样本数目为 32，重放记忆单元大小为 1000000，目标 Q 值每 10000 个样本更新一次,当前 Q 值每行动数个样本更新一次。

参见图 2 和图 3，如其中的图例所示，一种基于深度带权双 Q 学习的大范围监控方法，包括如下步骤：

S1、提供一大范围空间 20 和一机器人 10，机器人 10 在大范围空间 20 中通过双 Q 学习方法在工作状态下从当前状态到达目标状态；

工作状态时，大范围空间 20 中有不明物体 30；

目标状态时，不明物体 30 在机器人 10 的监控范围内；

S2、机器人 10 设置其初始状态为当前状态 s；

S3、机器人 10 检测并判断当前状态 s 是否为工作状态，如否，进入 S4，如是，进入 S5；

S4、机器人 10 待机后到达下一状态 s'，进入 S11；

S5、机器人 10 检测并判断当前状态 s 是否为目标状态，如否，进入 S6，如是，进入 S7；

S6、机器人 10 选择并执行当前动作 a 后到达下一状态 s'，进入 S8；

S7、机器人 10 选择并执行当前动作 a 后到达下一状态 s' 并监控不明物体 30，进入 S8；

S8、机器人 10 根据下一状态 s' 得到奖赏值 R，进入 S9；

S9、机器人 10 等概率的选择更新 Q_A 值或 Q_B 值并进行更新，进入 S10

S10、机器人 10 判断其 Q 值表是否收敛，如否，进入 S11，如是，进入 S12；

S11、机器人 10 重置下一状态 s' 为当前状态 s，回到 S3；

S12、机器人 10 制定最优监控策略，进入 S13；

S13、机器人 10 重置下一状态 s' 为当前状态 s，进入 S14；

S14、机器人 10 通过深度估值网络检测并判断当前状态 s 是否为工作状态，如否，进入 S15，如是，进入 S16；

S15、机器人 10 待机后到达下一状态 s'，返回 S13；

S16、机器人 10 检测并判断当前状态 s 是否为目标状态，如否，进入 S17，如是，进入 S18；

S17、机器人 10 根据最优监控策略到达下一状态 s'，回到 S13；

S18、机器人 10 选择并执行当前动作 a 后到达下一状态 s' 并监控不明物体 30，回到 S13。

上述技术方案中，在同一大范围空间中，机器人 10 仅初次选择并执行当前动作 a 前初始化其 Q 值、学习率 α、目标折扣 γ、动作选择方式、权重 β、以及深度估值网络的结构和参数 θ 等，本实施例中，初始化后，Q 值为 0、自由参数 c 为 1、学习率 α 为 0.8、目标折扣 γ 为 0.95、动作选择方式为 $\grave{o}-greedy$ 方式、权重 β 为 0.5。

上述技术方案中，大范围空间 20 划分为若干子空间 21，机器人 10 选择并执行当前动作 a 后，静止于当前子空间或移动至与当前子空间相邻的子空间，每个子空间 21 不大于机器人 10 的监控范围。

上述技术方案中，机器人 10 检测当前状态 s 时，通过其传感器获知不明物体 30 的大概位置 loc_i 和机器人 10 的精确位置 loc_a，记作 $s=\langle loc_i, loc_a \rangle$。上述状态具有马尔科夫性质，具有马尔科夫性质的状态的未来状态只与当前状态有关，与之前的状态没有关系。

上述技术方案中，机器人10监控不明物体30时，通过其摄像头12获取不明物体的图像信息。

上述技术方案中，机器人10选择当前动作 a 时，有较大概率选择最大Q值所代表的动作，较小概率选择其他任意动作。

一种实施方式中，机器人10的摄像头12为360度可旋转摄像头。

一种实施方式中，机器人 10 还设有报警器（图中未视出），该报警器电连

接机器人 10 的主控芯片 13。机器人 10 根据图像信息，通过深度估值网络进行特征提取和分类，自行判断是否有不明物体在监控范围内，若有则通过其报警器进行报警。

