**DOI**:10.11992/tis.201510031
**网络出版地址**:http://www.cnki.net/kcms/detail/23.1538.TP.20160315.1252.020.html

# 基于强化学习的多定位组件自动选择方法

梁爽[1],曹其新[1],王雯珊[1],邹风山[2]
(1.上海交通大学 机器人研究所,上海 200240; 2.新松机器人有限公司 中央研究院,辽宁 沈阳 110000)

**摘　要**:在一个大规模的动态环境中,针对机器人各种定位传感器的局限性,提出了一种基于强化学习的定位组件自动选择方法。系统采用分布式架构,将机器人不同的定位传感器与定位方法封装为不同的组件。采用强化学习的方法,寻找最优策略,实现多定位组件的实时切换。仿真结果表明,该方法可以解决大型环境中,单一定位方法不能适用于整个环境的问题,能够依靠多定位组件提供可靠的机器人定位信息;环境发生改变时,通过学习的方法不需要重新配置组件,且与直接遍历组件后切换组件的方法相比,极大地减小了延时。

**关键词**:移动机器人;定位;强化学习;中间件:Monte Carlo 方法;多传感器;模块化;分布式系统

**中图分类号**: TP242.6　**文献标志码**:A　**文章编号**:1673-4785(2016)02-0149-07

**中文引用格式**:梁爽,曹其新,王雯珊,等. 基于强化学习的多定位组件自动选择方法[J]. 智能系统学报, 2016, 11(2): 149-154.
**英文引用格式**:LIANG Shuang, CAO Qixin, WANG Wenshan, et al. An automatic switching method for multiple location components based on reinforcement learning[J]. CAAI transactions on intelligent systems, 2016, 11(2): 149-154.

# An automatic switching method for multiple location components based on reinforcement learning

LIANG Shuang[1], CAO Qixin[1], WANG Wenshan[1], ZOU Fengshan[2]
(1. Research Institute of Robotics, Shanghai Jiaotong University, Shanghai 200240, China; 2. SIASUN Robot and Automation CO., LTD, Shenyang 110000, China)

**Abstract**: To address the limitations of location sensors in large-scale dynamic environments, an automatic switching method for multiple robotic components based on reinforcement learning is proposed. This system uses distributed architecture and encapsulates different location sensors and methods into different middleware components. Reinforcement learning is employed to find the optimal strategy for deciding how to switch between components in real time. The simulation result shows that this method can solve problems that a single location method cannot in a large-scale environment and can provide reliable location information depending on multiple location components. This method can also effectively reduce the time delay compared with a method that first traverses all the components directly and then switches components.

**Keywords**: mobile robot; location; reinforcement learning; middleware; Monte Carlo; multi-sensor; modularization; distributed system

定位是移动机器人最重要的功能之一,其目的在于确定机器人在环境中的位置。定位是指利用先验环境地图信息、机器人位姿的当前估计与传感器获取的观测值,经过一系列的处理和转换,产生更加准确的对机器人当前位姿的估计[1]。机器人的定位方式与机器人使用的传感器有关。目前,在移动机器人上使用较多的传感器有里程计、摄像头、激光雷达、超声波等。不同的传感器有不同的优缺点,如

**收稿日期**:2015-10-29.　**网络出版日期**:2016-03-15.
**基金项目**:国家自然科学基金项目(61273331).
**通信作者**:曹其新. E-mail:qxcao@sjtu.edu.cn.

视觉传感器信息量大、感应时间短，但获得的数据噪声大、信息处理时间长，且定位目标容易受背景与光照条件的影响[2]；激光传感器在测距范围和方向上具有较高的精度，但价格昂贵；超声波传感器处理信息简单、成本低、速度快，但角度分辨力较低。各定位方法一般都有特定的工作环境要求，单一的定位方法很难适用于大规模动态环境。为了提高机器人的定位精度，常常采用多个传感器结合的方式。本文提出一种模块化开发的架构，将几个不同的定位系统模块化，采用强化学习方法，通过对系统的试错-奖励寻找最优策略，从而在动态环境中自动选择可靠的定位模块，不同的定位模块之间可以动态地切换、自由组合，以弥补各自的不足。

