Proceedings of the 33rd Chinese Control Conference
July 28-30, 2014, Nanjing, China

基于强化学习和意图推理的目标跟踪

李洁, 苏剑波
上海交通大学自动化系, 教育部系统控制与信息处理重点实验室, 上海 200240
E-mail: {lijiunderstand, jbsu}@sjtu.edu.cn

摘　要: 基于视觉传感的特定目标跟踪由于视觉信息处理存在迟滞,使得目标跟随的及时性和性能都受到很大的影响.本文提出一种新的方法, 通过灰色预测模型对目标运动意图进行建模,修正机器人动作来补偿不确定时延对系统性能造成的影响. 首先,考虑环境条件和目标跟踪的非线性非高斯的特点,引入粒子滤波算法进行运动的预测估计,得到目标物体在视觉传感视野中的位置状态; 进而基于强化学习算法不需要环境模型的特点,结合实际场景需要,设计加入人意图推理的强化学习算法,得到目标物体位置状态到跟踪动作空间的映射.本文提出的思路和方法在NAO仿人机器人平台上实现了跟随人运动的功能.实验验证了该方法的有效性和可靠性.

关键词: 灰色预测; 粒子滤波; 强化学习; 意图推理; NAO

Tracking Strategy Based on Reinforcement Learning and Intention Inference

LI Jie, SU Jianbo
School of Electronic Information and Electric Engineering , Shanghai Jiao Tong University, Shanghai 200240, P. R. China
E-mail: {lijiunderstand, jbsu}@sjtu.edu.cn

Abstract: Due to hysteresis in the visual information processing, the timeliness and performance is affected in the specific target tracking problem. This paper investigates a novel approach, target motion modeling through the grey prediction model, to compensate the uncertainty with the correction action of the robot. Firstly, with consideration of the nonlinear and non-gaussian characteristics of the environment and target tracking, the particle filter algorithm is introduced for motion prediction and estimation,and the target position in the robot vision can be calculated. Then, an intention inference based reinforcement learning control method is used to estimate the mapping from sensory information to appropriate robot action based on the characteristics of the method,which, do not need the model of the environment and the robot.At last, the method is used in humanoid platform NAO to realize robust people tracking problem. Experimental results show the validity of the proposed method.

Key Words: grey prediction; particle filter; reinforcement learning; intention inference; NAO

1　引言

人类的生活环境复杂多变,在复杂的环境中跟踪特定目标或人是机器人应用研究和人机交互领域的热点之一,具有重要的学术和应用价值.跟随目标行走并为其提供持续的服务已经成为服务机器人的一项重要能力[1−2].文[3]设计了一个仿人机器人在日常生活环境中跟随人类行走并照顾老人和小孩. 已有移动机器人为医护人员提供电子帮助的报导[4], 能够自主跟随医生并为其提供表格、物资以及患者资料等数据.文献[5]将移动机器人跟随技术用于机场环境来帮助旅客搬运行李.在复杂多变的环境中实时地完成跟踪任务需要考虑很多问题,例如光照的任意变化对视觉传感的影响,其他人或目标很可能在目标附近运动形成干扰等.因此,机器人对特定目标或人的稳定实时的跟随,就具有重要的实用和理论价值.而对特定人的跟踪因为人的外形的千差万别和随机可变性,以及运动的随机、随意性,又有特殊的难度,因而构成了相对独立的学术研究领域[6−8].

目前关于移动机器人利用视觉信息来跟随人的运动已经有很多研究[9−12],这些方法侧重于利用实时的图像匹配处理技术获取目标人的位置信息, 并控制机器人运动.这种简单的控制方法没有考虑机器人的移动步长导致机器人的跟随左右摇摆,不稳定.已有研究利用PID去控制机器人的移动[13],但PID参数一旦整定后,在整个控制过程中都是固定不变的,无法适应于实际应用中各种变化的情况. 文献[14]提出建立一个查找表用来储存状态和动作的映射完成机器人的跟随.但查表法仅仅适用于有限的情况,不能处理连续状态与连续动作空间的映射问题.

