

# 基于强化学习的机器人自适应路径规划方法

北京理工大学机器人研究中心 童亮 龚建伟 熊光明 陆际联 高峻尧等

Robotics Research Center, Beijing Institute of Technology.

L. Tong, J.W. Gong, G.M. Xiong, J.L. Lu

转载此文请署名作者 并标明来自龚建伟技术主页 [www.gjwtech.com](http://www.gjwtech.com)

此文工作已在学术期刊上正式发表

基于强化学习的机器人自适应路径规划方法.....	1
1 路径规划问题.....	1
2 常用机器人路径规划方法.....	2
2.1 传统路径规划方法.....	2
2.2 智能路径规划方法.....	5
3 动态环境的机器人自适应路径规划.....	7
3.1 改进 CMAC 网络的 Q-Learning 系统.....	8
3.2 路径规划控制总体模型设计.....	9
3.2 学习系统的环境感知及行为模型.....	10
3.3 强化函数的设定.....	11
4 系统学习结果及讨论.....	12
5 小结.....	14

## 1 路径规划问题

使机器人具有智能,在无人干预的情况下,自主完成任务已经成为机器人技术发展和在各领域广泛应用的迫切需求。要使机器人具有一定的智能,传统的实现方法是将人的知识和经验直接移植到机器人上,一般靠人事先编程来建立知识库和推理机制。由于真实环境多是动态的、不确定的和复杂的,一方面人不可能

预见到全部情况,另一方面对于复杂的环境和任务,手工编码也是一个非常繁重的工作,有时甚至是不可能实现的。学习是人类获取知识的主要形式,也是人类具有智能、提高智能水平的基本途径。因此人们希望机器人具有从环境中学习的能力,即自动获取知识、积累经验、不断更新和扩展知识的能力。近几年,使机器人具有自学习、自适应能力成为一个研究的热点,强化学习由于其比较符合人类和动物的学习过程、与Brooks提出的行为主义思想一致、可以不需要环境模型实现无导师的在线学习等特点,已经成为最受关注的一种方法。

机器人路径规划是自动机器人控制中一个基本的计算问题。目前已经提出多种路径规划算法并且仍有许多研究者热衷于这方面的研究工作。虽然这些算法的有关性质,如正确性、完备性、结果的最优性以及它们的时间和空间的复杂性已经进行了充分的研究,但如何选择一种适应于非特定环境及参数最优化方法却没有引起人们的足够重视。在我们的观点中,路径规划被认为是一种控制问题。事实上,路径规划和控制有许多共同点:它们都有起点和目标点、它们都依赖于特定的误差标准和机器的动力学性质、它们都必须在多维状态空间中生成一条到达目标点的路径等。

## 2 常用机器人路径规划方法

路径规划技术是机器人研究领域中的一个重要分支。所谓机器人的最优路径规划问题,就是依据某个或某些优化准则(如工作代价最小、行走路线最短、行走时间最短等),在其工作空间中找到一条从起始状态到目标状态的能避开障碍物的最优路径。根据控制方法的不同,机器人路径规划方法大致可以分为两类:传统方法和智能方法。

### 2.1 传统路径规划方法

传统的路径规划方法是根据优化的目标分为四个类型:一种是根据目标点的吸引力,一种是路径最短,一种是考虑通过环形障碍的步骤,另一种是路径中的空间最大化。具此分为以下几种类型:

(1) 人工势场法<sup>[106]</sup>是路径规划中应用比较普遍的一种方法。其基本原理就是在机器人所处离散环境中的每一点 $P$ 赋一个势场值 $v(p)$ , $v(p)$ 的值是目标点的引力和障碍物的斥力的叠加。

$$v(p) = b_{target} \cdot \delta_{target} + b_o \sum_i \frac{1}{\delta_{o_i}}$$

$b_{target}$  和  $b_o$  是距离影响因子,  $\delta_{target}$  (与目标的距离) 和  $\delta_{o_i}$  (与障碍物的距离)。因此机器人的路径规划就是从起始点沿着势场最快下降的方向达到目标点。这种基本算法的缺点是容易产生局部最小从而产生局部最优问题。通过对这种基本算法的扩展, 这一问题已经得到很好的解决。这一算法产生的路径是以牺牲安全、与最近障碍物的距离和路径的最优长度为代价的。

