DOI：10.13245/j.hust.181212

基于强化学习的机器人路径规划算法

张福海 李 宁 袁儒鹏 付宜利

(哈尔滨工业大学机器人技术与系统国家重点实验室，黑龙江 哈尔滨 150001)

摘要 提出了一种基于强化学习的机器人路径规划算法，该算法将激光雷达所获取的移动机器人周围障碍物信息与目标点所在方位信息离散成有限个状态，进而合理地设计环境模型与状态空间数目；设计了一种连续的报酬函数，使得机器人采取的每一个动作都能获得相应的报酬，提高了算法训练效率．最后在Gazebo中建立仿真环境，对该智能体进行学习训练，训练结果验证了算法的有效性；同时在实际机器人上进行导航实验，实验结果表明该算法在实际环境中也能够完成导航任务．

关键词 移动机器人；强化学习；路径规划；连续报酬函数；导航实验

中图分类号 TP242 **文献标志码** A **文章编号** 1671-4512(2018)12-0065-06

Robot path planning algorithm based on reinforcement learning

Zhang Fuhai Li Ning Yuan Rupeng Fu Yili

(The State Key Laboratory of Robotics and System，Harbin Institute of Technology，Harbin 150001，China)

Abstract A path planning algorithm for mobile robot was studied based on reinforcement learning．The algorithm discretized the obstacle information around the mobile robot acquired by the LIDAR(laser intensity direction and ranging)and the position information of the target point into finite state，and then rationally designed the number of the environmental model and state spaces．In addition，a continuous reward function was studied，which made each action taken by the robot get corresponding reward and improved the efficiency of algorithm training．Finally，a simulation environment was established in Gazebo to learn and train the agent．The training results verify the effectiveness of the algorithm．Simultaneously，a navigation experiment was conducted on an actual robot．The results show that the algorithm can also complete the navigation task in the actual environment．

Key words mobile robot；reinforcement learning；path planning；continuous reward function；navigation experiment

在机器人自主导航中，路径规划是一个非常重要的部分．机器人路径规划问题可以描述为在机器人自身位姿已知的情况下，根据一个或者多个优化目标，例如工作代价最小、轨迹长度最短、运动时间最少等，在机器人工作环境中寻找到一条从当前点到指定目标点的最优路径[1]．机器人路径规划算法可以根据对工作环境的理解程度不同而分为全局路径规划和局部路径规划：**a.** 全局路径规划，当全局静态环境地图已知时，在静态环境条件下按照特定的算法搜寻一条无碰撞路径；**b.** 局部路径规划，主要考虑机器人在动态环境中的避障，机器人只了解环境的部分信息或者对环境信息完全不了解，则根据传感器获取的信息不断地更新环境信息．局部路径规划算法考虑了机器人本身运动参数、路径的方向与障碍物等信息，所以近年来局部路径规划问题逐渐成为研究的重点．

局部路径规划常见的算法有人工势场法[2-3]、向量直方图法[4]、遗传算法[5-6]、模糊逻辑法[7]、强化学习法[8-9]等．强化学习算法是一种完全不需要环境与机器人自身的先验知识的学习方法，机器人边感

收稿日期 2018-05-09．

作者简介 张福海(1977-)，男，讲师，E-mail：zfhhit@163.com．

基金项目 黑龙江省自然科学基金资助项目(LC2017022)．

知当前环境的状态边行动，根据状态和行动，环境状态迁移到新的状态，相应地，新的状态的“奖惩”报酬信息返还给机器人，机器人根据报酬信息决定下一个行动．文献[10]将环境模型定义为三类，即机器人周围目标点所在象限、机器人周围最近障碍物所在象限及机器人与障碍物连线和机器人与目标点连线的夹角大小，并按照机器人与障碍物目标点的距离分为成功、失败、安全与不安全状态，由状态的转移来定义报酬函数．该方法缺点是只考虑了最近障碍物信息，简单场景算法学习能力较强，复杂场景适应能力较弱．文献[11]引入神经网络拟合Q_table，提高了算法的收敛性，但其对周围障碍物的分布情况没有明确划分，容易陷入局部最小状态．文献[12]提出了基于强化学习与神经网络相结合的移动机器人避障算法，状态为机器人前方传感器获取的障碍物信息，将状态与动作输入网络，输出为其对应的Q值，选取Q值最大的动作执行，算法能够较好地完成避障任务，但因为缺乏目标点的信息，所以无法完整地进行路径规划．