自然界中人类适应环境的能力主要来自于学习和进化,即对环境进行感知,并作出决策和行动,强化学习符合人类解决问题的心理习惯. 强化学习是智能体从环境状态到行为映射的学习,以使得累积奖励回报函数值达到最大[15].该方法不需要提供学习范例,而是通过与环境的交互和感知,在试探的过程中,根据环境反馈的信息学会达到其目标的最优动作或动作序列.Q学习方法是常见的强化学习方法,其易于理解,应用广泛[16].目前强化学习算法已经在机器人学中广泛使用,使系统具有灵活性和适应性.强化学习算法可以解决状态空间到动作空间映射的组合爆炸, 具有较强的在线自适应性和对复杂系统的自学习能力.将强化学习的方法引入移动机器人跟随人的任务中,能够处理连续状态连续

本文得到国家自然科学基金资助(61221003).

动作的映射问题, 使机器人实现对未知环境的探索,适应动态多变的环境.文献[17]使用Q学习算法实现了在连续状态空间下移动机器人对运动小球的跟踪,但强化学习算法仅能得到某一时刻状态静止时对应的动作序列,缺乏对下一时刻状态的预测,使得跟随的及时性和性能受到影响.

复杂环境中机器人的跟随很大程度上依赖于视觉传感信息,而视觉信息处理存在一定的迟滞,使得目标跟随的性能受到影响.人的行走具有一定的连续性和惯性,在很短的时间内,根据以往观测状态建立预测模型,可以得到人某一时刻的基本运动方向和速度.将人行走的意图推理加入控制算法修正动作序列,能够减小时延造成的跟随扰动,改进用户与机器人之间的交互.目前人机交互领域常用的意图推理方法是贝叶斯方法.贝叶斯方法是一种概率推理方法,能够学习变量之间的因果关系,从而建立起对问题领域的理解,预测可能的结果[18].但是其需要先验概率和似然函数,并且不能处理不确定性问题.人的行走意图具有一定的不确定性,灰色预测是一种对含有不确定因素的系统进行预测的方法[19], 能够对人的行走进行建模从而实现人的意图推理,目前使用最广泛的是GM(1,1)模型.新陈代谢GM(1,1)模型是在建模时置入新信息,并去掉一个最老的信息建立GM(1,1)模型,如此反复,依次递补,直到完成预测目标.此模型能及时考虑系统发展中的扰动信息,在补充新信息的同时能够去掉随时间推移而意义减少的老信息,更能反映系统当前特征.

综上所述,本文提出了一种基于强化学习和意图推理的目标跟踪方法,首先通过粒子滤波算法实现了对目标人的实时跟踪,得到了人在机器人视野中的位置状态信息,然后通过强化学习的方法,得到位置状态与机器人动作空间的一个映射, 使得机器人可以通过与环境的不断交互,感知环境状态并通过在线学习能够采取最优动作达到目标状态,让目标物始终位于机器人视野的中央,从而实现机器人自主跟随人的任务. 由于强化学习得到的是目标物体位置状态不变时所应采取的最优动作或者动作序列,忽略了在人机交互过程中人是运动的事实, 使得机器人在跟踪人快速移动时产生较大的误差. 意图推理通过新陈代谢GM(1,1)模型对人的状态序列进行建模,加入跟踪目标在瞬时时刻的方向和速度信息来修正强化学习得到的动作,减小了系统时延,达到了较好的跟踪效果.

文章结构为第二节简单介绍了控制方法的框架结构,第三节详细说明了加入人的意图推理的强化学习方法.第四节介绍了系统实现与实验结果及分析. 第五节对文章进行总结.

2 算法描述

为了完成跟随人运动的任务,首先要得到目标人的位置信息. 粒子滤波利用非参数化的蒙特卡罗方法实现贝叶斯滤波,可以有效解决实际环境中受光照强弱、目标姿态变化、大面积遮挡等因素而导致的目标难以识别并定位的问题[20].