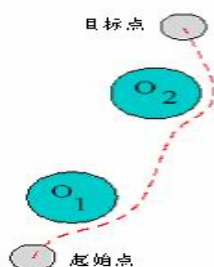


图 1 人工势场法机器人路径规划

(2) 反向梯度法是另一类型的路径规划算法, 其目的是求取最短路径<sup>[107]</sup><sup>[108]</sup>。这种算法将环境划分为一定数量的小格子并且计算出每个小格子到达目标点的最短路径距离。目标点被标注一个为 0 的值, 所有其它的小格子都被初始化一个很大的值, 算法从目标点开始并且遍历与其邻接的每一个状态, 不断重复。一个邻接状态  $p_i$  的状态  $p_{i+1}$  如果处在障碍物中, 其值  $v(p_{i+1})$  被设置为无穷大, 否则的话,  $v(p_{i+1}) = \min(v(p_i) + 1, v(p_{i+1}))$ 。这类算法的缺点在于它没有考虑路径的曲率的概念, 所以可能导致机器人需要对它的方向和速度进行频繁的调整, 而且这类算法的基本实现方法中比较偏爱靠近障碍物的路径。对于这些缺点, 可以通过增加人工障碍物或者可以将机器人的导航看作一个 Markov 决策过程<sup>[109]</sup>来解决, 在这个过程中, 机器人的行为结果是非确定性的。而后者可能会增加计算机的计算开销。同时, 这类路径规划方法如果应用于大于三维的状态空间时会产生严重的计算问题。

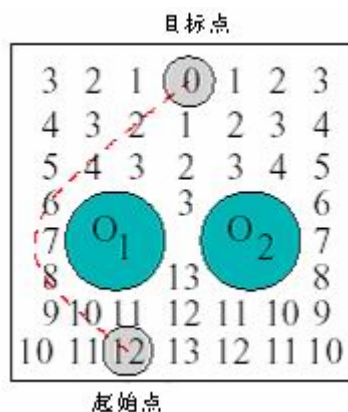


图 2 反向梯度法机器人路径规划

(3) 环绕障碍法是获得环绕障碍物的一个航行动作序列，这一序列与通向目标点的直线相交。A\*算法<sup>[110]</sup>可以利用给定的环绕障碍物的点计算出最短路径。这一类算法包括Viapoint方法<sup>[111]</sup>和弹性波段算法（动态障碍）<sup>[112]</sup>。这类算法在起始点或目标点特别靠近障碍物时容易产生问题。

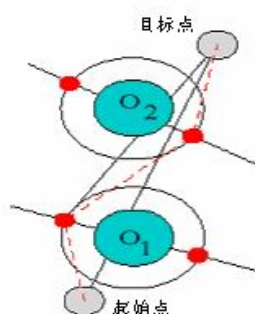


图 3 环绕障碍法机器人路径规划

(4) 自由空间算法这类路径规划算法的宗旨是求得距离障碍物最远的路径。其中一个最著名的算法就是Voronoi路径规划算法，这一算法引导机器人朝着与最近障碍物相等距离的路径行驶。另一种该类型的算法是在动态障碍物环境中最大化安全空间方法。Buck<sup>[113]</sup>在这一算法中通过计算得到与至少两个障碍物具有相同距离的评价点作为通向目标点的中间点。由于与障碍物的平均距离很大，所以这种算法得到的路径比通过其它方法得到的路径具有更小的曲率，但得到的路径会比其它方法得到的路径要长。

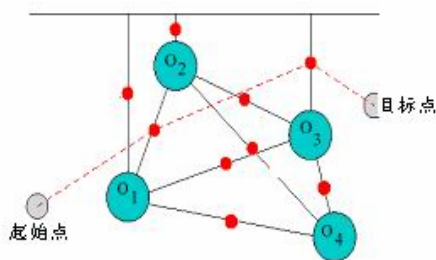


图4 最大空间法机器人路径规划

大部分机器人路径规划中的全局规划都是基于上述几种方法进行的,但是以上这些传统方法在路径搜索效率及路径优化方面尚有待于进一步改善。现在通常使用的搜索技术包括:梯度法、A3 等图搜索方法、枚举法、随机搜索法等,梯度法易陷入局部最小点,图搜索方法、枚举法不能用于高维的优化问题,而随机搜索法则计算效率太低。

## 2.2 智能路径规划方法

近年来,随着遗传算法等智能方法的广泛应用,机器人路径规划方法也有了长足的进展,许多研究者把目光放在了基于智能方法的路径规划研究上。其中,应用较多的算法主要有模糊方法、神经网络和遗传算法等。