本文提出一种基于强化学习的移动机器人路径规划算法，利用离散化的激光雷达信息，合理地设计环境模型与状态空间数目；研究连续的报酬函数，加快算法的收敛速度；最后在Gazebo环境中进行仿真学习训练，同时在实际机器人上进行导航实验，验证了算法的有效性.

1 局部路径规划方法设计

1.1 强化学习算法基本原理

Q_learning(Q学习)是典型的model-free(无模型)强化学习算法，该算法特点是每个状态下所采取动作的Q值可以通过Q_table表格的形式存储，方便获取每个状态与动作下的Q值．它是迭代更新式的学习方法，通过价值函数Q逼近目标函数Q^*，而不需要环境的任何先验知识的一种model-free学习算法．Q_learning的更新迭代形式为

$$Q(s,\ a)=Q(s,\ a)+\alpha(r(s,\ a)+\gamma\max_{a'\in A}Q(s',\ a')-Q(s,\ a)),\quad(1)$$

式中：r为状态s下采取行动a时的报酬；s'为下一个状态；α为学习率，代表学习新知识的程度，取0～1之间；γ为折扣率，代表考虑未来报酬的程度，取0～1之间．

1.2 动作状态空间设计

机器人的环境模型既要充分考虑周围障碍物信息，又要充分考虑目标点所在位置才能避免碰撞．由于连续性的高维状态空间会使得强化学习算法难以收敛，即状态与动作数量十分巨大，因此须要将机器人环境状态空间离散化．环境模型由机器人、目标点及障碍物组成，状态定义为

$$s=(R_{\rm g},\ R_{\rm o1},\ R_{\rm o2},\ R_{\rm o3},\ R_{\rm o4}),\quad(2)$$

式中：$R_{\rm g}$描述机器人要到达的目标点方位情况；$R_{\rm o1}$～$R_{\rm o4}$用来描述机器人前方障碍物的分布情况．

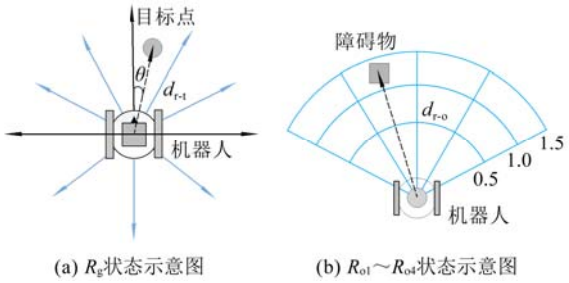
由于差动机器人仅能向机器人航向角所在方向运动，导航目标点与机器人的连线与机器人前进方向的夹角θ对于机器人路径规划很重要，可以控制机器人不断调整方向朝向目标点移动，因此将θ按照一定角度离散成七个状态(见图1)，即机器人周围目标点所在位置情况的状态$R_{\rm g}$定义如下：

$$R_{\rm g}=\begin{cases}1 & (-25°<\theta\leqslant 25°);\\ 2 & (25°<\theta\leqslant 80°);\\ 3 & (80°<\theta\leqslant 130°);\\ 4 & (130°<\theta\leqslant 180°);\\ 5 & (-180°<\theta\leqslant -130°);\\ 6 & (-130°<\theta\leqslant -80°);\\ 7 & (-80°<\theta\leqslant -25°).\end{cases}\quad(3)$$

