基于增强式学习的仿生机器鱼避障控制①

沈志忠②*,** 曹志强** 谭 民** 王 硕**

(*北京科技大学信息工程学院 北京 100083)

(**中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100080)

摘 要 研究了仿生机器鱼的自主避障控制,提出了一种集成超声和红外传感器的仿生机器鱼系统设计方案,为机器鱼安装了3个超声传感器和3个红外传感器。针对此仿生机器鱼系统,提出了一种基于增强式学习的仿生机器鱼自主避障控制策略,给出了状态集和行为集,采用当场奖励和延时奖励相结合的方法,通过学习获得了有效的状态行为组合。仿真实验验证了学习结果的有效性。

关键词 仿生机器鱼,避障,学习

0 引 言

鱼类作为自然界最早出现的脊椎动物,经过亿万年的自然选择,进化出了非凡的水中运动能力,将鱼类的优点和机器人技术结合为人类研制新型水下推进器提供了新的思路[1]。仿生机器鱼就是参照鱼类游动的推进机理,利用机械、电子元器件或智能材料来实现水下推进的一种运动装置。它可以在水下目标搜索、追踪等方面发挥作用,具有潜在的应用前景。从1994年MIT的Triantafyllou研究组成功研制了世界上第一条真正意义上的仿生机器鱼RoboTuna[2]之后,美国、日本出现了一些不同目的的仿生机器鱼实验平台或原理样机。国内多家研究单位也开展了对机器鱼的研究工作,取得了很多研究成果。北京航空航天大学机器人研究所提出了"波动推进理论"及其分析方法,设计研制了仿生"机器鳗鱼"实验模型,之后又研制了仿生"机器海豚",获取了鱼的摆动推进深层次机理[3];哈尔滨工程大学研制了仿生机器章鱼[4];哈尔滨工业大学开展了水下机器人仿鱼鳍推进机理的研究[5];中科院沈阳自动化研究所研制了两关节的仿生机器鱼模型;中科院自动化所复杂系统与智能科学重点实验室研制开发了多种具有不同功能的仿生机器鱼,并对多仿生机器鱼的协调和协作展开了研究,建立了多仿生机器鱼系统(multiple robot fish system)。

为了更好地控制仿生机器鱼,Barett[6]基于遗传算法开发了自优化的运动控制器,利用进化原理来搜索机器鱼运动学模型的7个主要参数的值,以获取最高的推进效率。Morgansen[7]应用非线性控制理论的方法来产生系统输入,实现了平面瘿科机器鱼简单的轨迹跟踪。目前研究的大多数仿生机器鱼需要接受上位机的控制,缺乏有效的信息获取手段,无法适应复杂多变的水环境。因此,研究集成多种传感器的仿生机器鱼系统,提高机器鱼的自主运动能力和避障功能,具有非常重要的意义。

本文针对实际应用需要,开发了集成超声和红外传感器的仿生机器鱼系统,为了有效地处理这些传感信息,在已有研究成果的基础上,提出了一种基于增强式学习的仿生机器鱼自主避障策略,给出了状态集和行为集,采用当场奖励和延时奖励相结合的方法,通过学习获得了有效的状态-行为组合。

1 增强式学习

增强式学习,又叫强化学习或再励学习,是一种实时的、在线的学习方法。它采用试错法(trial-and-error),不用建立环境和任务的精确数学描述。因此,我们不需要告诉机器鱼如何达到它的目标,只需告诉它目标是什么。通过学习,机器鱼能够从获取的关于系统状态、动作、奖励的有用的经验中掌握一套优化的策略、知识。图1是一个针对仿生机器鱼的增强式学习模型[8],通常由以下几部分组成:环境状态集合 $S(s\in S)$,机器鱼行为集 $B(b\in B)$和增强信号 $r(R$ 为增强函数),此外还有环境。

根据不同的问题,增强式学习算法有很多种,常

① 863计划(2004AA4201104,2005AA420040)和国家自然科学基金(50475179)资助项目。

② 男,1978年生,博士;研究方向:仿生机器鱼;通讯作者,E-mail:zzshen78@163.com

(收稿日期:2005-11-18)