颜色特征具有较强的适应性,对旋转、尺度等都具有不变性,因此特别适用于跟踪任务. 采用核函数下的颜色直方图作为量测基础,使得距离目标中心越近的像素点被赋予越大的权值,以提高目标描述的准确性.由于室内环境中光照变化对图像的RGB模型影响较大,可以选用更能反映图像色彩本质特性的HSV模型的色度分量做处理.由于HSV中三个分量是完全独立的,计算量比RGB模型要同时考虑三个分量更小.

在粒子滤波算法得到跟踪目标在机器人视场中的位置信息后,机器人的跟踪控制框架如图1所示,主要有3个步骤:(1)粒子滤波算法实现对目标人的实时跟踪,得到目标在机器人视野中的位置信息; (2)利用强化学习算法得到位置状态信息和机器人动作空间的映射;(3)使用基于对人的意图推理得到的信息去修正强化学习的结果,得到最终的控制动作.

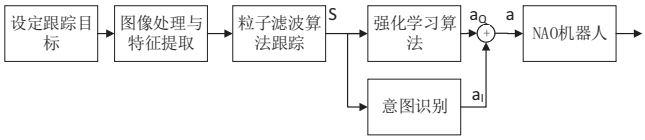


图1 基本算法组成

3 加入意图推理的强化学习方法

在机器人进行目标跟踪的过程中,为了得到从视觉反馈的目标位置状态到机器人运动动作的映射,利用强化学习的思想,并结合人机交互过程中人运动变化的特点来设计加入意图推理的强化学习方法.本节先给出基本的强化学习方法,然后再讨论加入意图推理的强化学习方法.

3.1 基本强化学习方法

强化学习的目标就是学习一个最优的策略π^*,使得智能体在状态s时采取最优行为a,使其导致的奖惩$r(s,a)$与后续的状态s'的评价函数$V^*(s')$的折扣之和达到最大:

$$\pi^*(s) = \underset{a\in A(s)}{argmax} \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V^*(s')]. \qquad (1)$$

其中,$P^a_{ss'}$为状态s下选择行为a,使状态变化到下一状态s'的状态转移概率; $R^a_{ss'}$为状态s下选择行为a,环境变化到s',系统获得的奖惩的期望值; $V^*(s')$为系统在最优策略π^*下系统获得的累积折扣奖惩的期望:

$$V^*(s') = E^*\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s'\right\}. \qquad (2)$$

由于$P^a_{ss'}$和$R^a_{ss'}$需要通过先验知识或系统模型来获得,限制了式(1)的应用范围. Q学习算法通过更新规则学习每个状态动作对所对应的Q值,来获得一个最优的策略,避免了$P^a_{ss'}$和$R^a_{ss'}$的计算, 其实现方法是按照递归公式进行的,考虑到实时跟随任务的要求,本文采用加入替代迹的基于人机交互的$Q(\lambda)$学习算法:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a). \qquad (3)$$

$$\delta_t = r_{HCI} + \gamma \underset{a'}{max} Q_t(s',a') - Q_t(s,a). \qquad (4)$$

其中,α为学习率（一般取$0 < \alpha < 1$）,γ为折扣因子,r_{HCI}是通过人机交互给予的智能体的瞬时奖惩,使

智能体获得当前状态与目标状态距离的度量,提高在状态空间中进行搜索的效率. 通过人对智能体行为的理解,根据不同的评价标准对当前状态与目标状态的偏差进行度量,给出各种奖惩的加权和作为奖惩. $e_t(s,a)$为替代迹：

$$e_t(s,a)=\begin{cases}1 & s=s_t\&a=a_t\\ \gamma\lambda e_{t-1}(s,a) & others\end{cases} \tag{5}$$

式中,λ为迹的衰减参数,$\gamma\lambda$为衰减率,通过其衰减的次数实现短期时间内的状态动作的“记忆功能”,加快算法的收敛速度.