将机器人前方120°范围内障碍物的分布按照角度与距离离散成四个表征障碍物分布的状态，机器人方向角所指方向为0°方向，以0.5 m分辨率离散连续空间为以下四个状态，当机器人与障碍物连线夹角在[20°，60°]时为状态$R_{\rm o1}$，其定义如下

$$R_{\rm o1}=\begin{cases}1 & (0.0<d_{\rm r\_o}\leqslant 0.5);\\ 2 & (0.5<d_{\rm r\_o}\leqslant 1.0);\\ 3 & (1.0<d_{\rm r\_o}\leqslant 1.5);\\ 4 & (1.5<d_{\rm r\_o}),\end{cases}\quad(4)$$

式中$d_{\rm r_o}$为机器人与该角度区域内最近障碍物的距离．当机器人与障碍物连线夹角在[-20°，0°)，[0°，20°)和[−60°，−20°)时的状态$R_{\rm o2}$，$R_{\rm o3}$和$R_{\rm o4}$的定义与$R_{\rm o1}$类似．输入状态共有7×4×4×4×4=1 792个．



(a) $R_{\rm g}$状态示意图 (b) $R_{\rm o1}$～$R_{\rm o4}$状态示意图

图1 环境模型状态示意图

1.3 动作状态空间设计

在矢量空间中的速度采样点主要分为前进、左

转与右转三个动作集合，这样设计的动作相比较于其他固定移动速度、只输出转向角的算法更符合机器人运动学特性，使机器人运动更加灵活，主要包括三部分：前进的速度(线速度 v，角速度 ω)为(0.3，0.0)，左转的速度为(0.1，-0.6)，右转的速度为(0.1，0.6).

实际选取动作过程中同样考虑机器人运动特性、电机负荷能力等因素，设置动态窗口，在窗口中选取动作，保证输出动作的可执行性、运动的连续性等. 根据电机的负荷能力，机器人存在最大加减速度的限制，因此在控制周期内机器人实际能达到的速度即动态窗口，动态窗口 V_{d} 定义如下

$$V_{\mathrm{d}}=\{(V,\ \omega)\,|\,V\in[v_{\mathrm{c}}-\dot{v}_{\mathrm{a}}\Delta t,\ v_{\mathrm{c}}+\dot{v}_{\mathrm{a}}\Delta t]\bigcap \omega\in[\omega_{\mathrm{c}}-\dot{\omega}_{\mathrm{a}}\Delta t,\ \omega_{\mathrm{c}}+\dot{\omega}_{\mathrm{a}}\Delta t]\};\tag{5}$$

式中：V为机器人一个采样周期内线速度范围；ω 为机器人一个采样周期内角速度范围；v_{c}为机器人当前的线速度；ω_{c} 为机器人当前的角速度；$\dot{v}_{\mathrm{a}}$ 为机器人最大线加速度；$\dot{\omega}_{\mathrm{a}}$ 为机器人最大角加速度；Δt 为机器人采样周期. 考虑动态窗口约束的机器人运动，可根据机器人当前速度值，与电机所能承受的最大加减速度限制，在机器人所能承受的范围内，选择可行的速度.

1.4　报酬函数设计

报酬函数是机器人在其状态下采取行动的估计值，表明在特定状态下采取动作的好坏程度，通常须要手动设计报酬函数. 若报酬函数是连续的，即在训练过程中报酬值时刻存在，则算法可以有效利用这些信息进行不断学习. 所以本研究设计了一种连续的报酬函数，即

$$r(s_t,\ a_t)=\begin{cases} r_{\mathrm{rea}} & (d_{\mathrm{r-t}}(t)\leqslant d_{\mathrm{win}});\\ r_{\mathrm{col}} & (d_{\mathrm{r-o}}(t)\leqslant d_{\mathrm{col}});\\ c_t(d_{\mathrm{r-t}}(t-1)-d_{\mathrm{r-t}}(t))+\eta_{\mathrm{o}}\dfrac{d_{\mathrm{r-o}}(t)}{d_{\mathrm{r-o}}(t-1)}, \end{cases}\tag{6}$$

式中：r_{rea} 为机器人到达目标点获得的报酬值；r_{col} 为机器人发生碰撞报酬值；η_o 和 c_t 为系数；$d_{\mathrm{r-t}}(t-1)$ 为 t-1 时刻机器人与目标点间的距离；$d_{\mathrm{r-t}}(t)$ 为 t 时刻机器人与目标点间的距离；$d_{\mathrm{r-o}}(t)$ 为 t 时刻机器人与最近障碍物间的距离；d_{win} 为机器人到达目标点的阈值；d_{col} 为机器人与障碍物发生碰撞的阈值.