用的有 Dyna 算法、Q 学习算法、Sarsa 学习算法等。本文采用 Q 学习和 Sarsa 算法进行策略学习。

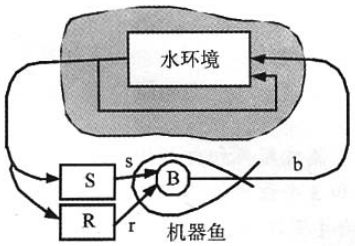

图 1 仿生机器鱼增强式学习模型

Q 学习算法无需任何模型，Q 值的更新是算法的核心，描述如下[9]：

步骤 1：对于当前状态 s，按一定的策略选取行为 b。

步骤 2：机器鱼执行被选取的行为，状态转换为 s'。同时机器人从环境中获得一个奖赏 r，于是得到 $\{s, b, s', r\}$。

步骤 3：更新 $Q(s,b)$ 值：

$$Q(s,b) := Q(s,b) + \alpha \cdot (r + \gamma \cdot \max_u Q(s',u) - Q(s,b))$$

其中 α 是学习率，γ 是再励值。

Sarsa 算法同 Q 学习算法都属于 TD(temporal difference)学习算法，不同之处在于前者属于有策略(on-policy)学习，而后者属于无策略(off-policy)学习。有策略学习的特点在于，它在选择动作和更新状态动作值时采用的是相同的策略；而无策略学习可以不同。例如，上面介绍的 Q 学习算法，可以采用贪婪(greedy)策略更新 Q 值，同时又不妨选择 ε-贪婪(ε-greedy)策略用于动作选择。Sarsa 算法的 Q 值更新方程可以描述如下：

$$Q(s,b) := Q(s,b) + \alpha(r + \gamma \cdot Q(s',b') - Q(s,b))$$

其中，$Q(s',b')$ 的选取按照与动作选取相同的策略进行。

2 基于增强式学习的仿生机器鱼避障控制

机器鱼避障控制的任务可以描述如下：

机器鱼在一个存在障碍物的环境中游动，超声和红外传感器获得环境信息，通过对这些信息的处理，机器鱼可执行不同的运动模式以实现无碰自主游动。

2.1 仿生机器鱼模型

为了使机器鱼得到更多的环境信息，在机器鱼上需安装多个超声传感器，考虑到机器鱼的体积以及任务的需求，为机器鱼安装 3 个超声传感器，每个超声传感器的波束角是 30 度。另外，为了弥补超声传感器的盲区，为机器鱼安装 3 个红外传感器，同时作为应急情况下采取动作的信号来源。超声传感器和红外传感器的布置方式如图 2 所示。

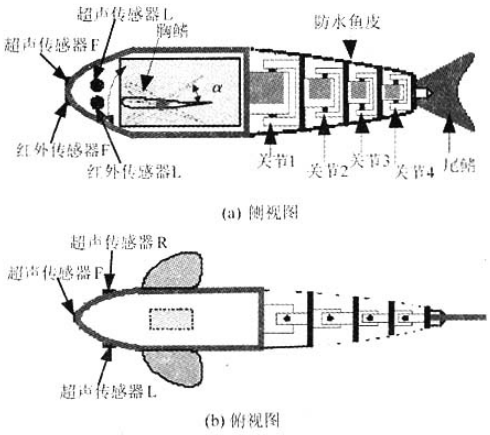

图 2 仿生机器鱼模型

2.2 机器鱼的状态集和行为集

一个超声传感器和一个红外传感器组合可以获得某一方向的相对完整信息，为了防止机器鱼在距离障碍物很远的地方就执行避障动作，使机器鱼能够更好地实现自主游动，以便执行目标搜索、目标追踪等任务，在距离障碍物较远的情况下，机器鱼不必采取避障行为，这样就扩大了机器鱼探测环境的范围。

本文对传感器所获数据进行了分段处理。假定超声传感器的最大探测距离为 L_m，执行有效避障行为的临界距离为 L_v，机器鱼到障碍物的危险距离为 L_h，假定超声传感器的盲区和红外传感器的探测距离相等为 L_I，这四者的距离关系是 $L_m > L_v > L_h > L_I$。对于障碍物与机器鱼之间距离小于 L_I 的应急情况，采用基于红外传感器的避障策略[10]。