一种实施方式中，机器人 10 通过ò− *greedy* 方式选择当前动作 *a*。动作为机器人 10 的运动方向，即向上、向下、向左、向右以及静止。

一种实施方式中，奖赏值 *R* 的设置如下：

$$R(s,a)=\begin{cases}10 & loc_i = loc_a \\ -0.1 & loc_i \neq loc_a\end{cases}$$

其中，loc_a 为机器人的精确位置，loc_i 为不明物体的大概位置，即当不明物体在机器人的监控范围内时给予正奖赏，而机器人未观测到不明物体时给予负奖赏。

一种实施方式中，机器人 10 不断更新其学习率 α，$\alpha(s,a)=\frac{1}{n(s,a)^{0.8}}$，其中，所述机器人执行当前动作 *a* 时不明物体也在移动构成双移状态，n 为双移状态下执行 a 动作的次数。

对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本实用新型的精神或范围的情况下，在其它实施例中实现。因此，本实用新型将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权 利 要 求 书

1.一种基于深度带权双 Q 学习的大范围监控方法，其特征在于，包括如下步骤：

S1、提供一大范围空间和一机器人，所述机器人在所述大范围空间中，在工作状态下通过双 Q 学习方法从当前状态到达目标状态，所述机器人的 Q 值表包括 Q_A 表和 Q_B 表，Q 值通过深度估值网络参数 θ 进行计算，其中，

Q_A 值的更新公式如下：

$$\beta_A = \frac{\left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|}{c + \left|Q_B(s',a^*;\theta) - Q_B(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_A Q_A(s',a^*;\theta) + (1-\beta_A)Q_B(s',a^*;\theta)\right] - Q_A(s,a;\theta);$$

$$Q_A \leftarrow Q_A(s,a;\theta) + \alpha(s,a)\delta;$$

Q_B 值的更新公式如下：

$$\beta_B = \frac{\left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|}{c + \left|Q_A(s',a^*;\theta) - Q_A(s',a_L;\theta)\right|};$$

$$\delta = R(s,a) + \gamma\left[\beta_B Q_B(s',a^*;\theta) + (1-\beta_B)Q_A(s',a^*;\theta)\right] - Q_B(s,a;\theta);$$

$$Q_B \leftarrow Q_B(s,a;\theta) + \alpha(s,a)\delta;$$

其中，β_A，β_B 表示权重；s' 表示下一状态；a^* 表示下一状态的最优动作；a_L 表示下一状态的最差动作；c 为自由参数，c≥0；δ 表示时间差分；R 表示奖赏值；γ 表示目标折扣，$0 \le \gamma \le 1$；s 表示当前状态，a 表示当前动作；α 表示学习率，α 在区间（0,1）内，θ 表示深度估值网络参数；

