doi:10.6041/j.issn.1000-1298.2014.02.010

基于强化学习的农业移动机器人视觉导航*

周　俊　陈　钦　梁　泉
(南京农业大学江苏省智能农业装备重点实验室, 南京 210031)

摘要: 以强化学习为基础,结合模糊逻辑理论研究了农业移动机器人通过自主学习获取导航控制策略的方法。首先使用机器视觉检测环境障碍并获取障碍物相对于移动机器人的方向和距离信息。然后应用强化学习设计了机器人自主获取导航控制策略方法,使机器人能够不断适应动态变化的导航环境。最后基于模糊逻辑离散化连续的障碍物方向和距离信息,构建了离散化的环境状态,并据此制定了自主导航学习 Q 值表。在自制的轮式移动机器人平台上开展了试验,结果表明机器人可以在实际导航环境中自动获取更优的导航策略,完成预期的导航任务。

关键词: 农业机器人　强化学习　模糊逻辑　视觉导航
中图分类号: TP242.6　　**文献标识码**: A　　**文章编号**: 1000-1298(2014)02-0053-06

引言

与工业应用场景不同,农业移动机器人的作业环境多数是开放且动态变化的,无法准确预知障碍几何形状、运动速度以及空间分布等信息,使得预先规划好的导航策略方法既繁琐又难以有效,这就需要机器人具有一定的自主学习能力来动态适应这种开放的变化环境。

目前,解决动态环境中机器人自主导航问题常用的方法有遗传算法[1]、人工势场法[2]、人工神经网络法[3]、蚁群优化算法[4]和模糊控制法[5-6]等。这些方法多数尚处于算法仿真阶段,有效性还需进一步检验。

强化学习通过智能体与环境之间的交互作用成为提高智能体适应未知环境能力的重要方法[7]。因此,本文以双目立体视觉作为机器人自主导航环境感知传感器,采用强化学习方法来增强农业移动机器人自主导航对动态场景的适应性。此外,为了克服强化学习方法不易于处理连续状态空间且容易出现维数爆炸问题,结合模糊逻辑方法,完成感知信息到移动机器人行为的离散化映射,提出基于强化学习与模糊逻辑相结合的反应式控制方法。最后在自制的移动机器人平台上进行试验。

1　自主导航系统

采用自行研制的移动机器人平台,如图 1 所示,该平台由两轮差动驱动本体、双目立体视觉和光电旋转编码器等几部分组成。双目立体视觉通过 1394 接口与主机连接,有效视距为 7 m,视野角度为 40°。光电旋转编码器分别置于移动机器人 2 个驱动轮的轮轴上,根据其读数可以实时解算移动机器人的速度和航向角信息。

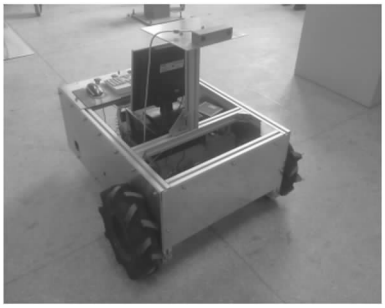

图 1　移动机器人平台

Fig. 1　Mobile robot platform

要实现移动机器人在动态作业环境中安全导航,必须检测环境中可能存在的各类运动和静止障碍,可以利用双目立体视觉来检测障碍物相对机器人的角度和距离信息[8-9]。如图 2 所示,在二维路面上定义移动机器人坐标系的 xoz 平面,其中 d_{ro} 和 θ_{ro} 分别是障碍物相对移动机器人的距离和方向信息。

2　基于强化学习和模糊逻辑的导航控制

由于无法事前规划,移动机器人的自学习能力对其在未知环境中安全导航尤其重要。在机器学习范畴,根据反馈的不同,学习技术可以分为监督学

收稿日期: 2012-12-23　修回日期: 2013-03-12
* 国家自然科学基金资助项目(31071325)和江苏省自然科学基金资助项目(BK2010458)
作者简介: 周俊,教授,主要从事农业机器人、机器视觉与模式识别研究,E-mail: zhoujun@njau.edu.cn