### (1) 模糊逻辑的机器人路径规划

模糊控制算法模拟驾驶员的驾驶思想,将模糊控制本身所具有的鲁棒性与基于生理学上的“感知、动作”行为结合起来,适用于时变未知环境下的路径规划,实时性较好。模糊逻辑方法是在线规划中通常采用的一种规划方法,包括建模和局部规划。Hartmut Surmann 等<sup>[114]</sup>提出一种未知环境下的高级机器人模糊导航方法,由 8 个不同的超声传感器来提供环境信息,然后利用基于模糊控制的导航器来计算这些信息,规划机器人路径。

该方法在环境未知或发生变化的情况下,能够快速而准确地规划机器人路径,对于要求有较少路径规划时间的机器人是一种很好导航方法。但是,其缺点是当障碍物数目增加时,该方法的计算量会很大,影响规划结果。

### (2) 基于神经网络方法的机器人路径规划<sup>[115]</sup>

人工神经网络是一种仿效生物神经系统的信息处理方法。神经网络的优点主要体现在它可以处理难以用模型或规则描述的过程和系统;对非线性系统具有统一的描述;具有较强的信息融合能力和系统容错能力。基于神经网络的多传感器

信息融合正是利用了神经网络这些特性,将传感器的数据信息作为神经网络的输入进行处理,可以获得移动机器人对障碍物影像的比较精确的估计。神经网络系统受学习样本的影响很大,选择代表性强的样本集是十分困难的,而让样本集覆盖整个样本空间是不现实的,因而样本的选择与设计是一大难题。

### (3) 基于遗传算法的机器人路径规划

J.Holland<sup>[116]</sup>在 20 世纪 60 年代初提出了遗传算法,以自然遗传机制和自然选择等生物进化理论为基础,构造了随机化搜索算法。它利用选择、交叉和变异来培养控制机构的计算程序,在某种程度上对生物进化过程做数学方式的模拟。它不要求适应度函数是可导或连续的,而只要求适应度函数为正。同时作为并行算法,它的隐并行性适用于全局搜索,多数优化算法都是单点搜索算法,很容易陷入局部最优,而遗传算法却是一种多点搜索算法,因而更有可能搜索到全局最优解。由于遗传算法的整体搜索策略和优化计算不依赖于梯度信息,所以解决了一些其他优化算法无法解决的问题,但遗传算法运算速度不快,进化众多的规划要占据较大的存储空间和运算时间。但由于常规遗传算法本身所存在的一些缺陷(如解的早熟现象、局部寻优能力差等),保证不了对路径规划的计算效率和可靠性的要求。为提高路径规划问题的求解质量和求解效率,又有研究者提出了在利用遗传算法进行路径规划的基础上,引入模拟退火算法,抑制了遗传算法的早熟现象,克服了其局部寻优能力较差的缺点,形成一种遗传模拟退火算法来解决机器人路径规划问题。

遗传算法用于复杂环境下路径规划存在以下缺陷:(一)路径个体编码设计若不合理,会导致进化缓慢、进化过程中产生非法个体;(二)若遗传算子选择不合理,进化效果不明显;(三)若规划过程中没有利用背景知识,进化效率不高。

### (4) 基于混合方法的机器人路径规划方法

L. H. Tsoukalas<sup>[117]</sup>等提出一种用于半自主移动机器人路径规划的模糊神经网络方法。所谓半自主移动机器人就是具有在人类示教基础上增加了学习功能器件的机器人。这种方法采用模糊描述来完成机器人行为编码,同时重复使用神经网络自适应技术。由机器人的传感器提供局部的环境输入,由内部模糊神经网络进行环境预测,进而可以在未知环境下规划机器人路径。此外,也有人提出基于模糊神经网络和遗传算法的机器人自适应控制方法。将规划过程分为离线学习和在线学习两部分。

另外,根据机器人对环境信息掌握的程度、障碍物的不同,移动机器人的路

径规划又可分为以下几类：

- (1) 已知环境下静态环境路径规划；
- (2) 未知环境下静态环境路径规划；
- (3) 已知环境下动态环境路径规划；
- (4) 未知环境下动态环境路径规划。