工作状态时，所述大范围空间中有不明物体；

目标状态时，所述不明物体在所述机器人的监控范围内；

S2、所述机器人设置其初始状态为当前状态 s；

S3、所述机器人检测并判断当前状态 s 是否为工作状态，如否，进入 S4，如是，进入 S5；

S4、所述机器人待机后到达下一状态 s'，进入 S11；

S5、所述机器人检测并判断当前状态 s 是否为目标状态，如否，进入 S6，

如是，进入 S7；

S6、所述机器人选择并执行当前动作 a 后到达下一状态 s'，进入 S8；

S7、所述机器人选择并执行当前动作 a 后到达下一状态 s' 并监控所述不明物体，进入 S8；

S8、所述机器人根据下一状态 s' 得到奖赏值 R，进入 S9；

S9、所述机器人等概率的选择更新 Q_A 值或 Q_B 值并进行更新，进入 S10

S10、所述机器人判断其 Q 值表是否收敛，如否，进入 S11，如是，进入 S12；

S11、所述机器人重置下一状态 s' 为当前状态 s，回到 S3；

S12、所述机器人制定最优监控策略，进入 S13；

S13、所述机器人重置下一状态 s' 为当前状态 s，进入 S14；

S14、所述机器人检测并判断当前状态 s 是否为工作状态，如否，进入 S15，如是，进入 S16；

S15、所述机器人待机后到达下一状态 s'，返回 S13；

S16、所述机器人通过深度估值网络检测并判断当前状态 s 是否为目标状态，如否，进入 S17，如是，进入 S18；

S17、所述机器人根据所述最优监控策略到达下一状态 s'，回到 S13；

S18、所述机器人选择并执行当前动作 a 后到达下一状态 s' 并监控所述不明物体，回到 S13。

其中，在同一大范围空间中，所述机器人仅初次选择并执行当前动作 a 前初始化其 Q 值、学习率 α、目标折扣 γ、深度估值网络的结构和参数 θ、动作选择方式以及权重 β。

2.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述大范围空间划分为若干子空间，所述机器人选择并执行当前动作 a 后，静止于当前子空间或移动至与当前子空间相邻的子空间，每个所述子空间不大于所述机器人的监控范围。

3.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述机器人检测当前状态 s 时，通过其传感器获知所述不明物体的大

概位置loc_i和所述机器人的精确位置loc_a，记作$s=\langle loc_i, loc_a\rangle$。

4.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述机器人通过其摄像头获取图像信息，通过深度估值网络进行特征提取和分类，自行判断是否有不明物体在监控范围内，若有则通过其报警器进行报警。

5.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述机器人通过ò− $greedy$方式选择当前动作a。

6.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述机器人选择当前动作a时，有较大概率选择最大 Q 值所代表的动作，较小概率选择其他任意动作。

7.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述奖赏值R的设置如下：

$$R(s,a)=\begin{cases}10 & loc_i=loc_a\\ -0.1 & loc_i\neq loc_a\end{cases}$$

其中，loc_a为机器人的精确位置，loc_i为不明物体的大概位置，即当不明物体在机器人的监控范围内时给予正奖赏，而机器人未观测到不明物体时给予负奖赏。

8.根据权利要求 1 所述的基于深度带权双 Q 学习的大范围监控方法，其特征在于，所述机器人不断更新其学习率α，$\alpha(s,a)=\dfrac{1}{n(s,a)^{0.8}}$，其中，所述机器人执行当前动作$a$时不明物体也在移动构成双移状态，n 为双移状态下执行 a 动作的次数。

9.一种基于深度带权双 Q 学习的大范围监控机器人，其特征在于，所述机器人的 Q 值表包括Q_A表和Q_B表，Q 值通过深度估值网络参数θ进行计算，其中，

Q_A值的更新公式如下：

$$\beta_A=\frac{\left|Q_B(s',a^*;\theta)-Q_B(s',a_L;\theta)\right|}{c+\left|Q_B(s',a^*;\theta)-Q_B(s',a_L;\theta)\right|};$$

$$\delta=R(s,a)+\gamma\left[\beta_A Q_A(s',a^*;\theta)+(1-\beta_A)Q_B(s',a^*;\theta)\right]-Q_A(s,a;\theta);$$

$Q_A \leftarrow Q_A(s,a;\theta)+\alpha(s,a)\delta$；

Q_B值的更新公式如下：

$$\beta_B = \frac{\left|Q_A(s',a^*;\theta)-Q_A(s',a_L;\theta)\right|}{c+\left|Q_A(s',a^*;\theta)-Q_A(s',a_L;\theta)\right|}$$；

$$\delta = R(s,a)+\gamma\left[\beta_B Q_B(s',a^*;\theta)+(1-\beta_B)Q_A(s',a^*;\theta)\right]-Q_B(s,a;\theta)$$；

$Q_B \leftarrow Q_B(s,a;\theta)+\alpha(s,a)\delta$；

其中，β_A，β_B表示权重；s'表示下一状态；a^*表示下一状态的最优动作；a_L表示下一状态的最差动作；c 为自由参数，c≥0；δ表示时间差分；R表示奖赏值；γ表示目标折扣，$0\le\gamma\le1$；s表示当前状态，a表示当前动作；α表示学习率，α在区间（0,1）内；θ表示深度估值网络参数；