## 1 组件化系统架构

本文采用 RT 中间件技术，将机器人各个功能模块及相应的算法封装成不同的 RT 组件。RT 中间件通过将不同的机器人组件连接起来，从而构建组件分布式的系统以增强机器人系统的灵活性与软件的重用性[3]。RT 中间件根据生命周期来定义 RT 组件的动作，如图 1 所示，RT 组件在被创建、未激活、激活、错误几个状态间切换，通过执行不同的状态对应的功能函数，来实现组件的功能。系统的每个组件都有输入与输出端口，用以和其他组件进行数据传输。

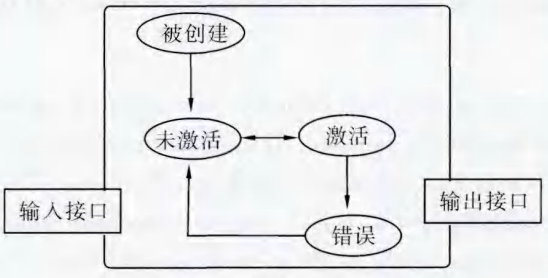


**图 1 RT 组件的结构**

**Fig.1 RT-Component Architecture**

本文创建 7 个组件：室内摄像头定位组件、走廊摄像头定位组件、室外摄像头定位组件、激光传感器定位组件、定位选择组件、电机驱动组件与机器人导航组件，各组件之间通过组件接口进行数据传输。分布式的机器人系统如图 2 所示，机器人本体安装有一个激光传感器，环境中分布安装有个环境摄像头，每个传感器模块及算法都采用中间件技术进行封装，保证各个传感器组件之间无影响，定位选择组件的输入接口可以连接到不同定位组件输出接口，从而选择相应的定位组件。

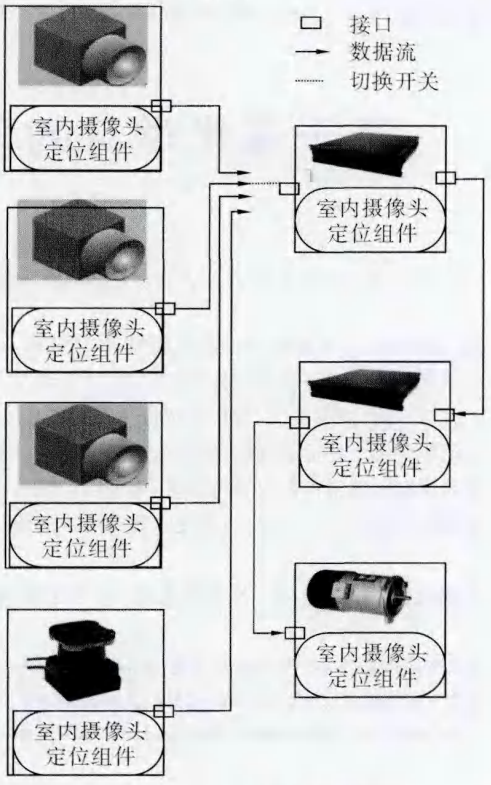


**图 2 模块化的机器人系统**

**Fig.2 Modular robotic system**

### 1.1 基于激光测距仪的定位方法

激光测距仪是一种高精度、高解析度外部传感器[5]。本文使用激光测距仪感知机器人所处的环境信息，采用 MCL（Monte Carlo localization）定位方法，并使用 ROS（robot operating system）中的 AMCL 包完成基于激光测距仪的定位[6]。

MCL 方法是一种基于贝叶斯滤波的概率估计自定位方法。它使用一系列带有权值的粒子来表示机器人在运行环境中的可能位置。权值越大意味着粒子代表的位置与机器人实际位置的匹配程度越高。MCL 算法首先根据机器人当前的运动状态，对代表机器人位置的粒子进行移动，由里程计运动信息 $u_{t-1}$ 和前一时刻的机器人位姿 $l_{t-1}=(x_{t-1},y_{t-1},\theta_{t-1})$ 估计次粒子的位置；之后根据当前位置传感器得到的测量信息 $z_t$ 计算粒子权值；接着更新粒子分布，根据粒子权值的大小来复制粒子，权值较小的粒子即被筛除。最后根据粒子的权值及其分布情况来计算机器人的位置，在粒子最集中的地方，选取权值最大的粒子作为机器人的位置估计[7]。从而得到当前位置的估计值 $(x,y,z)$ 与相应的协方差矩阵 $\boldsymbol{C}$：

$$C = \begin{bmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{bmatrix} \quad (1)$$