若机器人距离目标点小于一定阈值 d_{win}，则给予一个正的报酬 r_{rea}；若机器人距离障碍物小于一定的阈值 d_{col}，则给予一个负的报酬 r_{col}，否则报酬值为机器人采取上一个动作所造成环境的改变，与目标点间距离的变化值；若该动作减小了到达目标点的距离，则获得一个正的报酬，否则获得一个负的报酬. 该设计是为了使机器人不断向目标点移动，使机器人每采取一个动作都能及时获得反馈，保证报酬函数的连续性，加快算法的收敛速度.

2　实验

2.1　实验平台

本实验室基于机器人操作系统(ROS)自主研发的一款移动机器人系统，其采用底层硬件与顶层软件相结合的方式来实现机器人自主导航.

2.2　仿真实验

Gazebo 仿真场景为 10 m×7 m 的仿真实验场景，机器人的起点位置为(0 m，0 m)，目标点位置为(4.5 m，4.5 m)，场景的周围由墙组成，内部随机设置一些障碍物. 机器人按照本文所提出的强化学习算法，在该场景内不断地学习探索，一旦机器人与障碍物发生碰撞或者到达目标点，则整个场景复位. 仿真实验参数设置如下：学习率 α=0.2，r_{rea}=1，c_t=−1，d_{win}=0.25，折扣率 γ=0.9，r_{col}=−1，d_{col}=0.05，总训练次数为 5 000 次.

仿真实验动作的选取策略为 ε - greedy 策略，共训练 5 000 次，ε 初始设置为 1，前期每 100 次 ε 减小 0.1；在 ε 降低到 0.1 之后，每 100 次 ε 减小 0.01 一直降低到 0.05 保持不变，即保持 5%概率的动作选择随机性，增加探索环境的可能性. 图 2 为强化学习算法成功率(ξ)曲线图(n 为训练次数)，经过约 30 h 的训练之后，机器人无碰撞地从起点出发移动到指定目标点，算法成功率收敛到 95%. 在训练初期，机器人没有充分地学习，随机选择动作概率 ε 较大，机器人主要在探索环境，经常发生碰撞，到达目标点次数较少，而随着训练次数的上升，随机选择动作概率 ε 逐渐减小，经过充分“试错”学习之后，机器人由探索环境状态逐渐转为利用知识状态，算

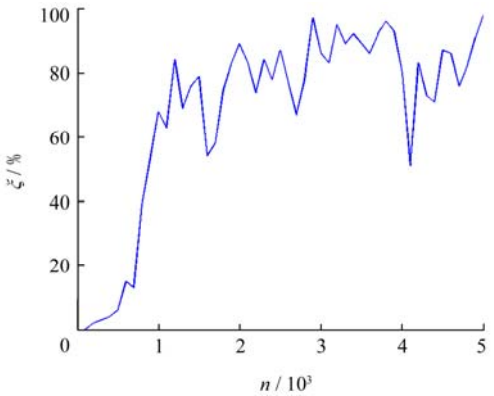


图 2　成功率曲线

法进入快速收敛阶段，成功率迅速升高.

图 3 为训练过程中路径图，(a)～(d)分别为训练第 423 次、第 1 566 次、第 3 532 次与第 4 879 次的结果. 由于处于训练前期，随机概率较大，因此图 3(a)算法没有收敛，机器人碰撞障碍物. 其他三种情况经过训练都可以避开障碍物到达终点.