所述机器人还设有实时检测其自身精确位置和不明物体大概位置的传感器和监控所述不明物体的摄像头，所述传感器和所述摄像头分别电连接所述机器人的主控芯片。

10.根据权利要求 9 所述的基于深度带权双 Q 学习的大范围监控机器人，其特征在于，所述机器人还设有报警器，所述报警器电连接所述机器人的主控芯片。

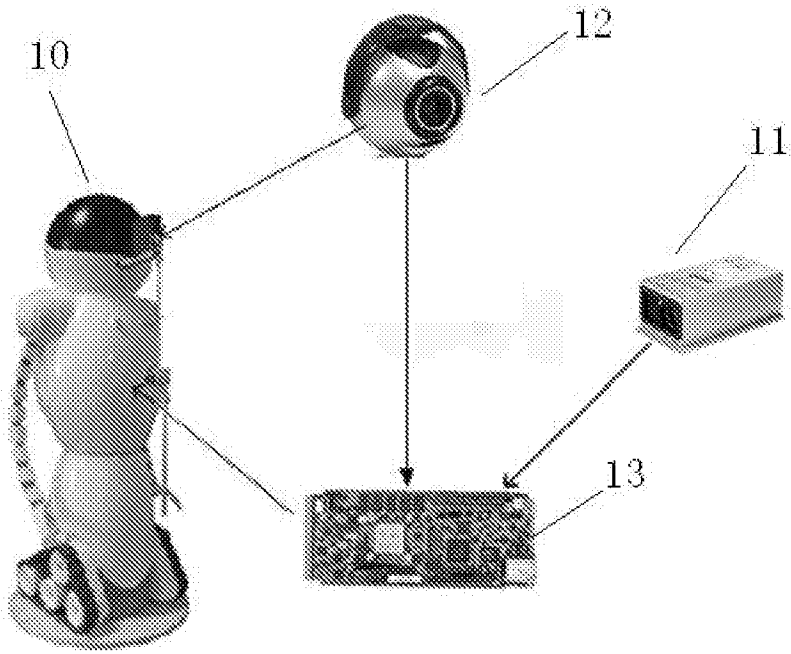


图 1

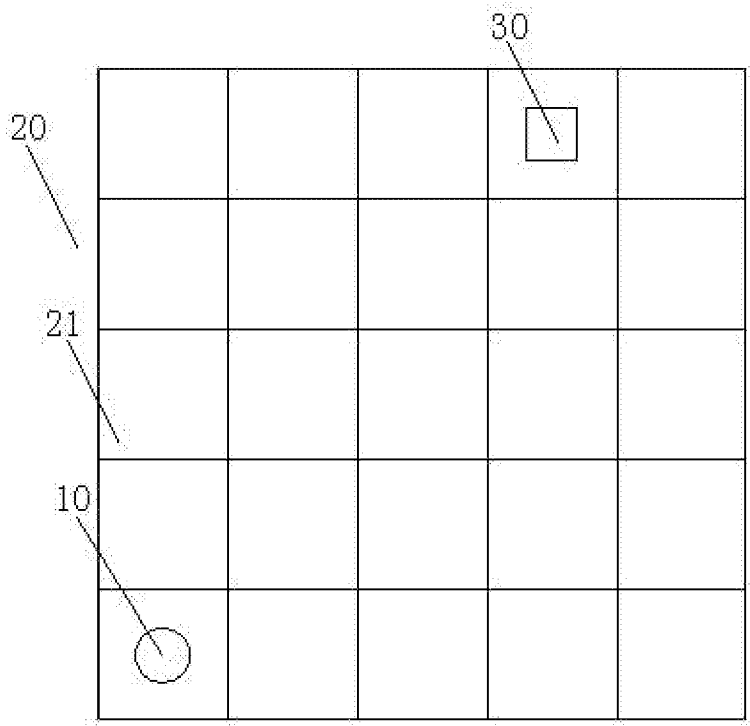


图 2

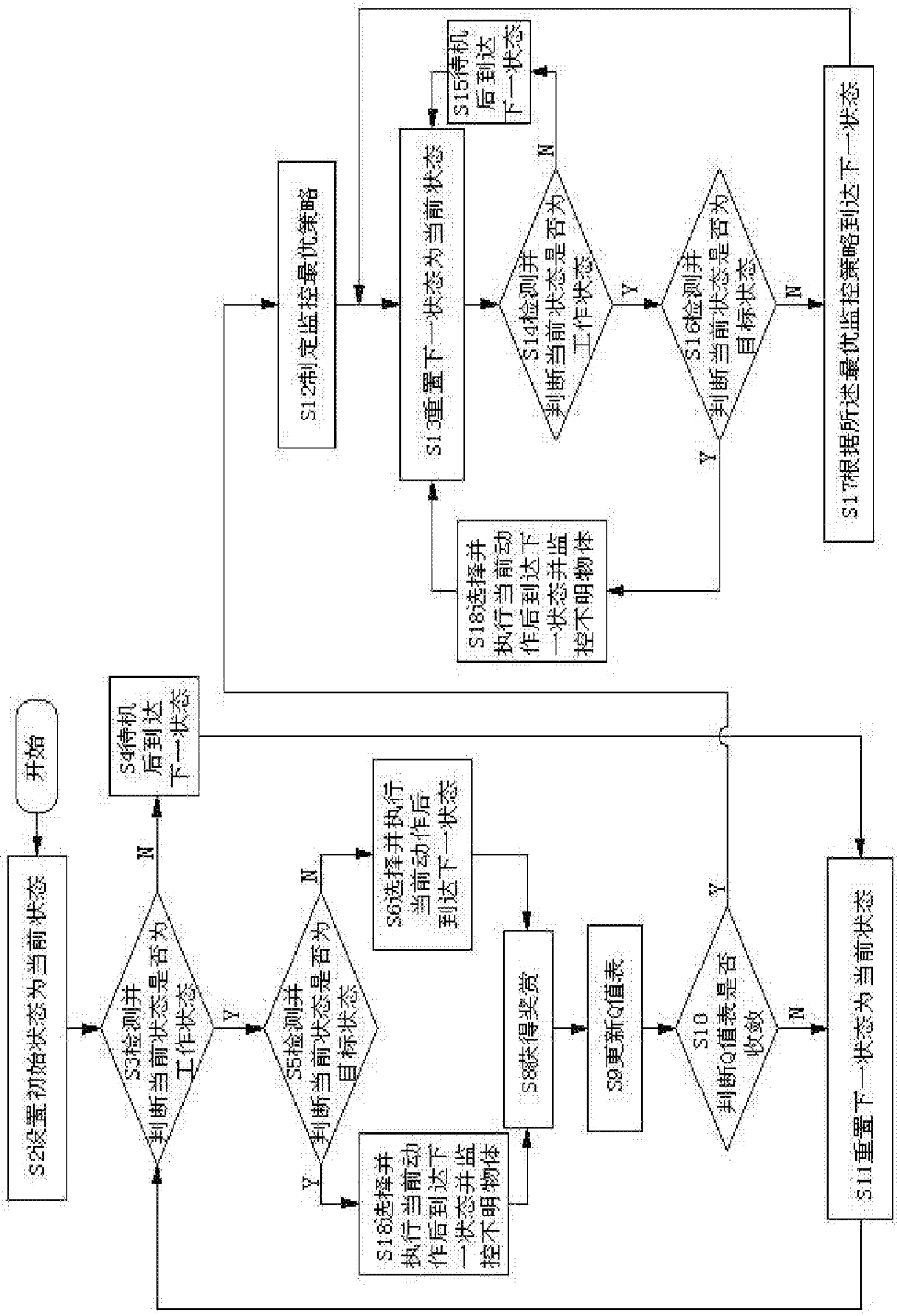

开始
S2设置初始状态为当前状态
S3检测并判断当前状态是否为工作状态
N
S4待机后到达下一状态
Y
S5检测并判断当前状态是否为目标状态
N
S6选择并执行当前动作后到达下一状态
Y
S18选择并执行当前动作后到达下一状态并监控不明物体
S8获得奖赏
S9更新Q值表
S10判断Q值表是否收敛
Y
N
S11重置下一状态为当前状态
S12制定监控最优策略
S13重置下一状态为当前状态
S14检测并判断当前状态是否为工作状态
N
S15待机后到达下一状态
Y
S16检测并判断当前状态是否为目标状态
Y
S18选择并执行当前动作后到达下一状态并监控不明物体
N
S17根据所述最优监控策略到达下一状态

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/082105

A. CLASSIFICATION OF SUBJECT MATTER

G05B 13/00 (2006.01) i; G05D 1/12 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G05B; G05D; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPODOC, WPI, CNPAT, IEEE, CNKI: 苏州大学, 深度, 估计, 网络, 双, Q, 学习, 更新, 权值, 机器人, 奖赏, 状态, 动作, deep, depth, network, double, dual, learning, weight+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 107292392 A (SOOCHOW UNIVERSITY), 24 October 2017 (24.10.2017), description, paragraphs [0071]-[0115]	1-10