也可根据对环境信息掌握的程度不同将移动机器人路径规划分为两种类型，一个是基于环境先验完全信息的全局路径规划，另一个是基于传感器信息的局部路径规划，后者环境是未知或部分未知的，即障碍物的尺寸、形状和位置等信息必须通过传感器获取。全局路径规划是指根据先验环境模型找出从起始点到目标点的符合一定性能的可行或最优路径，它涉及的基本问题是世界模型的表达和搜寻策略。

对于前面提到的路径规划方法，其应用大部分集中在对已知静态环境下的机器人进行路径规划，对于环境为动态，特别是动态的未知环境的研究还不深入，本章研究的目的是根据强化学习的特点，建立一种动态环境下的机器人实时路径规划方法。

### 3 动态环境的机器人自适应路径规划

自适应路径规划是机器人应用研究的一个重要方面，是机器人在给定的环境中在与环境的不断交互过程中，规划出一条从特定的起始点到目标点并且满足一定的最优标准的、与障碍物不能碰撞的路径。

通过前面的介绍，我们可以得知，对于基于遗传算法、模糊算法和神经网络算法等的人工智能路径规划方法都与优化算法相联系，其结果都是全局最优路径规划。人工智能对于环境信息提供不太完备的情形下使用比较合适，但最优化算法对于实时控制，由于其复杂性、耗时性等原因往往不太实用。

在机器人对环境状态完全感知的情况下，应用人工势场法进行路径规划比人工智能方法具有更高的效率，但这种方法存在局部极小问题，虽然许多研究人员对这一问题进行了研究并提出了有效的解决方法<sup>[118][119][120]</sup>，但对于实时的路径规划，这些方法还是显得复杂，对于存在移动障碍物的场合不太实用。

本文所提出的方法采用了与值函数逼近相集合的强化学习算法，是机器人在与环境的交互过程中通过适应性学习自动获得控制策略。强化学习是一种非监督的在线学习方法，它已经广泛应用于智能控制问题。它是环境状态集合到控制行

为集合进行映射的反应型控制方法,所以机器人是通过与环境的不断交互过程中得到环境信息和控制行为的映射关系来提高路径规划的效率。因为它不需要环境的精确模型,所以特别适合于解决动态环境的控制问题。

### 3.1 改进 CMAC 网络的 Q-Learning 系统

函数逼近器的离散化结构可以是各种形状的,但均匀的网络结构仍是最常用的方法。考虑到对象通常具有空间上的非一致性,即在不同的区域控制作用对系统输出的影响不同。如果对敏感度较大的区域做精细量化,而对其他区域只作相对粗略的量化,则逼近器的逼近效果势必更为理想。非均匀分割方法有多种,比如可以用进化算法来合理量化状态空间的非均匀分割,但这种方法需要大量的计算时间,影响系统的实时性,而这是实时路径规划中最主要的问题。本文根据领域知识,对状态空间的不同区域进行不同精度的量化。对于反映了机器人各个方向上的障碍物距离的传感器信息,机器人离障碍物距离越近时避障行为就越重要,因此对距离越小的状态越细分,对于趋向目标的行为,机器人运动方向与目标的夹角越小量化精度越高。对于反应环境详细信息的参数,结合非均匀 BOX 方法进行不同精度的量化。

Q-Learning 算法和它的特点前面已经介绍过,由于只对  $Q(x, a)$  值进行预测,它的使用比 actor-critic 系统要直观的多。由于最优策略的选择仅仅用贪婪方法,它不需要学习和存储一个策略。所谓的贪婪策略是指当行动值收敛到它们实际值的时候,选择行动值最大的行动。

$Q$  值被存储在从输入向量得到的状态向量  $x$  的 CMAC 网络中,由于每个离散的行动  $a$  需要一个 CMAC 网络,所以 CMAC 网络的数量是  $A = \max_x |A(x)|$  并存储所有的  $Q(x, a)$  值。

任一时刻智能体在状态  $x_t$  采取行动  $a_t$ , 对于状态  $x_t$  和行动  $a_t$  的行动值当前预测值  $\hat{Q}_t(x_t, a_t)$  的更新如下

$$\hat{Q}_{t+1}(x_t, a_t) \leftarrow \hat{Q}_t(x_t, a_t) + \alpha[r_t + \gamma \max_{l \in A(x_{t+1})} \hat{Q}_{t+1}(x_t, l) - \hat{Q}_t(x_t, a_t)]$$

(1)