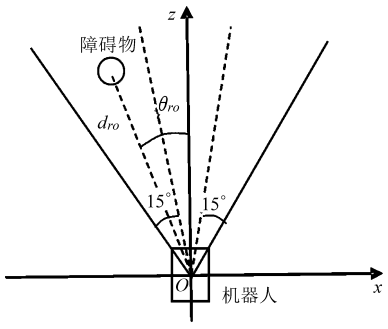

图 2 移动机器人坐标系

Fig. 2 Coordinate system of mobile robot

习、非监督学习和强化学习 3 大类。其中强化学习是一种实时在线的学习方法,通过试错方法不断获得经验知识,然后根据这些知识来完善行动策略进而完成目标任务[10-11]。

2.1 强化学习算法

Q 学习算法 1992 年由 Watkins 提出,是强化学习中最为重要的算法之一,这里在此基础上对动态环境中移动机器人的自主导航进行了研究。该算法可以让智能系统拥有在 Markov 决策过程中利用经历的动作序列选择最优动作的能力,不需要建立环境模型。

Q 学习算法的目的是,在状态转移概率和所获得奖惩未知的情况下来估计最优策略的 Q 值。Q 值函数是在环境状态 s_t 时执行动作 a_t 的评价函数,与此后按最优动作序列执行时得到的强化信号折扣的和,即

$$Q(s_t,a_t)=r_{t+1}+\gamma\max_{a\in A}Q(s_{t+1},a) \tag{1}$$

式中 r_{t+1}——在状态 s_t 执行动作 a_t 到达状态 s_{t+1} 时得到的瞬时奖惩值

γ——折扣率,保证返回的奖惩是有限的,$\gamma\in[0,1]$

A——状态 s_{t+1} 时可执行的动作集

式(1)只有在得到最优策略的前提下才成立,而在学习阶段两边误差为

$$\Delta Q(s_t,a_t)=r_{t+1}+\gamma\max_{a\in A}Q(s_{t+1},a)-Q(s_t,a_t) \tag{2}$$

所以 Q 学习的更新规则为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\Delta Q(s_t,a_t)$$

整理后

$$Q(s_t,a_t)\leftarrow(1-\alpha)Q(s_t,a_t)+\alpha(r_{t+1}+\gamma\max_{a\in A}Q(s_{t+1},a))$$

式中 α——学习率

α 控制着学习的速度,α 越大则收敛越快,但过大的 α 可能引起不成熟收敛。

Q 学习算法中,最基本的形式为单步学习。学习过程中,首先观察现在的状态 s_t,选择并执行一个动作 a_t;然后观察下一个状态 s_{t+1},收到一个瞬时奖惩信号 r_{t+1},根据整理后的 Q 学习更新规则更新 Q 值;最后进入下一时刻。

与单步学习相对应的是多步学习,区别在于它假定在当前环境状态 s_t 选取动作 a_t 后,进入下一个状态 s_{t+1},此时得到的瞬时奖惩信号 r_{t+1} 不立即进行 Q 值更新,而是根据策略从状态 s_{t+1} 的动作中选取 a_{t+1},并执行它进入状态 s_{t+2},得到回报 r_{t+2}。依此类推,可以得到 n 步 Q 学习算法,这时 Q 值更新规则为

$$Q(s_t,a_t)\leftarrow(1-\alpha)Q(s_t,a_t)+\alpha[r_{t+1}+\gamma r_{t+2}+\cdots+\gamma^{n-1}r_{t+n}+\gamma^n\max_{a\in A}Q(s_{t+n},a)]$$

2.2 基于模糊逻辑的环境状态离散化

在实际应用中,大多数系统本身的输入输出量是连续的状态空间。但常用的 Q 学习算法采用的是基于查找表的结构,不易处理连续状态空间问题,且易出现维数爆炸问题。为了使 Q 学习算法能够在实际移动机器人平台上顺利运行,避免因连续状态空间导致维数灾难问题,这里采用模糊逻辑对移动机器人的输入状态信息进行离散化处理。