二维平面中 $\mathrm{cov}(x,x)$、$\mathrm{cov}(y,y)$ 分别代表 $x$ 与 $y$ 方向上的方差。

### 1.2 基于环境摄像头的定位方法

机器人在摄像头视野内运动时,可由摄像头得到机器人的位置信息,本文采用颜色跟踪的方法,用 CamShift 算法实现机器人跟踪定位。CamShift 跟踪算法是以颜色直方图为目标模式的目标跟踪算法,它具有较高的运算效率[8]。它的基本思想是首先建立起被跟踪目标位置的颜色概率模型。选择出示搜索窗口的大小和位置后,统计出该区域内每个像素点的直方图,再将此直方图作为概率查找表。在目标跟踪的过程中,针对摄像头所得图像的每一个像素,查找生成的概率查找表,从而将视频图像转化为目标颜色概率分布图,同时用当前帧定位的结果来预测下一帧图像中目标的中心和大小。

对得到的结果进行卡尔曼滤波,同样得到当前位置的估计值 $(x,y,z)$ 与相应的协方差矩阵 $\boldsymbol{C}$。

## 2 基于 Monte Carlo 的强化学习算法

强化学习是一种基于试错机制的学习方法。所谓强化学习就是智能系统从环境到行为映射的学习,以使动作累积得到的奖励信号值最大,从而寻找到一个最优策略。强化学习并不直接告诉强化学习系统该如何产生正确或最优的动作,而是依靠自身的经历学习,通过学习,在行动的环境中获得知识,改进行动方案以适应环境[9]。

定义系统状态集合为 $S$,组件选择动作集合 $A$。在任意时刻 $t$,系统从当前状态状态 $s_t \in S$ 选择下一步执行动作 $A(s_t)$,系统变迁到新的状态 $s_{t+1} \in S$ 并获得相应的奖惩值 $r_t$。定义目标函数为从初始状态到目标状态后累积得到的奖励值。强化学习的目的就在于搜索出一个最优策略 $\pi^*: S \rightarrow A$,使得目标函数的值最大,即 $\pi^*(s) = \max_a \pi^a(s)$ [10]。强化学习的框图如图 3 所示。

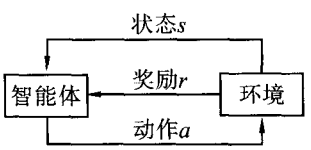


**图 3　强化学习的结构**

**Fig.3　The structure of reinforcement learning**

Monte Carlo 方法是通过估计值函数来学习的一种常用的强化学习算法,与其他方法相比,它不需要环境的完整模型,只需要与环境在线或者仿真下交互得到的状态、动作和奖励的采样序列。Monte Carlo 的具体算法如下[11]:

1) 初始化强化学习系统的 $Q(s,a)$ 与策略 $\pi(s)$:对于所有的 $s \in S$, $a \in A(s)$, $Q(s,a)$ 与 $\pi(s)$ 取任意值;

2) 初始化移动机器人的状态;

3) 根据状态信息与 $\pi(s)$,生成一个动作;

4) 执行动作使机器人到达一个新的位置,同时获得相应的奖励值,并添加到 $\mathrm{Returns}(s,a)$;

5) 根据奖励值更新 $Q(s,a)$, $Q(s,a) = \mathrm{average}(\mathrm{Returns}(s,a))$;

6) 选择最大 $Q$ 值对应的动作 $\pi(s) = \arg\max_a Q(s,a)$;

7) 重复 3) ~6)。

## 3 基于强化学习的定位组件选择方法

由于不同环境下不同定位组件的定位效果不同,需要在不同的定位方法间动态地切换、自由组合,来弥补各自的不足。例如走廊上由于几何特征不丰富,因此激光定位效果不佳,需要切换为走廊环境摄像头定位。