$x_{t+1}$  是下一个状态,  $\gamma$  是折扣因子,  $\alpha$  是步长参数,  $A(x_{t+1})$  是状态  $x_{t+1}$  的可能执行行动,  $r_t$  是智能体从环境中得到的回报。其它状态和行动的状态-行动值保持不变。对行动值的预测偏差也可以用下式表示

$$\varepsilon_t^Q \leftarrow r_t + \gamma \max_{l \in A(x_{t+1})} \hat{Q}_{t+1}(x_t, l) - \hat{Q}_t(x_t, a_t)$$

(2)

量值  $\varepsilon_t^0$  在 Q-learning 和  $TD(\lambda)$  算法结合时使用。

在状态  $x$  选择行动  $a$  的概率通过 Boltzmann 分布来进行

$$p(a | x) = \frac{e^{\hat{Q}(x,a)/T}}{\sum_{l \in A(x)} e^{\hat{Q}(x,l)/T}}$$

(3)

以上算法系统的设计将强化学习算法和 CMAC 网络结合到了一起。

### 3.2 路径规划控制总体模型设计

基于强化学习的路径规划方法是一种未知环境下的实时规划方法，其总体控制模型设计如图 5 所示。机器人通过传感器感知环境，按一定的搜索策略如随机策略或贪婪策略，选择一个控制动作，并执行控制动作，使得系统状态转移并得到一个即时强化信号，机器人根据新的状态搜索新的控制动作，如此重复，直到到达目标点。机器人为获得最优路径所决策的原则是使得累积报酬最大，因此，机器人在每一个时间步进行决策的时候，不仅要考虑当前的即时强化，还要使期望的滞后强化尽可能地大。这使得机器人在局部感知的规划时考虑全局影响而试图获得全局的最优路径。该文将强化学习与基于行为的反射式控制相结合，将机器人的行为分解成趋向目标、固定障碍物避障、动态障碍物避障三种基本行为。在学习过程中，动作的选择考虑对三种基本行为的影响，使得三者协调融和，这反映在动作执行后这三种行为状态的改变对强化学习强化函数也就是强化信号的影响上。

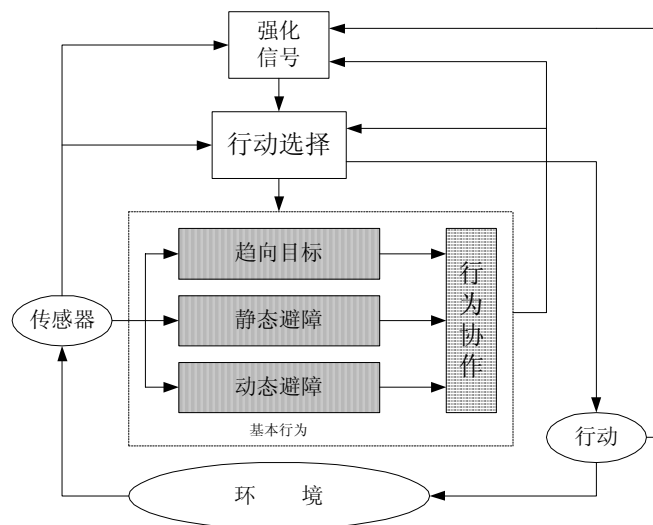


图 5 机器人自适应动态路径规划控制模型

### 3.2 学习系统的环境感知及行为模型

(1) 机器人静态障碍物避障模块设计:

为了即时而准确获得机器人所处环境的信息,对于解决静态障碍物的避障问题机器人的传感器系统设计如图 6

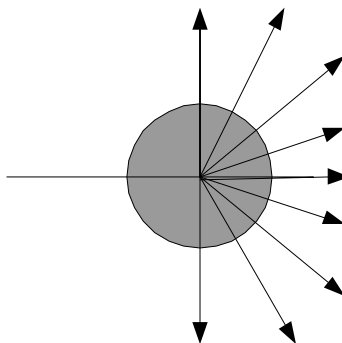


图 6 机器人距离传感器设置

机器人有 9 个测距传感器 (声纳或超声波), 其位置设置如图 6, 因为机器人没有向后退的动作, 所以机器人后部不设传感器, 传感器得到的障碍物距离为  $d_{roi} (i = 1, 2, \dots, 9)$ 。

(2) 机器人动态障碍物避障模块设计:

对于强化学习系统, 参数的选择必须能够代表环境的特征。根据机器人动态路径规划的特点, 机器人还有可以进行 360 度感知动态障碍物的一个传感器 (如图像传感器), 它可以通过传感器信息确定机器人与动态障碍物的距离、障碍物运动速度和运动方向, 机器人感知动态障碍物的参数如图 7 所示, 而且假设机器人在任何时候都可以通过特定的传感器获得这些参数, 这些参数包括:

$d_{ro}$  - 机器人和最近的动态障碍物间的距离;

$v_o, \theta_o$  - 最近障碍物的速度和相对于机器人运动的运行方向角度  $\theta_o$ ;

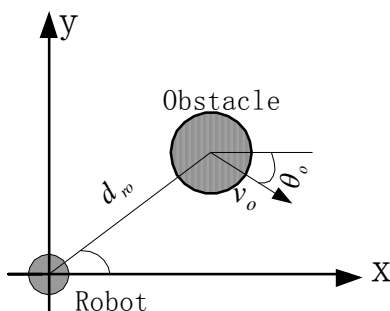


图 7 动态障碍物环境的状态描述参数

(3) 机器人趋向目标模块设计:

机器人趋向目标模块近考虑机器人运行方向与目标的相对角度  $\theta_{rg}$ ，对于机器人趋近目标是通过强化信号来反映的。

因此，机器人所处环境的状态向量为  $S = \{d_{roi}, d_{ro}, v_o, \theta_o, \theta_{rg}\}^T$ 。

(4) 机器人可执行行为空间

为了使机器人学习得到的路径光滑并符合实际机器人运行要求，我们将机器人动作空间分为 7 个离散的动作，为  $-40^\circ$ ， $-20^\circ$ ， $-5^\circ$ ， $0^\circ$ ， $5^\circ$ ， $20^\circ$  和  $40^\circ$  角度向前行驶一定距离。

### 3.3 强化函数的设定

从以上的介绍中我们可以知道，机器人路径规划学习系统是一个多目标系统。由于强化学习系统对静态障碍物和动态障碍物的环境信息获取方法的不同，为了训练的简化，我们将静态障碍物和动态障碍物的避障问题用两个模块来完成。所以对整个学习系统来说，机器人的学习目标有三个：第一个目标是避开静态障碍物，另一个目标是避开动态障碍物，最后一个目标是在最短时间内到达目标点，因此系统的强化信号也包括三个方面：

$$r_{os} - \text{机器人与静态障碍物碰撞与否}$$

$$\begin{cases} r_{os} = -1 & \text{if closer to static obstacle} \\ r_{os} = 1 & \text{if faraway from static obstacle} \\ r_{os} = -5 & \text{if collision with static obstacle} \\ r_{os} = 0 & \text{otherwise} \end{cases}$$

$$r_g - \text{机器人是否接近目标}$$

$$\begin{cases} r_g = 1 & \text{if robot closer to goal} \\ r_g = -1 & \text{if robot faraway from goal} \\ r_g = 0 & \text{otherwise} \end{cases}$$

$$r_{od} - \text{机器人与动态障碍物碰撞与否}$$

$$\begin{cases} r_{od} = -1 & \text{if closer to dynamic obstacle} \\ r_{od} = 1 & \text{if faraway from dynamic obstacle} \\ r_{od} = -5 & \text{if collision with dynamic obstacle} \\ r_{od} = 0 & \text{otherwise} \end{cases}$$

机器人从环境中得到的总回报值为

$$r_{og} = w_{os} r_{os} + w_g r_g + w_{od} r_{od}$$

其中  $w_{os}$ 、 $w_g$  和  $w_{od}$  是机器人相对于静态障碍物、目标点和动态障碍物回报的加权值， $w_{os} + w_g + w_{od} = 1$ ， $0 < w_{os} < 1$ ， $0 < w_g < 1$ ， $0 < w_{od} < 1$ 。

对于以上各模块的加权值的选定，可以根据定义中与动态障碍物和静态障碍物的相关性，对传感器可感知区域内是否感知到动态障碍物和静态障碍物都存在、之一存在或都不存在来赋予不同的权值，权值的大小应能反映不同模块对不同环境的重视程度。

## 4 系统学习结果及讨论

系统学习的任务是：从给定的初始、和目标位置开始，以尽可能少的时间(即步数)到达目标位置，并且不发生和障碍物的碰撞。在系统的学习过程中，到达目标、和障碍碰撞及达到预定的最大步数任何一个条件得到满足时为一个学习周期。每个学习步骤由以下几个环节组成：