在双目立体视觉的 40°视野中,障碍物相对于移动机器人的方向信息是一个连续变化量,如图 2 所示,因此利用模糊逻辑方法将它划分为{左侧,中间,右侧},具体的模糊隶属度函数如图 3a 所示。同样,在双目立体视觉 7 m 有效视距内,将障碍物相对于移动机器人的距离信息模糊划分为{冲突,危险,威胁,安全},对应的模糊隶属度函数如图 3b 所示。

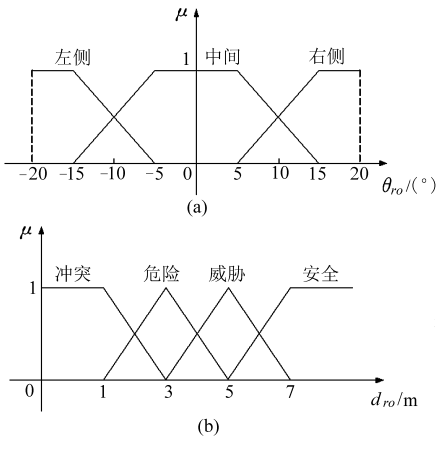

图 3 模糊隶属度函数

Fig. 3 Functions of fuzzy membership

(a) 方向 (b) 距离

根据上述方向信息和距离信息确定移动机器人在导航环境中所处的状态,如表 1 所示。例如当距离信息为危险且方向信息为左侧时,环境状态定义

为 s2。这样总共可以得到 s1 ~ s8 等 8 种有限的环境状态，避免强化学习时出现维数爆炸问题。

表 1　环境状态定义

Tab. 1　Definition of environment states

方向信息	距离信息			
	冲突	危险	威胁	安全
左侧	s1	s2	s5	s8
中间	s1	s3	s6	s8
右侧	s1	s4	s7	s8

2.3　强化学习导航动作选择方法

当移动机器人处于某一状态时，它需要从动作集中抽取一个对其有利的动作执行。一方面，移动机器人需要尽可能地探索不同的动作，以找到最优的策略；另一方面，又要考虑选择可以得到奖赏值最大的动作。探索对学习是非常重要的，只有通过探索才能确定最优策略，而过多的探索会降低系统的性能，影响学习速度。

当移动机器人视野中存在障碍物并构成威胁时，机器人采用贪心策略中的 Boltzmann 策略[11]选择动作进行避障，动作集中包括直行、左转和右转。相对于其他策略的收敛速度，Boltzmann 探索策略可以更快地对算法进行收敛。Boltzmann 探索策略指定了当前状态 s 下执行动作 a_i 的概率为

$$P(s,a_i)=\frac{e^{\frac{Q(s,a_i)}{T}}}{\sum_{a_k\in A}e^{\frac{Q(s,a_k)}{T}}} \tag{3}$$

$$T=T_0 n^{-\beta}+1 \tag{4}$$

式中　T——温度参数

n——学习循环次数

T_0、β——调节参数，分别取 2 和 0.5

这样，在学习的初始阶段，较大的 T 值减小了不同动作的入选概率差异，动作选择倾向于随机选择，以便探索出较好的导航避障动作。随着学习的继续，T 逐渐变小，此时动作选择倾向于 $P(s,a_i)$ 最大的动作，保证以前的学习成果不被破坏。

根据状态定义及可选择动作定义 Q 值表，如表 2 所示。试验开始时，赋初值为零，随着学习训练次数的增加，某一状态下某一最优动作的 Q 值逐渐增大，在导航控制时被选中的概率也就增大。特别指出的是，当移动机器人与障碍物发生冲突时，不选择任一动作而是直接停止运动，因此对应状态 s1 的 Q 值始终不进行更新。