根据超声传感器的数据值 L_d，将机器鱼某方向的障碍物分布情况分为 s_0, s_1, s_2 等 3 种状态，这里假定超声传感器没有探测到障碍物时 L_d 为一足够

大的数。这 3 种状态如下:

s_0 机器鱼没有探测到障碍物或者障碍物距离机器鱼很远,即 $L_d > L_m$ 或 $L_d > L_v$;

s_1 机器鱼距离障碍物较远,即 $l_v > l_d > l_h$;

s_2 机器鱼距离障碍物较近,即 $l_h > l_d > l_I$。

为了叙述方便,本文把 s_0 状态简单理解为在该方向不存在障碍物。每一个方向的传感器有 3 种状态,3 个超声传感器的状态组合为 $ZT_LZT_FZT_R$,其中 ZT_L 为超声传感器 L 的状态,ZT_I 为超声传感器 F 的状态,ZT_R 为超声传感器 R 的状态,ZT_L,ZT_F,ZT_R 在 s_0,s_1,s_2 中取值。例如:$s_0s_1s_0$ 表示机器鱼前方的传感器探测到障碍物,且距离较远,其它方向没有探测到障碍物。

机器鱼可能的状态有 27 种,然而 3 个方向的超声传感器状态对机器鱼的影响是不一样的,由于机器鱼的游动方向主要表现为前向运动,所以在机器鱼的前方有无障碍及其距离对机器鱼运动的影响是最重要的。为了减少状态集的数量和提高学习速度,本文对 27 种状态进行了合并,合并规则如下:

$ZT_{L(R)} = s_1$

(当 $ZT_{L(R)} = s_1$ 或 $ZT_{L(R)} = s_2$ 时)

合并后的状态共有 12 种,见表 1。

表 1 合并后的状态集

状态	描述	状态	描述
$S_1 - s_0s_0s_0$	没有障碍物	$S_2 - s_0s_0s_1$	右侧有障碍物
$S_3 - s_0s_1s_0$	前方有障碍物且距离头部较远	$S_4 - s_0s_1s_1$	前方有障碍物且距离头部较远,右侧有障碍物
$S_5 - s_0s_2s_0$	前方有障碍物且距离头部较近	$S_6 - s_0s_2s_1$	前方有障碍物且距离头部较近,右侧有障碍物
$S_7 - s_1s_0s_0$	左侧有障碍物	$S_8 - s_1s_0s_1$	左侧和右侧有障碍物
$S_9 - s_1s_1s_0$	左侧有障碍物,前方有障碍物且距离头部较远	$S_{10} - s_1s_1s_1$	左侧和右侧均有障碍物,前方有障碍物且距离头部较远
$S_{11} - s_1s_2s_0$	左侧有障碍物,前方有障碍物且距离头部较近	$S_{12} - s_1s_2s_1$	左侧和右侧均有障碍物,前方有障碍物且距离头部较近

考虑到机器鱼的性能,本文为机器鱼设计了如下行为:

b_1 机器鱼速度为零向右静止转弯;

b_2 机器鱼速度为零向左静止转弯;

b_3 机器鱼以速度 V 向右转弯;

b_4 机器鱼向前直游;

b_5 机器鱼以速度 V 向左转弯;

b_6 漫游。

上述行为中,$b_1 \sim b_5$ 是机器鱼学习的目标,而 b_6 并不需要机器鱼通过学习获取,它作为机器鱼本身所应该具有的基本能力,当机器鱼的 3 个传感器都没有探测到障碍物时,机器鱼自动选择该行为。