- (1) 环境状态向量作为参数估计网络的输入，通过网络计算得到行动决策；
- (2) 根据决策随机选择控制行为；
- (3) 执行控制行为并根据增强信号和状态转换信息得到调整值和参数估计网络的误差值；
- (4) 训练状态评价网络，同时训练被执行的控制行为所对应的网络。

训练环境为在一个  $400 \times 400$ （像素）的环境中存在形状和位置未知的障碍物，要求机器人从已知起始点以安全和最佳到达目标点。对于距离传感器其最小感知距离为 20 像素，最大感知距离为 80 像素，根据距离远近对避障影响的不同，等分为三个不同的量化精度区域，分割区域分别为 4、2 和 1 个，因此每个传感器的分割区域为 8 个。对动态障碍物，因为要考虑到速度因素，因此其感知距离为 20 像素到 100 像素，分割区域为 8，角度和速度的分割区域分别为 6 和 4。对  $\theta_{rg}$  的分割也采用非均匀分割方法，分别分割为 5 和 3，总体分割为 8。CMAC 的感知域为 5。由于存储需求非常大，这里采用哈希编码。

训练时，指定机器人的起点、终点，并任意给定环境障碍物信息（包括动态障碍物和静态障碍物），为了简化训练过程，其中动态障碍物的速度随机选定为机器人速度的 0.5 到 1.5 倍，事实上只要训练时间和环境样本足够多的情况下，任何参数的环境设置，机器人都能够通过学习获得无碰撞的移动轨迹。

图 7 给出了用训练好的网络对静态环境的路径规划所得到的轨迹，在路径中随机增加一个动态障碍物，机器人仍能够成功地到达目标点，如图 8 所示，可见算法的有效性。

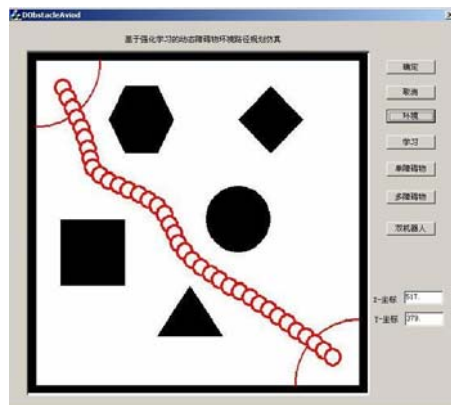


图 7 静态环境路径规划结果

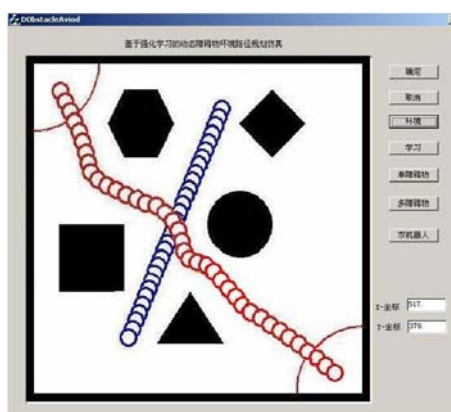


图 8 动态环境路径规划结果

在上面的训练学习中，为了对系统学习过程和效果进行进一步的研究，我们对学习过程进行了量化研究。为考察控制策略的优化情况，每经过 100 个学习周期，对到达目标所用的平均步数和成功到达目标的次数进行统计，根据实验数据分别得到图 9 和图 10。

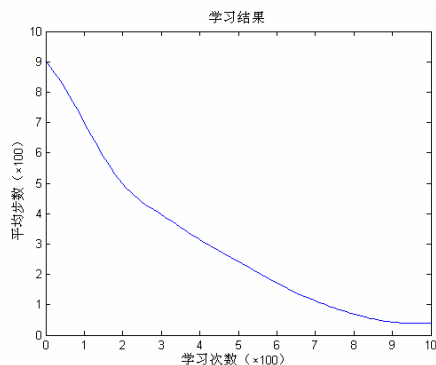


图 9 学习次数与平均步数关系

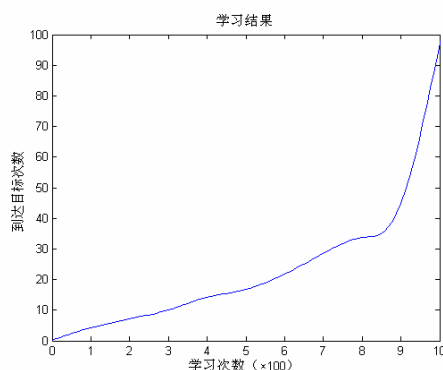


图 10 学习次数与得到目标次数关系

图 9 表明到达目标所用的平均步数随着学习呈减小趋势；图 10 表明每 100 个学习周期中成功到达目标的次数随着学习呈增加趋势。这说明控制策略随着学习不断得到优化。