2.4　奖惩值函数

根据状态转移，设定瞬时奖惩值。当移动机器人进入某一状态时，设定奖惩值为

表 2　Q 值

Tab. 2　Q value

状态	动作		
	直行	左转	右转
s1	$Q(s_1,a_1)$	$Q(s_1,a_2)$	$Q(s_1,a_3)$
s2	$Q(s_2,a_1)$	$Q(s_2,a_2)$	$Q(s_2,a_3)$
s3	$Q(s_3,a_1)$	$Q(s_3,a_2)$	$Q(s_3,a_3)$
s4	$Q(s_4,a_1)$	$Q(s_4,a_2)$	$Q(s_4,a_3)$
s5	$Q(s_5,a_1)$	$Q(s_5,a_2)$	$Q(s_5,a_3)$
s6	$Q(s_6,a_1)$	$Q(s_6,a_2)$	$Q(s_6,a_3)$
s7	$Q(s_7,a_1)$	$Q(s_7,a_2)$	$Q(s_7,a_3)$
s8	$Q(s_8,a_1)$	$Q(s_8,a_2)$	$Q(s_8,a_3)$

$$r=\frac{c_1\mu_1+c_2\mu_2+c_3\mu_3+c_4\mu_4}{\mu_1+\mu_2+\mu_3+\mu_4}+(\mu_1+\mu_4-1)\frac{c_5\mu_5+c_6\mu_6+c_7\mu_7}{\mu_5+\mu_6+\mu_7} \tag{5}$$

式中　$c_1 \sim c_7$——常数

μ_1、μ_2、μ_3、μ_4——冲突、危险、威胁和安全的模糊隶属度

μ_5、μ_6、μ_7——左侧、中间和右侧的模糊隶属度

当移动机器人进入状态 s1 ~ s4 时，$\mu_4\equiv 0$；当移动机器人进入状态 s5 ~ s8 时，$\mu_1\equiv 0$。分别定义 $d_{ro}(t)$、$d_{ro}(t+1)$ 为 t、$t+1$ 时刻移动机器人与障碍物的距离，根据移动机器人到障碍物的距离变化设定式(5)中 $c_1 \sim c_7$ 的数值，即可得到强化学习算法的瞬时奖惩值，如表 3 所示。

设定学习率 $\alpha=0.5$，折扣率 $\gamma=0.8$，强化学习算法的值函数根据整理后的 Q 学习规则更新 Q 值。采用 Visual C + + 编写算法程序及人机交互界面，移动机器人在自主控制模式下时，每隔 0.2 s 完成一次状态信息获取和控制信号发送，进行导航控制。

3　试验及结果分析

为了充分验证结合强化学习与模糊逻辑的自主导航控制方法应用于移动机器人未知环境下进行探索并完成相应导航任务的有效性，基于图 1 所示的移动机器人平台进行了室外试验。试验通过双目立体视觉检测地平面上物体的高度来判断导航环境中是否存在有效的障碍物，所有高度不对移动机器人的运动构成威胁的物体全部视为无效的障碍物。图 4a 是导航环境中无有效障碍物时的情况，图 4b 中的深色物体为障碍物，机器人必须要进行避让。

设定移动机器人自主导航控制总体任务是：当移动机器人与环境中的障碍物发生冲突时，为了安全考虑，移动机器人立即停止运动；当环境中不存在障碍物、障碍物不构成威胁或者已经完成避障时，移