基于人机交互的$Q(\lambda)$学习过程如图2所示.

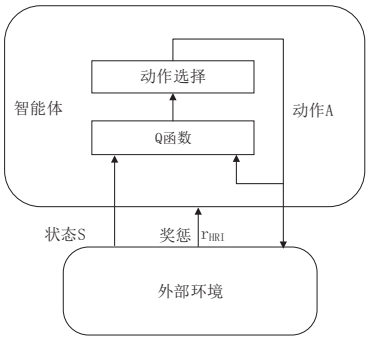


图2　基于人机交互的$Q(\lambda)$学习过程

3.2　加入意图推理的强化学习方法

通过以上的基本强化学习算法,可以得到机器人状态空间到其动作空间的映射,即可以得到机器人在某一时刻到达目标状态的最优动作序列. 在人与仿人机器人交互的过程中,机器人移动的速度较慢,存在一定的时滞,不能立即完成规划好的最优动作;另一方面,人总是存在一定的运动速度,因此基本的强化学习所学到的动作策略很可能在实际交互时产生较大的误差.本文在人机交互过程中,使用灰色预测方法对人的运动意图进行建模, 提出了一种加入意图推理的强化学习方法.

在实际的跟随任务中,考虑到机器人行走的速度,令机器人的控制周期为t,假设在t时刻发出指令动作. 新陈代谢GM(1,1)模型利用在t时刻之前Δt时间内人所处的位置状态序列$s_1, s_2, ..., s_t$的变化来预测下一时刻人的位置状态,表征人的行为序列变化,并通过对其进行建模, 得到人的行走意图,包括方向和速度的信息.

算法步骤如下：

首先,对位置状态序列s_1,s_2,...,s_t做平滑处理,消除噪声的影响,并且减小粒子滤波算法的误差. 然后,基于灰色预测具有不需要大量样本,样本不需要有规律的分布,计算量小,准确度高等特点,建立新陈代谢GM（1,1）模型对平滑后的数据做处理, 得到包含$t+1$时刻预测值的位置序列.对序列中的数据做处理,得到t时刻位置状态的平均变化率,通过比较后分类,得到人的行走意图,即快速左移,快速右移,慢速左移,慢速右移,静止不动.

同时创建了一个行为对策规则表A_I来实现人的行走意图与机器人所应采取的修正动作序列的映射,如表1所示.

Algorithm 1 The Intention Inference Algorithm

for each $i \in [1,t]$ **do**
　Smooth the data $s_1, s_2, ..., s_t$;
end for
Use the GM(1,1) model to predict the data $\hat{s}_1, \hat{s}_2, ...\hat{s}_{t+1}$
for each $i \in [1,t+1]$ **do**
　$k_i = \hat{s}_i - \hat{s}_{i-1}$;
　$kaverage = \frac{\sum_{i=t}^{t+1} k_i}{(t+1)-t+1}$;
　Determine the intention by comparing $kaverage$ with k_1 and $k_2(0 < k_1 < k_2)$;
　if $kaverage \geq k_2$ **then**
　　Intention is fast and left;
　else if $k_1 \leq kaverage < k_2$ **then**
　　Intention is slow and left;
　else if $kaverage \leq -k_2$ **then**
　　Intention is fast and right;
　else if $-k_2 < kaverage \leq -k_1$ **then**
　　Intention is slow and right;
　else
　　Intention is static;
　end if
end for

表 1: 行为对策规则表A_I

属性	左移	右移	不动
快速	x_1	x_3	x_5
慢速	x_2	x_4	x_5

然后通过查表法获得t时刻基于人的意图推理所得到的动作a_I,将其与强化学习得到的动作a_Q进行加权修正即可得到最终的优化动作.

加入意图推理的强化学习算法如算法2所示.

