005-环境感知-Learning How Pedestrians Navigate:A Deep Inverse Reinforcement Learning Approach

序号：005
名称：Learning How Pedestrians Navigate:A Deep Inverse Reinforcement Learning Approach
学习行人如何导航：一种深度逆强化学习的方法
作者：Muhammad Fahad, Zhuo Chen, and Yi Guo
文献类型：会议IROS
年份：2018
关键词：增强学习
整理日期：2019年5月9日
论文链接：005-Learning How Pedestrians Navigate- A Deep Inverse Reinforcement Learning Approach.pdf

主要解决问题

机器人学习人类的导航方式。作者提出了深度逆强化学习（IRL）的方法来对行人的导航方式进行建模。

解决思路

作者提出了一种使用 最大熵深度逆强化学习（**maximum entropy deep inverse reinforcement learning [MEDIRL])** 来学习人类导航行为的方法。作者使用了一个在不受控的环境下收集的大型公开的行人轨迹数据集来作为专家演示。人类的导航行为由一个通过深度神经网络（DNN）估计的非线性奖励函数（reward function）来获取。已经开发好的MEDIRL算法用从人类运动轨迹中提取出的包括社会亲和图（social affinity map [SAM]）在内的特征作为输入。
作者使用学习到的奖励函数进行了模拟实验，并通过与数据集中真实测量的行人轨迹对比来评估其性能。评估结果显示作者提出的方法在与其他最先进的方法对比中有着可以接受的预测精度，并且可以生成与人类轨迹相似，有着例如防碰撞、领导与跟随、分来与汇合等自然社会导航行为的行人轨迹。

核心知识点

1、增强学习
https://www.jianshu.com/p/7a9f9225e2b2

- 增强学习就是将情况映射为行为，也就是去最大化收益。学习者并不是被告知哪种行为将要执行，而是通过尝试学习到最大增益的行为并付诸行动。也就是说增强学习关注的是智能体如何在环境中采取一系列行为，从而获得最大的累积回报。通过增强学习，一个智能体应该知道在什么状态下应该采取什么行为。RL是从环境状态到动作的映射的学习，我们把这个映射称为策略。
- 增强学习的构成要素
    - **policy**定义为在给定时刻t下，智能体表现的方式。具体来说，policy就是从环境状态到动作行为的一个映射。policy是增强学习的核心。
    - **reward function**定义为在增强学习问题中的目标。具体来说，reward就是从环境状态到奖励的一个映射。智能体的任务就是在长期的过程中，不断最大化所得的总奖励。注意，这里的reward

function指的是单次行为所造成环境状态改变所带来的奖励，所以为即时奖励。这个reward function定义了对智能体来说，什么是对的行为，什么是错误的行为。

- **value function**定义为在长期奖励。也就是说是reward function的累加。如果说reward function告诉智能体什么行为是好的坏的，那么value function告诉智能体的是长期以来你的行为是好的还是坏的。
- **a model of environment**定义为对环境行为的模拟。比如说，给定state和action，我们要估计下一个时刻的state和reward，这时候就需要model of environment作出预测。

2、马尔科夫决策过程

有限状态MDP可以定义为(S,A,T,γ,R)，其中状态空间$S=\{s_1,\ldots,s_O\}$，行为空间（action space）$A=\{a_1,\ldots,a_L\}$，状态变换概率T（即描述执行a_{i-1}时，从状态s_{i-1}转变到a_i的变化情况），折扣因子discount factor$\gamma \in [0,1)$，奖励函数R（取决于状态和行为）。

决策过程的核心是要找到policy π（可以看成是从S→A的一个映射），找到每一个状态对应的最优行为来最大化累计回报函数。

增强学习使用给定的回报函数来生成最优的从状态到行为的policy映射；IRL是其可逆问题，旨在从现有的policy、一个行为-状态序列或显示的状态序列中找到回报函数