表 3 奖惩值

Tab. 3 Rewards and punishments value

状态转移		奖惩值
t 时刻	$t+1$ 时刻	
s1 ~ s4	s1 ~ s4	$d_{ro}(t+1) \geqslant d_{ro}(t)$: $\dfrac{-2\mu_1+\mu_2+2\mu_3}{\mu_1+\mu_2+\mu_3}+(\mu_1-1)\dfrac{\mu_6}{\mu_5+\mu_6+\mu_7}$ $d_{ro}(t+1) < d_{ro}(t)$: $\dfrac{-2\mu_1+\mu_3}{\mu_1+\mu_2+\mu_3}+(\mu_1-1)\dfrac{\mu_5+2\mu_6+\mu_7}{\mu_5+\mu_6+\mu_7}$
	s5 ~ s8	$\dfrac{\mu_2+2\mu_3+2\mu_4}{\mu_2+\mu_3+\mu_4}+(\mu_4-1)\dfrac{\mu_6}{\mu_5+\mu_6+\mu_7}$
s5 ~ s8	s5 ~ s8	$d_{ro}(t+1) \geqslant d_{ro}(t)$: $\dfrac{\mu_2+2\mu_3+2\mu_4}{\mu_2+\mu_3+\mu_4}+(\mu_4-1)\dfrac{\mu_6}{\mu_5+\mu_6+\mu_7}$ $d_{ro}(t+1) < d_{ro}(t)$: $\dfrac{\mu_3+2\mu_4}{\mu_2+\mu_3+\mu_4}+(\mu_4-1)\dfrac{\mu_5+2\mu_6+\mu_7}{\mu_5+\mu_6+\mu_7}$
	s1 ~ s4	$\dfrac{-2\mu_1+\mu_3}{\mu_1+\mu_2+\mu_3}+(\mu_1-1)\dfrac{\mu_5+2\mu_6+\mu_7}{\mu_5+\mu_6+\mu_7}$

图 4 基于强化学习的导航试验场景

Fig. 4 Experiment environment of navigation based on reinforcement learning

(a) 机器人视野中无有效障碍物 (b) 机器人视野中存在障碍物

动机器人保持或恢复初始状态;当环境中障碍物的运动对移动机器人运动构成威胁时,采用自学习得到的导航策略进行安全避障。前两者比较简单,机器人可以根据预先设定导航规则动作,本文提出的强化学习方法重点针对上述的第 3 个导航任务提高机器人自主导航控制的智能水平。

图 5 为基于强化学习的导航试验结果。可以看出,当环境中不存在障碍物时,移动机器人保持初始运动状态运行。当然,由于地面不平干扰、机器人自身机构误差等原因必然会有时导致航向发生偏移,这在图 5a 的轨迹曲线中有明确反映。但移动机器人会根据光电编码器的反馈信息进行自平衡调整,

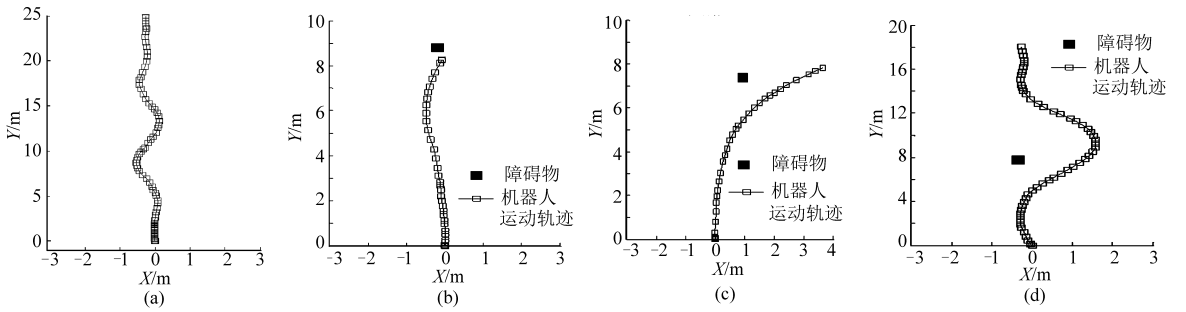

图 5 导航试验结果

Fig. 5 Experiment result of navigation based on reinforcement learning

(a) 无有效障碍物时移动机器人运动轨迹 (b) 强化学习初期的导航运动轨迹

(c) 强化学习中期的导航运动轨迹 (d) 强化学习后期的导航运动轨迹

恢复至初始运动状态继续运行，满足导航任务的要求。

当环境中存在有效障碍物时，移动机器人根据双目立体视觉获取障碍物在移动机器人坐标系下的坐标，实时计算移动机器人与障碍物的距离以及方向信息，根据前述方法判断移动机器人所处的环境状态，调用 Q 学习算法，根据 Boltzmann 策略选择动作，并下发控制信号给执行机构，进行避障。开始时机器人没有任何经验，Q 值表初始全部赋值为零。