现有的方法采用简单直观的贪心思想,即总是选择可靠度最高的那个组件[12]。由于在线切换需要遍历所有定位组件,效率低下。网络可能存在较大延迟,每连接一个组件需要几百毫秒甚至两三秒的时间。在大规模环境中在线切换算法会造成很大的延迟。同时离线配置每个组件的定位范围,工作量大,不能适应动态情况。离线配置需要各组件在不同区域的定位精度有一个可靠的估计,在定位组件众多、环境复杂的情况下很难配置。因此当环境改变,例如移动家具,光线条件变化等,组件定位效果会发生变化,造成离线设定的失败。

为了更加缩短组件间的切换时间,同时实现机器人的智能化,本文提出了一种基于强化学习的定位组件自动选择方法,不用手动设定各组件的定位范围,而是让机器人在环境中运动一段时间,自己找出最优的定位方案。

### 3.1 机器人状态与动作空间的描述

定义状态 $s(x,y,c,d) \in S$,其中 $(x,y)$ 表示机

器人当前坐标，$c \in C$，$C = \{1,2,3,4\}$ 表示当前使用的定位组件，共有室内摄像头定位组件、走廊摄像头定位组件、室外摄像头定位组件、激光传感器定位组件 4 种组件。$d \in D$，$D = \{1,2,3,4,5,6,7,8\}$ 表示机器人运动方向，即在世界坐标系下沿 -135°、-90°、-45°、0°、45°、90°、135°、180°前进，动作 $a \in A$ 表示切换组件的动作。

### 3.2　强化信号的确定

协方差矩阵可以用来判断定位信息是否可靠，本文通过设定阈值的方法来判断定位是否有效。定位方差过大，此时组件定位无效，令奖惩函数 $R(s,a)$ 表示在状态 $s$ 下执行动作 $a$ 后获得的奖惩值，此时 $R(s,a)$ 返回 -5。若定位方差较小，说明切换到有效的定位组件，则 $R(s,a)$ 返回 0。

期望函数 $Q(s,a) = E(\sum R(s_i,a_i))$ 表示在状态 $s$ 下执行动作 $a$ 的奖惩值的期望。策略 $\pi^a(s)$ 表示采取的一系列动作的集合，在强化学习收敛以后，最优策略 $\pi^*(s)$ 就是针对每一个状态 $s$，选择一个使 $Q$ 函数值最大的动作 $a$。

在选择定位组件问题中，初始时可以使用一个随机选择的策略。机器人在环境中按照当前策略运行，运行一段时间后即根据奖励和惩罚，修改选择策略，使得奖励值的期望值最大化，策略将收敛到最优策略。

## 4　仿真与实验结果

为了验证本方案，在 MATLAB 平台上进行了仿真实验。仿真环境中选用了两种传感器：激光传感器和摄像头传感器，其中激光传感器在特征点较多的房间内定位精度较高，而在特征点较为相似的走廊中定位精度很差；由于摄像头定位能力局限于其安装位置，因此 3 个摄像头有 3 个不同的有效定位区域：室内摄像头区域、走廊摄像头区域、室外摄像头区域。图 4 中的 4 幅图代表了不同传感器的有效定位区域，黑色部分代表本传感器的无效定位区域。

实验中，设置机器人的起始位置后，系统随机选择房间中的一点作为机器人的目标点，机器人通过强化学习算法判断所经过的区域内选择何种传感器可以保证最优策略。实验测试了选取不同迭代次数时本算法所得到的结果。

图 5 所示为不同迭代次数下，本算法对于定位组件的选择，横坐标与纵坐标分别代表机器人所处位置的横坐标与纵坐标，不同灰度的方块代表当前位置本算法所选择的定位组件。可以看到，迭代次数越大，定位方法的选择越符合最优策略，当迭代次数为 50 000 时，与图 4 对比可看出，不同组件所在的位置都处于组件的有效定位区域内，此时系统可以自动地切换到有效的定位组件。