2.3 奖励

由于超声传感器不能对障碍物进行精确的方向定位,在波束角范围内的物体均产生信号,为此综合机器鱼 3 个超声传感器的信号情况,将障碍物相对于机器鱼的方向角离散化为 7 种情况,作如下定义:

$$\Psi = \begin{cases} 30^\circ & ZT_L = s_1 \cap ZT_F = s_0 \cap ZT_R = s_0 \\ 15^\circ & ZT_L = s_1 \cap (ZT_F = s_1 \cup ZT_F = s_2) \cap ZT_R = s_0 \\ 0^\circ & ZT_L = s_0 \cap (ZT_F = s_1 \cup ZT_F = s_2) \cap ZT_R = s_0 \\ -15^\circ & ZT_L = s_0 \cap (ZT_F = s_1 \cup ZT_F = s_2) \cap ZT_R = s_1 \\ -30^\circ & ZT_L = s_0 \cap ZT_F = s_0 \cap ZT_R = s_1 \\ -30^\circ \cup 30^\circ & ZT_L = s_1 \cap ZT_F = s_0 \cap ZT_R = s_1 \\ -30^\circ \cup 0^\circ \cup 30^\circ & ZT_L = s_1 \cap (ZT_F = s_1 \cup ZT_F = s_2) \cap ZT_R = s_1 \end{cases}$$

假定 3 个超声传感器探测到的距离值分别为 d_L,d_F,d_R,则障碍物相对机器鱼的距离 d 定义为:

$$d = \min(d_L, d_F, d_F)$$

基于障碍物相对机器鱼的方向角为 Ψ,当机器鱼执行行为的方向偏离障碍物相对机器鱼的方向时,称机器鱼相对于障碍物方向偏离;反之称机器鱼相对于障碍物方向靠近。

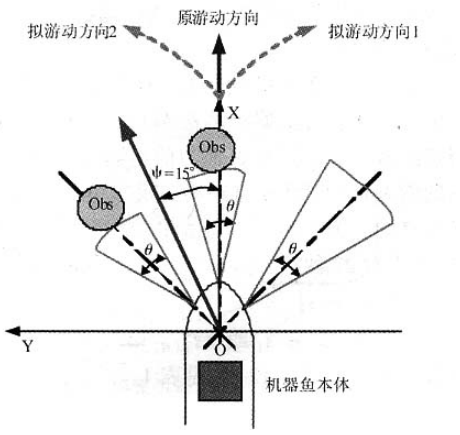

图 3 方向偏离和方向靠近描述

图 3 给出了一种机器鱼相对于障碍物游动的示意图,如果机器鱼执行的行为使得机器鱼按拟游动方向 1 运动,则为机器鱼相对于障碍物方向偏离;反之,如果机器鱼执行的行为使得机器鱼按拟游动方向 2 运动,则为机器鱼相对于障碍物方向靠近。

如何奖惩机器鱼是增强式学习中的一个重要环节,它影响到学习的好坏、快慢。在文献[11]中,Tucker Balch 给出了一套描述符,对增强式学习的奖励进行了分类。对于机器鱼的避障任务,本文采用的是当场奖励和延时奖励相结合的办法。

假定机器鱼当前状态为 s_t,前两个状态分别为 s_{t-1}和 s_{t-2},按照一定的策略选择行为 b,状态由 s_t 转换为 s_{t+1},则下面情况能够产生正增强(positive reinforcement):

情况 1 机器鱼从有障碍状态运动到没有障碍状态;

情况 2 机器鱼执行动作后方向偏离障碍物;

情况 3 机器鱼头部距离障碍物较远时,对能够扩大探测范围的动作进行奖励。

下面情况产生负增强(negative reinforcement):

情况 4 机器鱼执行动作后方向靠近障碍物;

情况 5 机器鱼头部距离障碍物较近时,执行动作后机器鱼相对于障碍物的距离减少。

以上 5 种奖励均为当场奖励(immediate reward)。

在满足第一种奖励的前提下,对状态之前的两个状态行为组合进行奖励(情况 6),这是延时奖励(delayed reward)。

3 仿真结果

仿真学习环境如图 4 所示,图中的黑色圆为障碍物,在学习过程中边界也当作障碍物来处理。本文分别采用 Q 学习和 Sarsa 算法来进行状态-行为的学习。在 Q 学习中,行为选取采用贪婪法,Q 值更新采用 ε-贪婪法。ε 初始值 ε_i 设为 0.95,ε 每代递减$\frac{\varepsilon_i - 0.05}{T}$,其中 T 为学习总代数,随着学习的进行 ε 逐渐降低至 0.05。学习率 α 和再励值 γ 分别被设为 0.03 和 0.9,学习总代数 T 为 65,每代最大运动步数是 3100 步。