A	CN 105867427 A (SOOCHOW UNIVERSITY), 17 August 2016 (17.08.2016), description, paragraphs [0018]-[0020]	1-10
A	CN 106598058 A (NORTH CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY), 26 April 2017 (26.04.2017), entire document	1-10
A	US 2017032245 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY), 02 February 2017 (02.02.2017), entire document	1-10
A	US 2010094786 A1 (HONDA MOTOR CO., LTD.), 15 April 2010 (15.04.2010), entire document	1-10

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 June 2018	05 July 2018

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
LI, Duo

Telephone No. 86-(10)-53961393

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/082105

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 107292392 A	24 October 2017	None	
CN 105867427 A	17 August 2016	None	
CN 106598058 A	26 April 2017	None	
US 2017032245 A1	02 February 2017	None	
US 2010094786 A1	15 April 2010	WO 2010045272 A1	22 April 2010

国际检索报告

国际申请号
PCT/CN2018/082105

A. 主题的分类

G05B 13/00(2006.01)i; G05D 1/12(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G05B; G05D; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

EPODOC, WPI, CNPAT, IEEE, CNKI: 苏州大学，深度，估计，网络，双，Q，学习，更新，权值，机器人，奖赏，状态，动作，deep, depth, network, double, dual, learning, weight+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 107292392 A (苏州大学) 2017年 10月 24日 (2017-10-24) 说明书第[0071]-[0115]段	1-10
A	CN 105867427 A (苏州大学) 2016年 8月 17日 (2016-08-17) 说明书第[0018]-[0020]段	1-10
A	CN 106598058 A (华北理工大学) 2017年 4月 26日 (2017-04-26) 全文	1-10
A	US 2017032245 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 2017年 2月 2日 (2017-02-02) 全文	1-10
A	US 2010094786 A1 (HONDA MOTOR CO., LTD.) 2010年 4月 15日 (2010-04-15) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2018年 6月 12日	2018年 7月 5日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 李铎 电话号码 86-(10)-53961393

表 PCT/ISA/210 (第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/082105

检索报告引用的专利文件	公布日（年/月/日）	同族专利	公布日（年/月/日）
CN 107292392 A	2017年 10月 24日	无	
CN 105867427 A	2016年 8月 17日	无	
CN 106598058 A	2017年 4月 26日	无	
US 2017032245 A1	2017年 2月 2日	无	
US 2010094786 A1	2010年 4月 15日	WO 2010045272 A1	2010年 4月 22日