4　系统实现与实验

本文研究的机器人对特定目标的跟踪算法可以通过下面的实验来验证其性能.设计的跟随任务如下,机器人通过与环境交互学习, 自主地选取相应的动作,使得目标物体始终位于机器人视野中央,从而完成目标的跟随.其中,为了加快强化学习算法的收敛速度,简化学习任务为机器人的横向移动,纵向的前后移动采用双目视觉系统得到的距离信息来进行控制.

4.1　实验平台

实验硬件平台如图3所示:由NAO机器人、USB摄像头和计算机控制系统组成.Nao机器人全身有26个自由度,另外使用两个USB摄像头安装在NAO机器人的头上,构成双目视觉, 通过标定程序得到目标相对于机器人的深度信息.PC机处理器为Pentium双核,2.60GHz.摄像头为罗技S5500,帧速为30fps.软件平台为Windows7,VS2010,采用Visual C++实现.

整个实验系统的工作由两个并行的工作过程构成：一是计算机负责视频采集、图像处理以及控制量

Algorithm 2 The Intention Inference Based Reinforcement Learning Algorithm

Initialization:

Set parameter γ,α,k,and the reward r;
Initialize matrix Q as zero matrix;

Iteration:

for each episode **do**
  Determine initial state s_t;
  Save state s_1 to s_t during a short period of time Δt;
  while not reach good state **do**
    Select $a_Q = argmaxQ(s_t, a)$ according to the $Q(\lambda)$ algorithm;
    Select a_I from the behavioral strategies rules table A_I according to the intention inference algorithm during time Δt;
    $a_{improve} = a_Q + k * a_I$;
    $s_t \leftarrow s_t'$;
  end while
end for

图3　系统硬件平台

的计算,二是NAO机器人负责的运动控制.整个系统的工作流程如图4所示.首先,计算机上的程序从USB摄像头中提取到当前的图像帧,然后对图像进行目标特征识别和定位,得到目标在图像平面中的位置坐标,然后将其用于加入意图推理的强化学习算法中,计算得到控制量并发送给NAO机器人.NAO机器人按照得到的控制量进行运动规划,控制各关节完成相应的运动.当计算机把控制量发送给机器人后,则进入下一个处理周期,采集新的信息,如此周而复始.

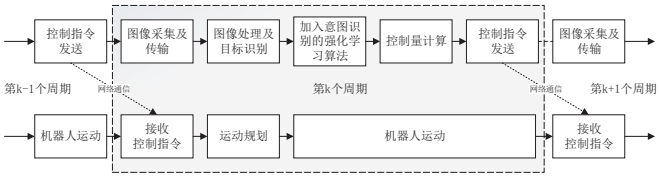


图4　系统工作流程图

4.2　环境状态的划分

文中使用的基本强化学习算法是建立在离散马尔可夫过程模型之上的,因此利用该强化学习理论来解决连续状态下的问题,方法之一是将连续空间离散化,得到有限个环境状态.根据目标物体在机器人视野中的位置,将图像平面划分为7个区域,如图5所示.目标状态为机器人视野的中央区域.

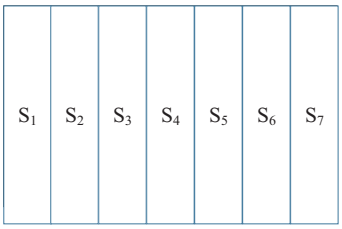


图5　状态划分

4.3　机器人动作行为的定义

通过研究NAO机器人的特点,选取机器人的动作集合如下:

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}. \tag{6}$$

a_1–定义为方向左移且步长Step=STEP(STEP为目标物体从一个区域到相邻区域的机器人移动步长);
a_2–定义为方向左移且步长Step=2STEP;
a_3–定义为方向左移且步长Step=3STEP;
a_4–定义为方向右移且步长Step=STEP;
a_5–定义为方向右移且步长Step=2STEP;
a_6–定义为方向右移且步长Step=3STEP;
a_7–定义为原地不动;

4.4　奖赏回报函数的设定

奖赏回报R是对强化学习中智能体在某一状态下采取某一动作的即时回报,它来自于环境的立即反馈,是智能体在学习的过程中获得的唯一可用于策略学习的有用信息.本文中通过人机交互对奖赏回报进行设定,对于接近目标状态的行为,给予较大的奖励,反之给予较小的奖励,系统能够通过奖励获得当前状态与目标状态之间的距离,有效地结合人的先验知识,减少搜索范围,降低学习的复杂度,从而提高任务学习的快速性.