根据光电编码器的反馈信息可以计算出移动机器人在有障碍环境场景中强化学习各个阶段的运动轨迹，如图 5b ~ 5d 所示。可以看出，在初期试验中，由于移动机器人对环境状态没有任何先验知识，机器人与障碍物发生了干涉（图 5b）。这时得到的 Q 值表如表 4 所示，可见不同的状态下，不同的动作已经得到了不同的奖惩值。在学习中期的试验中，机器人已经能够完成避障动作，但由于右转角度过大，最终导致在恢复预定导航路径时失败（图 5c），没有实现预定的导航任务，此时的 Q 值如表 5 所示。在后期试验中，移动机器人已经在前面试验中进行了大量的探索性尝试，通过不断更新 Q 值表使得移动机器人对环境状态有了更为充分的经验，逐步积累了完成预定导航任务所需的知识，完成了自主导航任务（图 5d）。最后得到的 Q 值表如表 6 所示，也即是这种环境下机器人通过自主学习获取的完成预定导航任务的最佳策略知识表达。

表 4　初期时 Q 值

Tab. 4　Q value at early stage of training

状态	动作		
	直行	左转	右转
s1	0	0	0
s2	-0.675 0	-0.473 0	1.033 1
s3	-0.042 9	-0.225 2	3.185 2
s4	-0.473 2	0.828 7	-0.35 50
s5	0.137 0	0.177 5	3.730 7
s6	1.708 0	4.305 5	6.128 0
s7	0.952 0	3.809 0	2.482 3
s8	1.403 8	1.313 2	3.782 2

表 5　中期时 Q 值

Tab. 5　Q value at medium stage of training

状态	动作		
	直行	左转	右转
s1	0	0	0
s2	0.526 6	-0.458 7	1.085 1
s3	-0.956 8	2.968 7	4.509 4
s4	0.628 8	1.368 9	-0.815 0
s5	2.056 9	0.177 5	3.230 7
s6	2.661 8	6.166 8	5.172 7
s7	2.461 2	5.857 7	1.723 0
s8	1.684 2	1.780 7	5.436 7

表 6　后期时 Q 值

Tab. 6　Q value at final stage of training

状态	动作		
	直行	左转	右转
s1	0	0	0
s2	0.526 6	-0.458 7	1.085 1
s3	-0.956 8	2.968 7	4.509 4
s4	0.628 8	1.368 9	-0.815
s5	2.056 9	0.177 5	6.230 7
s6	2.661 8	6.855 6	5.172 7
s7	2.461 2	5.857 7	1.723 0
s8	2.024 8	2.654 5	6.972 0

4　结束语

在强化学习基础上，结合模糊逻辑设计了移动机器人自主导航策略自学习方法，使机器人能够在未知环境中不断积累完成预定任务的知识，自动探索较好的解决问题策略。在自制的轮式移动机器人平台上开展了试验，结果表明机器人经过多轮学习以后，可以不断自主获取并积累导航避障知识，完成了事前给定的导航任务，算法能够较好地满足移动机器人导航的实时性要求。该算法可以增强农业机器人适应动态开放的农业环境的能力。

参　考　文　献

1　刘国栋，谢宏斌，李春光. 动态环境中基于遗传算法的移动机器人路径规划的方法[J]. 机器人，2003，25(4)：327 - 330.
Liu Guodong, Xie Hongbin, Li Chunguang. Method of mobile robot path planning in dynamic environment based on genetic algorithm [J]. Robot, 2003, 25(4): 327 - 330. (in Chinese)

2　Lu Y, Yi X Y, Cheng J L. A new potential field method for mobile robot path planning in the dynamic environments [J]. Asian Journal of Control, 2009, 11(2): 214 - 225.