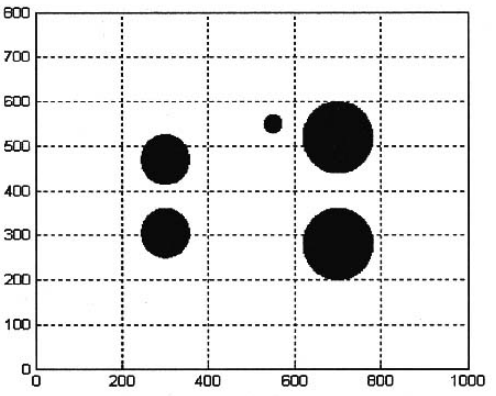

图 4 仿真环境

采用当场奖励和延时奖励相结合的办法进行学习,具体奖励值设定如下:

$$R = \begin{cases} 6 & \text{满足情况 1 时} \\ 10 & \text{满足情况 2 时} \\ 5 & \text{满足情况 3 时} \\ -15 & \text{满足情况 4 时} \\ -5 & \text{满足情况 5 时} \end{cases}$$

当满足情况 6 时,对前两个状态行为组合的奖励值分别为 6 和 3。

在 Sarsa 算法中,行为选取和 Q 值更新均采用 ε-贪婪法,算法中的参数值以及奖励值 R 和 Q 学习相同。

Sarsa 算法和 Q 学习算法的学习结果在表 2 中给出,括号内为 Q 学习的结果,可见两种算法的学习结果基本相同,其中对于状态 S_3 和 S_{10}两种算法学习结果不同。通过分析可知,状态 S_3 学习到 b_3 或 b_5,状态 S_{10}学习到 b_1 或 b_2 都认为是合理的。综

合学习结果可以得出:

(1) 左侧有障碍机器鱼执行右转策略,右侧有障碍机器鱼执行左转策略,两侧有障碍直行;

(2) 当机器鱼距离障碍物较远时,机器鱼执行前进中转弯策略进行避障,当障碍物进入机器鱼的危险区时,机器鱼执行静止转弯策略避障。

表 2 Sarsa 算法和 Q 学习算法的学习结果

状态	行为
S_{11}、S_{12}、S_{10}(S_{11}、S_{12})	b_1
S_5、S_6(S_5、S_6、S_{10})	b_2
S_7、S_9、S_3(S_7、S_9)	b_3
S_8(S_8)	b_4
S_2、S_4(S_2、S_4、S_3)	b_5

策略改变次数是算法收敛性的重要评价指标,该标准用于绘制出机器鱼策略改变的次数随训练次数的变化情况,图 5 给出了两种学习算法的策略改变次数随学习代数变化的曲线,可以看出,算法是收敛的。

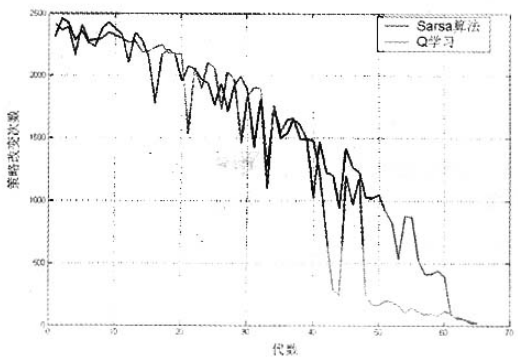

图 5 Sarsa 算法和 Q 学习算法的策略改变次数曲线

对利用 Sarsa 算法和 Q 学习算法学习到的策略进行了仿真实验,在 1000×800 大小的范围内采用所学到的策略让机器鱼进行自主运动,图 6 给出了采用这两个结果的典型运动轨迹。可以看出,机器鱼在仿真环境中实现了无碰运动。