4.5　实验结果与分析

基于粒子滤波算法的跟踪效果如图6所示,图像分辨率为320×240,粒子数为50,处理时间大约为每帧0.05s.其中红色的矩形框是粒子滤波算法中权重最大的粒子所在的位置,绿色的矩形框是所有粒子的加权期望所在的位置,即为最终确定的目标位置.目标人在室内复杂环境中行走,实验证明粒子滤波的跟踪效果较好,满足实际应用要求.

机器人得到目标位于其视野平面中的位置状态后,通过学习算法,采取相应的动作,使得目标位于其视野平面中央区域,从而完成跟随任务.

图7是利用新陈代谢GM(1,1)模型对Δt时间内人所处的位置状态序列$s_1, s_2, ..., s_t$(t取10)进行预测的结

图6　粒子滤波的跟踪效果

果.(a)图中在Δt时间内人的行走方向没有变化,(b)图中在Δt时间内人的行走方向改变.其中红色表示测量值,黑色表示预测值.并且根据前面所述算法对人的行走意图进行建模,得到所需信息.该方法能在已知数据较少的情况下达到较高的预测精度,对人的意图进行较好的判断.

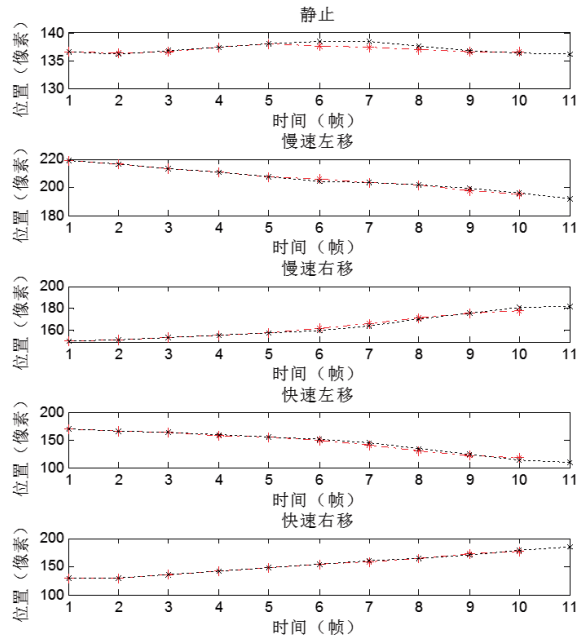
图8是采用基本的强化学习方法的实验效果图,黑线表示目标物体的质心在机器人视野中的横向位置,两条蓝线之间的区域为机器人视野的中央区域,即目标状态.实验中人的行走速度非常缓慢,在某一具体的时刻可以认为人的状态是静止的.通过图中曲线可以看到,人行走速度较慢时,基本的强化学习方法可以完成简单的跟踪任务.

图9是行走速度加快时采用基本强化学习方法的实验效果图,强化学习得到的动作序列没有考虑人的动态信息,使得机器人跟随人行走时具有一定的时滞.