3　肖本贤，齐东流，刘海霞，等. 动态环境中基于模糊神经网络的 AGV 路径规划[J]. 系统仿真学报，2006，18(9)：2 401 - 2 404.
Xiao Benxian, Qi Dongliu, Liu Haixia, et al. AGV path planning in the dynamic environment based on fuzzy neural network[J]. Journal of System Simulation, 2006, 18(9): 2 401 - 2 404. (in Chinese)

4　Wang M, James N, Liu K. Fuzzy logic-based real-time robot navigation in unknown environment with dead ends[J]. Robotics and

Autonomous Systems, 2008, 56 (7): 625 – 643.
5 刘宏林,罗杨宇,李成荣. 基于模糊控制器的未知环境下移动机器人导航[J]. 计算机仿真,2011,28(1):201 – 205.
Liu Honglin, Luo Yangyu, Li Chengrong. Mobile robot navigation based on fuzzy controller in unknown environment[J]. Computer Simulation, 2011, 28(1):201 – 205. (in Chinese)
6 Wang H J, Xiong W. Research on global path planning based on ant colony optimization for AUV [J]. Journal of Marine Science and Application, 2009,8(1):58 – 64.
7 Lewis F L, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control [J]. Circuits and Systems Magazine, IEEE, 2009,9(3):32 – 50.
8 杨福增, 刘珊, 陈丽萍,等. 基于立体视觉技术的多种农田障碍物检测方法[J]. 农业机械学报, 2012, 43(5): 168 – 202.
Yang Fuzeng, Liu Shan, Chen Liping, et al. Detection method of various obstacles in farmland based on stereovision technology [J]. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(5): 168 – 202. (in Chinese)
9 周俊,程嘉煜. 基于机器视觉的农业机器人运动障碍目标检测[J]. 农业机械学报,2011,42(8):154 – 158.
Zhou Jun, Cheng Jiayu. Moving obstacle detection based on machine vision for agricultural mobile robot [J]. Transactions of the Chinese Society for Agricultural Machinery, 2011, 42(8):154 – 158. (in Chinese)
10 高阳,陈世福,陆鑫. 强化学习研究综述[J]. 自动化学报,2004,30(1):86 – 100.
Gao Yang, Chen Shifu, Lu Xin. Research review on the navigation for outdoor agricultural robot [J]. Acta Automatica Sinca, 2004, 30(1): 86 – 100. (in Chinese)
11 Duan Y, Xu X H. Fuzzy reinforcement learning and its application in robot navigation [C] // Proceedings of 2005 IEEE/ International Conference on Machine Learning and Cybernetics, 2005,2: 899 – 904.

Vision Navigation of Agricultural Mobile Robot Based on Reinforcement Learning

Zhou Jun Chen Qin Liang Quan
(*Jiangsu Key Laboratory for Intelligent Agricultural Equipments*, *Nanjing Agricultural University*, *Nanjing* 210031, *China*)

Abstract: The method that agricultural mobile robot acquire the navigation strategies through autonomous learning was development based on reinforcement learning and fuzzy logic. Firstly, the machine vision was applied to detect obstacles in the navigation environment, and the corresponding direction and distance between the robot and the obstacle was calculated. Then the algorithm of acquiring the more optimal navigation strategies was introduced with the reinforcement learning, so the capability of the mobile robot of adapting the dynamic navigation environment was improved. Finally, the continuous values of the direction and the distance between the obstacles and the mobile robot were discretized with the fuzzy logic rules, and the discrete navigation environment states were obtained, then the Q value table was designed for the reinforcement learning. The experiment was carried out with the wheeled mobile robot, and the experimental results showed that the mobile robot was able to automatically acquire more optimal navigation strategies in the actual environment, and fulfill the expected navigation tasks.

Key words: Agricultural robot Reinforcement learning Fuzzy logic Vision navigation