## 5 小结

路径规划是智能机器人关键问题之一，它包括全局路径规划和局部路径规划，局部路径规划是路径规划的难点。当环境复杂时，很难得到好的路径规划结果，特别是动态环境中，传统的方法解决起来比较困难。在这种情形下，强化学习被认为是获取未知环境下自主机器人控制策略的合适的方法，这里将强化学习方法用于机器人学习控制，以实现在复杂动态环境下的机器人自适应路径规划。

本文提出了一种机器人基于强化学习在动态环境中的路径规划方法，通过利用改进的值函数逼近网络 CMAC，使得系统具有在连续状态中在线学习的特点，

因此使决策系统具有自适应性。通过自适应的学习, 机器人控制器建立起了环境状态到控制输出的直接映射, 可以使机器人在动态未知的环境中有效避开障碍物并以最短路径到达目标位置。从仿真的结果来看, 具有很好的适应性, 路径规划算法的设计相对来说也比较简单, 同时具有较好的实时性能, 是解决动态环境中路径规划问题的一种可行方法。

### 参考文献

- 
- [106] Khatib O., Real-time obstacle avoidance for manipulators and mobile robots. *Int J Robotics Research*, 5(1):90-98, 1986.
- [107] J. Lengyel, M. Reichert, B. Donald, and D. Greenberg: Real-time robot motion planning using rasterizing computer graphics hardware. *Proceedings of SIGGRAPH*, pages 327-335, August 1990.
- [108] K. Konolige: A Gradient Method for Realtime Robot Control. *Proc. of the IEEE/RSJ IROS 2000*.
- [109] L. Kaelbling, A. Cassandra, and J. Kurien, Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. *Proceedings of the IEEE/RSJ IROS 1996*.
- [110] P.E. Hart, N.J. Nilsson and B. Raphael, A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2): 100-107, 1968.
- [111] A. Schweikard, A simple path search strategy based on calculation of free sections of motions. *Engineering Applications of Artificial Intelligence*, 5 (1) 1 - 10, 1992.
- [112] J.-C. Latombe, *Robot Motion Planning*. Kluwer Academic Publishers, 1991.
- [113] S. Buck, R. Hanek, M. Klupsch, and T. Schmitt, Agilo RoboCuppers: RoboCup Team Description. *RoboCup 2000: Robot Soccer World Cup IV*, Springer, 2000.
- [114] Hartmut Surmann, Jens Wehking, Path planning for a fuzzy controlled autonomous mobile robot. *Fifth IEEE Int. Conf. On Fuzzy Systems Fuzz2IEEE' 96*. USA: New Orleans, 1996.
- [115] E.S. Plumer, Neural network structure for navigation using potential fields, in *Proc.*

- Int. Joint Conf. Neural Networks (IJCNN-92), vol. 1,pp. 327-332, 1992.
- [116] Holland J.H., Genetic algorithms and the optimal allocations of trails, SIAM Journal of Computing, 2(2):88-105, 1973.
- [117] Tsoukalas L.H , Houstis E.N , Jones G.V. Neuro-fuzzy motion planners for intelligent robots. Journal of Intelligent and Robotic Systems , 19 :339-356, 1997.
- [118] Akishita S., Hisanobu T. and Kawamura S., Fast path planning available for moving obstacle avoidance by use of Laplace potential, Intelligent Robots and Systems '93, IROS '93. Proceedings of the 1993 IEEE/RSJ International Conference on Volume: 1, Page: 673 -678 , 1993.
- [119] Makita Y., Hagiwara M. and Nakagawa M., A simple path planning system using fuzzy rules and a potential field, Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence, Proceedings of the Third IEEE Conference on, vol.2 Page: 994 -999, 1994.
- [120] Kun Hsiang Wu; Chin Hsing Chen and Jiann Der Lee, Genetic-based adaptive fuzzy controller for robot path planning, Fuzzy Systems, Proceedings of the Fifth IEEE International Conference on Volume: 3, Page: 1687 -1692, 1996.