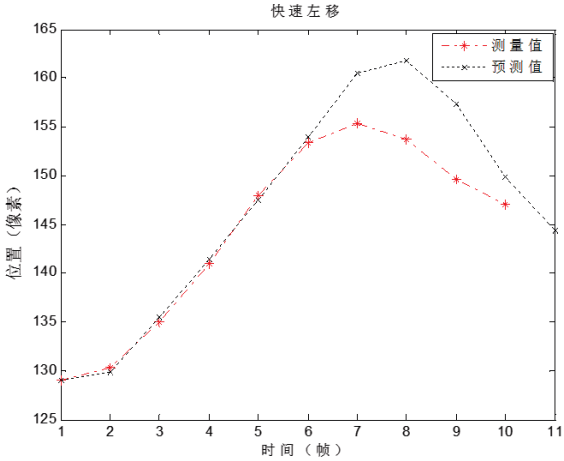
图10是加入意图推理的强化学习方法在人快速行走时的实验效果,可以看到机器人的跟踪效果优于基本的学习算法, 受人的速度的影响较小,能较好地完成跟随任务,基本满足实际应用的需要,同时该算法使得机器人的运动更加平滑和连贯.

5　结论

在人机交互的过程中,机器人能够在复杂的环境中跟随人行走,并完成更高级的交互任务是机器人学研究领域的一个重要问题. 本文提出了一种新的控制框架,使用粒子滤波算法实时地跟踪目标物体,得到其在机器人视野中的位置状态,并采用强化学习算法建立起状态变量与机器人动作的映射,使得机器人能够通过与环境的交互,自主地学习应采取的动作,达到目标状态从而完成跟随任务.同时,考虑到人在行走的过程中存在一定的速度、方向和惯性, 通过加入人的意图推理,修正强化学习得到的动作序列,减小机器人跟踪的时滞.实验验证了该方法的有效性.



(a)



(b)

图7　新陈代谢GM(1,1)模型预测结果.(a)表示Δt时间内行走方向不变.(b)表示Δt时间内行走方向改变.

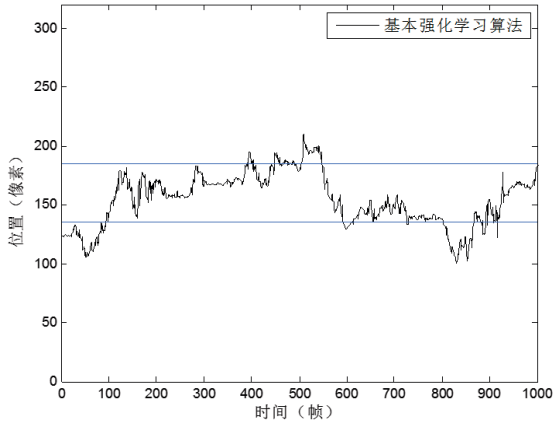


图8　基本强化学习算法跟踪结果

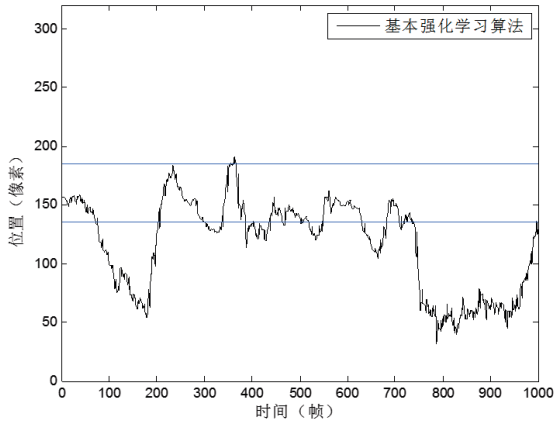


图9　基本强化学习算法跟踪结果

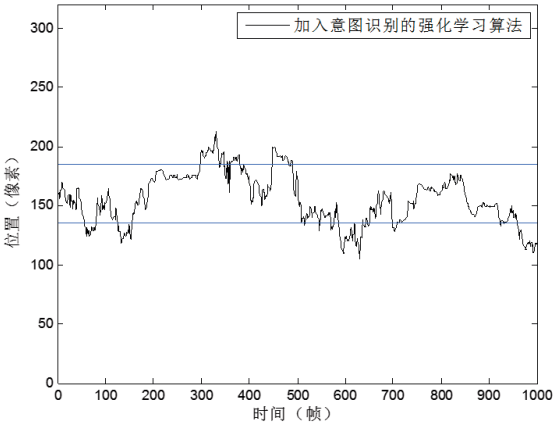


图10　加入意图推理的强化学习算法跟踪结果

参考文献

[1] J. Satake, J. Miura, Robust stereo-based person detection and tracking for a person following robot, in *Proceedings of the IEEE ICRA Workshop on Person Detection and Tracking*, Kobe, Japan, 2009.