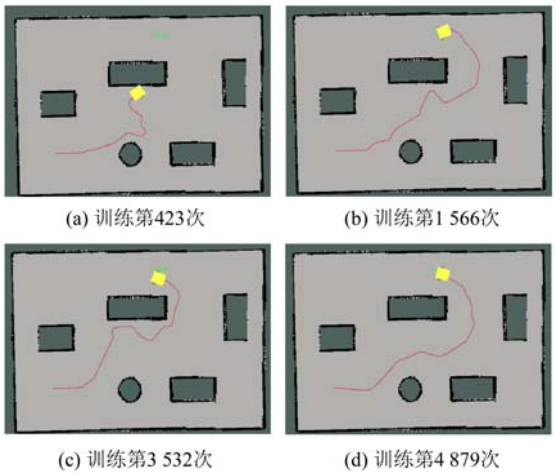


图 3　强化学习算法训练过程中的路径图

图 4 为上述训练过程中机器人每步所获得的 Q 值，图中(a)～(d)分别与图 3(a)～(d)对应，机器人每次选取当前状态下 Q 值最大的动作执行，(a)处于训练前期所有 Q 值都较低，并且机器人发生碰撞，特别是后面训练状态下 Q 值都非常低，表明机器人处于危险状态，选择该动作所获得的回报较低. (b)、(c)与(d)机器人均顺利避过障碍物成功到达终点，尤其训练后期机器人接近终点时，这些状态未来的报酬期望值较高，选择这些动作可以获得较高的报酬值，所以 Q 值较高，符合强化学习方法所能产生的结果.

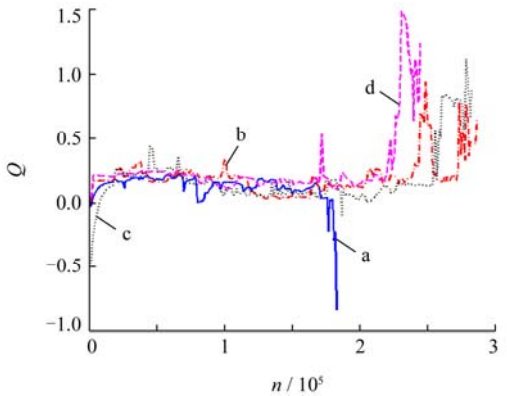


图 4　强化学习算法 Q 值

实验(a)～(d)四次仿真训练机器人运动路径长度分别为 5.79，10.68，10.56，9.93 m，可以看出除了实验(a)失败以外，其他几次实验随着次数的增加，机器人路径长度逐渐减小，表明随着训练的深入，算法能够选择更优的轨迹.

图 5 为训练过程中每次选择动作的平均 Q 值，由图 5 可以看出：训练前期，因为随机概率较大，机器人获得的知识与经验较少，机器人到达目标点次数较少，所以平均 Q 值较低；在训练达到 1 000 次之后，机器人经过学习到达目标点的次数逐渐增加，获得正报酬的次数越来越多，所以平均 Q 值呈现逐渐上升的趋势，即机器人选择动作获得的平均累计报酬值越来越高；之后随着训练次数增加，到达目标点的次数越来越多，获得报酬也越来越多，Q 值逐渐增加，最终算法逐渐收敛.

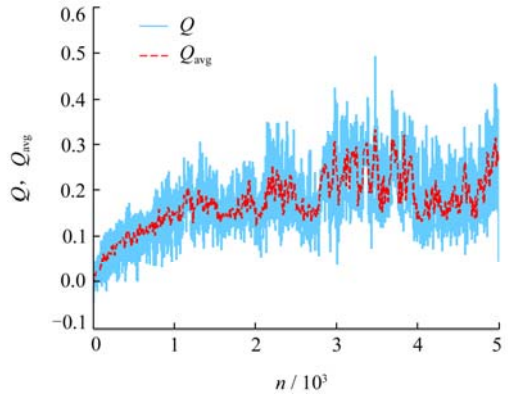


图 5　训练过程中的平均 Q 值

图 6 为训练过程中累计报酬值(r 和 r_{avg} 分别为报酬值和平均报酬值)，机器人到达终点可以获得最高报酬值 1，发生碰撞获得最低报酬值−1，机器人朝向目标点移动可以获得相应的正报酬. 由图 6 可以看出：训练前期因为随机概率较大，机器人处于探索环境阶段，经常发生碰撞，所以每次实验所获得的报酬值较低；随着训练的深入，随机概率减低，机器人进入利用知识阶段，到达目标点次数增加，每次训练获得的报酬值不断增加，即机器人能够较好地利用学到的知识避开障碍物，到达终点.