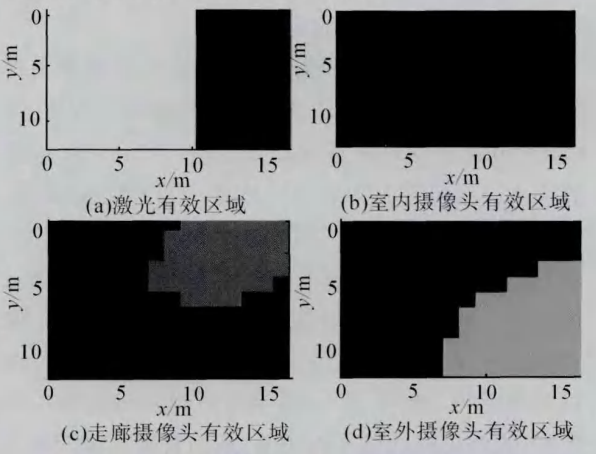


**图 4　环境内各传感器的有效定位区域**

**Fig.4　Effective area of each sensor in the environment**

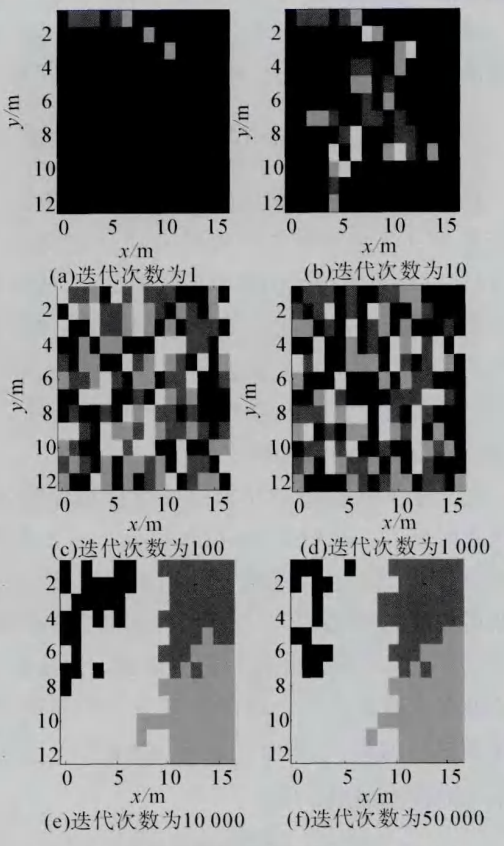


**图 5　迭代次数分别为 1、10、100、1 000、10 000、50 000 时本算法对定位区域的选择**

**Fig.5　The choice of localization components when the iterations is 1, 10, 100, 1 000, 10 000, 50 000**

为了测试系统的鲁棒性,将室内环境进行了小范围改变,使得激光有效定位区域发生了改变,此时各传感器的有效区域如图 6 所示。不同迭代次数时,本算法对于定位组件的选择如图 7 所示。