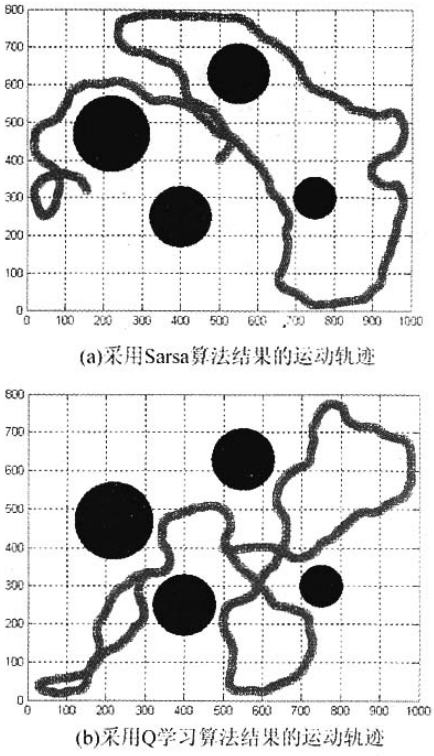

图 6 典型运动轨迹

4 结 论

机器鱼作为一种新型水下潜器,具有广阔的应用前景,机器鱼可以在水下目标搜索、目标追踪等方面发挥作用。本文针对我们研制开发的一种集成多个超声和红外传感器的仿生机器鱼系统,提出了一种基于增强式学习的机器鱼避障控制策略,机器鱼不需要知道外界精确模型,可以通过学习优化运动策略。本文对传感器的状态进行了组合,减少了状态集的数量,提高了学习速度,仿真实验验证了学习结果的有效性。

参考文献

[1] Colgate J E, Lynch K M. Mechanics and control of swimming: a review. *IEEE Journal of Oceanic Engineering*, 2004, 29(3): 660-673

[2] Triantafyllou M, Triantafyllou G S. An efficient swimming machine. *Scientific American*, 1995, 272(3): 64-70

[3] 梁建宏, 王田苗, 魏洪兴等. 水下仿生机器鱼的研究进展 II——小型实验机器鱼的研制. 机器人, 2002, 24(3): 234-238

[4] 彭之春，庞永杰. 机器鱼的运动仿真方法. 系统仿真学报，2004，16(12)：2643-2646

[5] 刘军考，陈在礼. 水下机器人新型仿鱼鳍推进器. 机器人，2000，22(5)：427-432

[6] Barrett D，Grosenbaugh M，Triantafyllou M. The optimal control of a flexible hull robotic undersea vehicle propelled by an oscillating foil. In：Proceedings of the IEEE Symposium on Autonomous Underwater Vehicle Technology，1996：1-9

[7] Morgansen K A，Duindam V，Mason R J，et al. Nonlinear control methods for planar carangiform robot fish locomotion. In：Proceedings of the 2001 IEEE International Conference on Robotics and Automation，2001：427-434

[8] Kaelbling L P，Littman M L，Moore A W. Reinforcement learning：a survey. *Journal of Artificial Intelligence Research*，1996，4：237-285

[9] Watkins C J C H，Dayan P. Q-learning. *Machine Learning*，1992，8(3)：279-292

[10] 桑海泉，王硕，谭民等. 基于红外传感器的仿生机器鱼自主避障控制. 系统仿真学报，2005，17(6)：1400-1404

[11] Tucker B. Behavioral diversity in learning robot teams：[Ph. D. thesis]. Atlanta：Georgia Institute of Technology，1998

Control of obstacle avoidance of biomimetic robot fish based on reinforcement learning

Shen Zhizhong[*,**]，Cao Zhiqiang[**]，Tan Min[**]，Wang Shuo[**]

([*] School of Information Engineering，University of Science and Technology Beijing，Beijing 100083)

([**] Laboratory of Complex Systems and Intelligence Science，Institute of Automation，Chinese Academy of Sciences，Beijing 100080)

Abstract

Autonomous control of obstacle avoidance of biomimetic robot fish is studied. The design of a kind of biomimetic robot fish system with ultrasonic and infrared sensors is given. Three ultrasonic sensors and three infrared sensors are equipped with the robot fish. Aiming at this robot fish system，an obstacle avoidance strategy based on reinforcement learning is proposed. The state and behavior sets are brought forward. Through adopting immediate and delayed rewards，valid state-behavior pairs are obtained through learning. Computer simulations show the validity of the learning results.

Key words：biomimetic robot fish，obstacle avoidance，learning