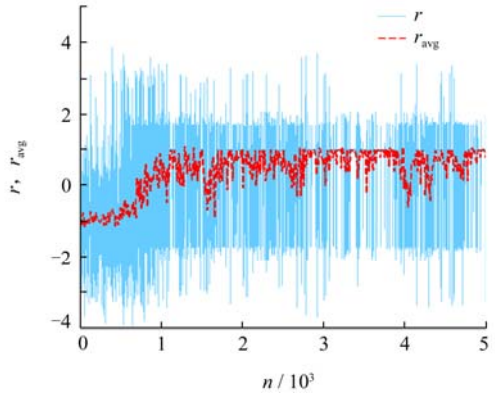


图 6　训练过程中报酬值

2.3　实物实验

实物实验场景仿照仿真场景，首先用 SLAM(即时定位与地图构建)算法对实验场景建立占据栅格地图(5.4 m×8.0 m)，然后设置机器人起点与终点位

置，最后运用本文所提出的算法进行实验.

图 7 为实验过程的图片，其右下角为实验过程中 Rviz 与强化学习程序的输出终端图. Rviz 是 ROS 官方的一款 3D 图形化工具，可以方便地对 ROS 的程序进行图形化操作. 强化学习程序终端输出的内容包括当前状态、执行动作、获得的报酬值等信息.

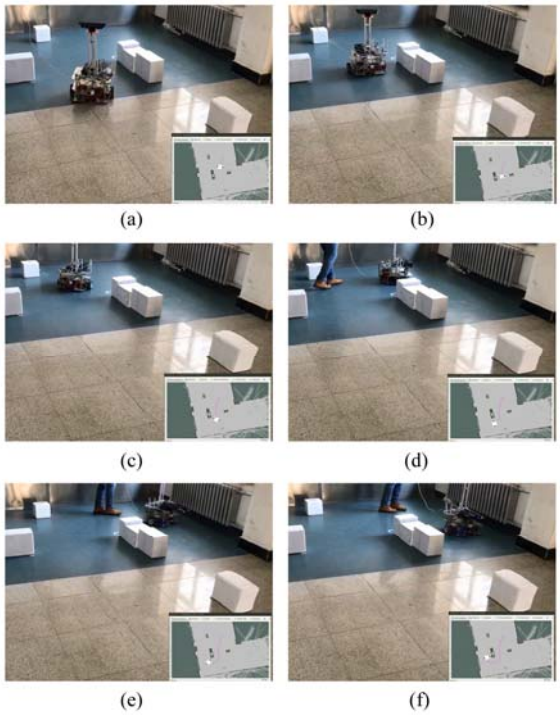


图 7　机器人实验过程

图 8 所示为实物实验过程中获取的机器人的 Q 值曲线，图 8 中曲线表明：机器人距离障碍物较近，处于危险状态时其 Q 值较低；当机器人距离障碍物较远、处于安全状态时其 Q 值较高，并且当接近终点附近时 Q 值最高.