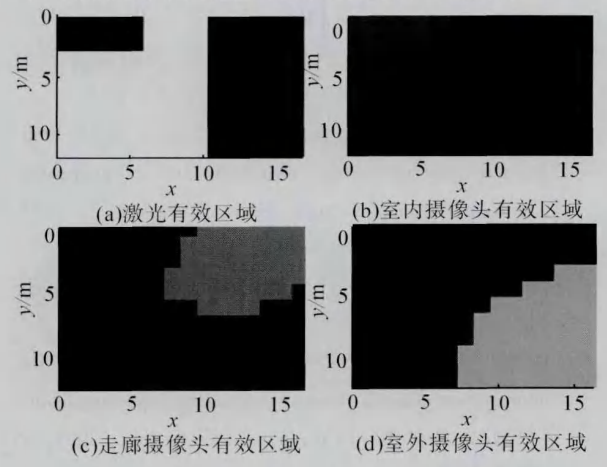


**图 6　环境改变后各传感器的有效定位区域**

**Fig.6　Effective area of each sensor when the environment is changed**

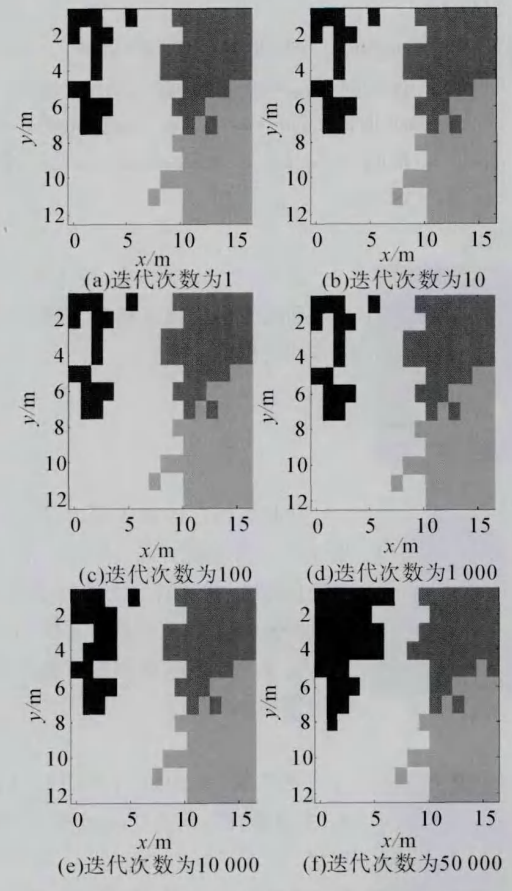


**图 7　迭代次数分别为 1、10、100、1 000、10 000、50 000 时本算法对定位区域的选择**

**Fig.7　The choice of localization components when the iterations is 1, 10, 100, 1 000, 10 000, 50 000**

与图 6 对比可以看到,在环境特征发生改变的区域内(左上角部分),机器人可以通过学习,将定位组件更改为有效的室内摄像头,提高了系统的鲁棒性。

将仿真中收敛后的结果在机器人上进行了测试,测试结果如图 8 所示,当机器人位于室内(室内摄像头视野外和室内摄像头视野内)、室外、走廊中时,机器人定位选择组件可以正确地切换到相应的定位组件。

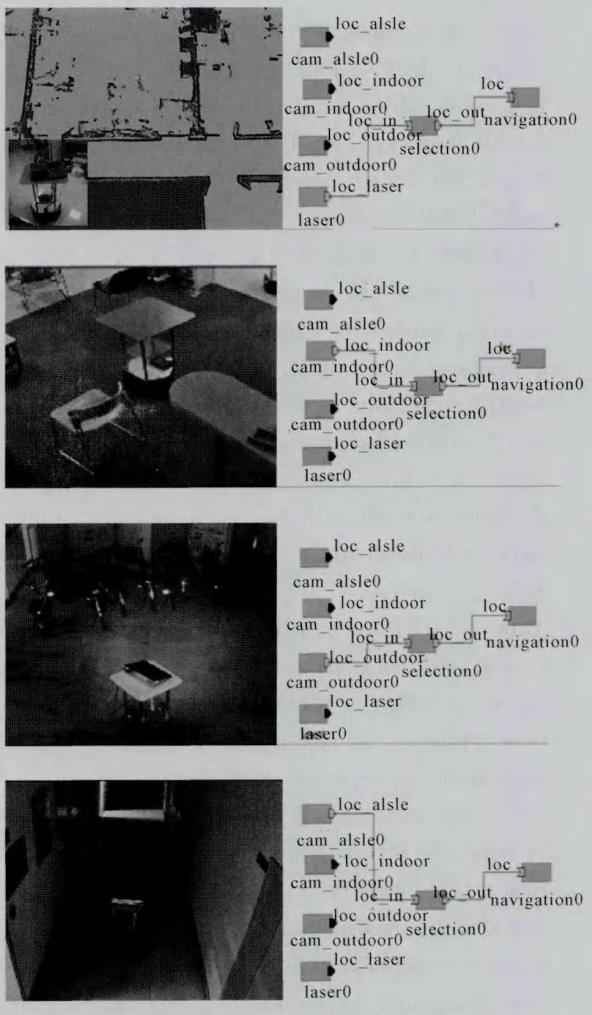


**图 8　实际环境中不同位置机器人定位组件的选择**

**Fig.8　The selection of localization component in different positions**

## 5　结束语

本文提出了一种基于强化学习的大规模动态环境中多定位组件自动选择方法,实验结果表明移动机器人到达未知环境时,在环境中运行一段时间后,可以成功地根据所处的环境切换到较为可靠的传感器。通过强化学习的方法,机器人在选择定位组件

的时候，直接查找 $Q$ 函数值最大的组件，而不是再遍历所有的组件，耗时从 $O(n)$ 减小到 $O(1)$，大大减少了切换的延迟。组件化的结构可以极大地提高系统的扩展性。

实验表明多次迭代后，系统可以收敛，今后的工作将着重放在进一步优化奖惩函数，以减少所需的迭代次数上。

## 参考文献：


[1]李群明，熊蓉，褚健. 室内自主移动机器人定位方法研究综述[J]. 机器人，2003，25(6)：560-567，573.
LI Qunming，XIONG Rong，CHU Jian. Localization approaches for indoor autonomous mobile robots：A review[J]. Robot，2003，25(6)：560-567，573.