[2] M. Tarokh, P. Ferrari, Robotic person following using fuzzy control and image segmentation, in *Journal of Robotic Systems*, 20(9): 557–568, 2003.

[3] T. Yoshimi, M. Nishiyama, T. Sonoura, et al, Development of a person following robot with vision based target detection, in *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, China, 2006: 5286–5291.

[4] Z. Chen, S.T. Birchfield, Person following with a mobile robot using binocular feature-based tracking, in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Diego, USA, 2007: 815–820.

[5] M. Kristou, A. Ohya, S. Yuta, Target person identification and following based on omnidirectional camera and LRF data fusion, in *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication*, Atlanta, USA, 2011: 419–424.

[6] M. Kleinehagenbrock, S. Lang, J. Fritsch, et al, Person tracking with a mobile robot based on multi-modal anchoring, in *Proceedings of the 11th IEEE International Workshop on Robot and Human Interactive Communication*, Berlin, Germany, 2002: 423-429.

[7] K. Qian, X. Ma, X. Dai, Simultaneous Robot Localization and person tracking using rao-blackwellised particle filters with multi-modal sensors , in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nice, France, 2008: 3452-3457.

[8] 戴景文, 刘丹, 苏剑波. 基于投影峰的眼睛快速定位方法, *模式识别与人工智能*, 22(4): 605–609, 2009.

[9] E. D’ Arca, N. M. Robertson, J. Hopgood, Person tracking via audio and video fusion, in *Data Fusion & Target Tracking Conference*, London, UK, 2012: 1–6.

[10] N. Bellotto, H. Hu, Vision and laser data fusion for tracking people with a mobile robot, in *Proceedings of the 2006 IEEE International Conference on Robotics and Biomimetics*, Kunming, China, 2006: 7–12.

[11] T. Wilhelm, H.-J. Boehme, H.-M. Gross, Sensor fusion for vision and sonar based people tracking on a mobile service robot. in *Proceedings of the International Workshop on Dynamic Perception*, Bochum, Germany, 2002: 315–320.

[12] W. Wang, J. Chang, Implementation of a mobile robot for people following, in *International Conference on System Science and Engineering*, Dalian, China, 2012: 112–116.

[13] C. Hu, X. Ma, X. Dai, K. Qian, Reliable people tracking approach for mobile robot in indoor environments, *Robotics and Computer-Integrated Manufacturing*,26: 174–179, 2010.

[14] H. Kwon, Y. Yoon, et al, Person tracking with a mobile robot using two uncalibrated independently moving cameras, in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*,Barcelona, Spain, 2005: 2877–2883.

[15] 张汝波, 顾国昌, 刘照德,王醒策. 强化学习理论、算法及应用, *控制理论与应用*, 17(5): 637–642, 2000.

[16] 高阳, 陈世福, 陆鑫. 强化学习研究综述, *自动化学报*, 30(1): 86–100, 2004.

[17] Y. Takahashi, M. Takeda, M. Asada, Continuous valued Q-learning for vision-guided behavior acquisition, *Proceedings of the 1999 IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*,Taipei, Taiwan, 1999: 255–260.

[18] 陈睿, 苏剑波. 非合作型人机交互中用户意图推理, *第32届中国控制会议论文集*,2012: 5960–5964.

[19] 邓聚龙. *灰色系统基本方法*. 武汉: 华中理工大学出版社, 1987.

[20] 王立琦, 陈海云, 燕小强. 一种改进的粒子滤波视频跟踪算法. *第13届中国系统仿真技术及其应用学术年会论文集*, 2011: 984–987.