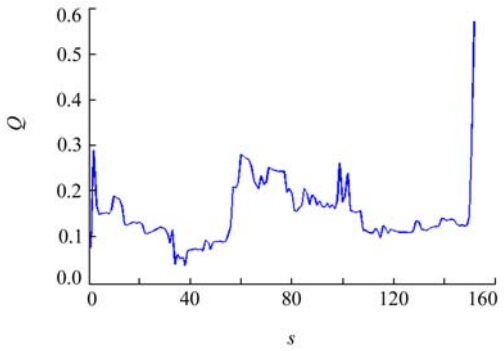


图 8　实验过程中机器人 Q 值

为验证算法的有效性和可靠性，进行了三次机器人自主导航实验，记录得到机器人运行路径与运动时间如表 1 所示. 由表 1 可以看出：机器人三次导航均能成功到达目标点，并且路径长度与运动时间相差不多，表明算法具有可实现性和较好的稳定性，即机器人可以通过在仿真环境中运用强化学习算法进行训练学习，在实际环境中同样可以利用学到的知识与经验完成路径规划任务.

表 1　实验运动路径长度与运动时间

实验序号	路径长度/m	运动时间/s
1	4.721	30.26
2	4.956	31.44
3	4.689	29.76

3　结语

本研究提出了一种基于强化学习的机器人局部路径规划算法，分析了强化学习算法应用于移动机器人路径规划上的优势，将激光雷达所获取的障碍物信息离散化为有限数量的区域，并合理地设计了环境模型与状态空间数目，既充分考虑了机器人周围障碍物的分布，又合理控制状态数量；设计了连续的报酬函数，使机器人每执行一个动作都能获得及时的反馈，加快了算法的收敛速度. 最后在 Gazebo 环境中建立算法仿真环境，对移动机器人进行仿真学习训练，仿真结果验证了算法的有效性. 同时用实际机器人也进行了导航实验，验证了算法在实际环境中也能够完成导航任务.

参　考　文　献

[1] Wu H，Tian G H，Li Y，et al. Spatial semantic hybrid map building and application of mobile service robot[J]. Robotics and Autonomous Systems，2014，62(6)：923-941.

[2] Kovács B，Szayer G，Tajti F. A novel potential field method for path planning of mobile robots by adapting animal motion attributes[J]. Robotics and Autonomous Systems，2016，82：24-34.

[3] Zhang Q，Yue S G，Yin Q J，et al. Dynamic obstacle-avoiding path planning for robots based on modified potential field method[C]// Proc of International Conference on Intelligent Computing Theories and Technology. Berlin：Springer-Verlag，2013：332-342.

[4] Babinec A，Dekan M，Duchoň F，et al. Modifications of VFH navigation methods for mobile robots[J]. Procedia Engineering，2012，48(1)：10-14.

[5] Tuncer A，Yildirim M. Dynamic path planning of mobile robots with improved genetic algorithm[J]. Computers and Electrical Engineering，2012，38(6)：1564-1572.

[6] Karami A H，Hasanzadeh M. An adaptive genetic algorithm for robot motion planning in 2D complex

environments[J]. Computers and Electrical Engineering，2015，43：317-329.

[7] Pandey A，Sonkar R K，Pandey K K，et al. Path planning navigation of mobile robot with obstacles avoidance using fuzzy logic controller[C]// Proc of International Conference on Intelligent Systems and Control. New York：IEEE，2014：36-41.

[8] Su M C，Huang D Y，Chow C H，et al. A reinforcement-learning approach to robot navigation[C]// Proc of International Conference on Networking，Sensing and Control. Taipei：IEEE，2004：665-669.

[9] Chen Y F，Liu M，Everett M，et al. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning[C]// Proc of International Conference on Robotics and Automation. Singgapore：IEEE，2017：285-292.

[10] Jaradat M A K，Al-Rousan M，Quadan L. Reinforcement based mobile robot navigation in dynamic environment [J]. Robotics and Computer-Integrated Manufacturing，2011，27(1)：135-149.

[11] Duguleana M，Mogan G. Neural networks based reinforcement learning for mobile robots obstacle avoidance [J]. Expert Systems with Applications，2016，62：104-115.

[12] Huang B Q，Cao G Y，Guo M. Reinforcement learning neural network to the problem of autonomous mobile robot obstacle avoidance[C]// Proc of International Conference on Machine Learning and Cybernetics. New York：IEEE，2005：85-89.