[2]张雪华，刘华平，孙富春，等. 采用 Kinect 的移动机器人目标跟踪[J]. 智能系统学报，2014，9(1)：34-39.
ZHANG Xuehua，LIU Huaping，SUN Fuchun，et al. Target tracking of mobile robot using Kinect[J]. CAAI transactions on intelligent systems，2014，9(1)：34-39.

[3]ANDO N，SUEHIRO T，KITAGAKI K，et al. RT-middleware：distributed component middleware for RT (robot technology)[C]//Proceedings of 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems，2005. (IROS 2005). Edmonton，Alta.，Canada：IEEE，2005：3933-3938.

[4]ANDO N，SUEHIRO T，KOTOKU T. A software platform for component based RT-system development：OpenRTM-aist[M]//CARPIN S，NODA I，PAGELLO E，et al. Simulation，Modeling，and Programming for Autonomous Robots. Berlin Heidelberg：Springer，2008：87-98.

[5]庄严，王伟，王珂，等. 移动机器人基于激光测距和单目视觉的室内同时定位和地图构建[J]. 自动化学报，2005，31(6)：925-933.
ZHUANG Yan，WANG Wei，WANG Ke，et al. Mobile robot indoor simultaneous localization and mapping using laser range finder and monocular vision[J]. Acta automatica sinica，2005，31(6)：925-933.

[6]GERKEY B P. AMCL[EB/OL]. http://www.ros.org/wiki/amcl，2011.

[7]刘洞波，刘国荣，胡慧，等. 基于激光测距的温室移动机器人全局定位方法[J]. 农业机械学报，2010，41(5)：158-163.
LIU Dongbo，LIU Guorong，HU Hui，et al. Method of mobile robot global localization based on laser range finder in greenhouse[J]. Transactions of the Chinese society for agricultural machinery，2010，41(5)：158-163.

[8]李振伟，陈翀，赵有. 基于 OpenCV 的运动目标跟踪及其实现[J]. 现代电子技术，2008，31(20)：128-130，138.
LI Zhenwei，CHEN Chong，ZHAO You. Moving object tracking method and implement based on OpenCV[J]. Modern electronics technique，2008，31(20)：128-130，138.

[9]张汝波，顾国昌，刘照德，等. 强化学习理论、算法及应用[J]. 控制理论与应用，2000，17(5)：637-642.
ZHANG Rubo，GU Guochang，LIU Zhaode，et al. Reinforcement learning theory，algorithms and its application[J]. Control Theory & Applications，2000，17(5)：637-642.

[10]黄炳强，曹广益，王占全. 强化学习原理、算法及应用[J]. 河北工业大学学报，2007，35(6)：34-38.
HUANG Bingqiang，CAO Guangyi，WANG Zhanquan. Reinforcement learning theory，algorithms and application[J]. Journal of Hebei University of Technology，2007，35(6)：34-38.

[11]SUTTON R S，BARTO A G. Reinforcement learning：an Introduction[M]. Cambridge，Mass.：MIT Press，1998：114-116.

[12]WANG Wenshan，CAO Qixin，ZHU Xiaoxiao，et al. An automatic switching approach of robotic components for improving robot localization reliability in complicated environment[J]. Industrial robot：an international journal，2014，41(2)：135-144.


**作者简介：**

梁爽，女，1993 年生，硕士研究生，主要研究方向为移动机器人路径规划及模块化机器人技术。

曹其新，男，1960 年生，教授，博士生导师，主要研究方向为智能机器人与模块化系统、机器视觉与模式识别、移动机器人、泛在机器人技术。被 EI、SCI 检索论文 90 余篇，获得发明和实用新型专利 50 余项。

王雯珊，女，1986 年生，博士研究生，主要研究方向为泛在机器人、任务规划。