3、逆强化学习

逆强化学习就是学习得到回报函数。

逆向强化学习的提出者Ng是这么想的：专家在完成某项任务时，其决策往往是最优的或接近最优的，那么可以这样假设，当所有的策略所产生的累积回报期望都不比专家策略所产生的累积回报期望大时，强化学习所对应的回报函数就是根据示例学到的回报函数。

逆向强化学习一般流程如下：

1. 随机生成一个策略作为初始策略；
2. 通过比较“高手”的交互样本和自己交互样本的差别，学习得到回报函数；
3. 利用回报函数进行强化学习，提高自己策略水平；
4. 如果两个策略差别不大，就可以停止学习了，否则回到步骤2。

逆向强化学习分类如下：

- 最大边际形式化：学徒学习、MMP方法、结构化分类、神经逆向强化学习。
- 基于概率模型的形式化：最大熵IRL、相对熵IRL、深度逆向强化学习。

4、行人运动建模

假设行人在离散化的格点中运动，初始位置为s1，可能的行为集合为$\{a_0,a_1,a_2,\ldots,a_8\}$，并对应着可能的相邻状态$\{ns_1,ns_2,\ldots,ns_8\}$，故行人从s1到sK的轨迹可以描述为$\zeta=\{(s_1,a_1),(s_2,a_2),\ldots,(s_K,a_K)\}$。

假设行人的行为符合马尔科夫决策过程，且每一次从s到s’的决策概率总和为1，回报函数为R。

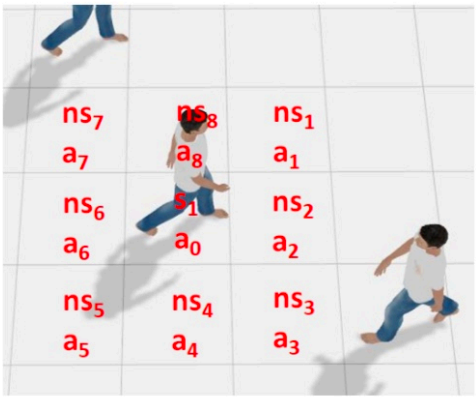


Fig. 1: Pedestrian motion model in a grid based workspace with adjoining states $ns_1, ns_2, ..., ns_8$ and corresponding actions $a_1, a_2, ..., a_8$ to reach those states.

程序功能分块说明

0、问题描述

专家策略 π_D 给出的描述为 $D = \{\zeta_1, \zeta_2, \ldots, \zeta_N\}$，每一次描述对应的特征因子为φ。

目标： 找到未知的回报函数R*能够接近D的结果。

1、最大熵深度逆强化学习MEDIRL

假设回报函数R*是关于特征因子 $\phi = \{\phi_1, \phi_2, \ldots, \phi_n\}$ 的非线性函数，并可以用深度神经网络计算：

$$R^* = g(\phi, \theta_1, \theta_2, \theta_3, \ldots, \theta_j), \qquad (1)$$

$$= g_1(g_2(\ldots(g_j(\phi, \theta_j), \ldots), \theta_2), \theta_1), \qquad (2)$$

其中， $\theta=[\theta_1,\theta_2,\ldots,\theta_j]$ 是深度神经网络的权重，$g_j()$是非线性函数。

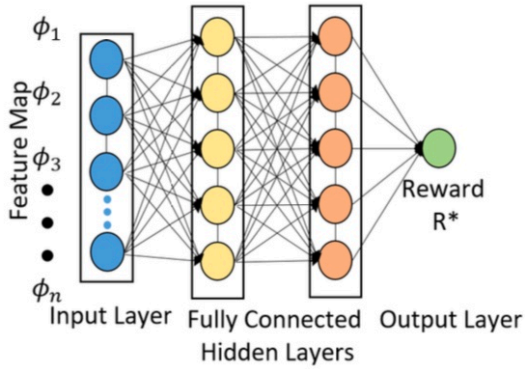


Fig. 2: DNN for reward function approximation based on the feature space represented by $\phi = [\phi_1, \phi_2, \phi_3, ..., \phi_n]$.

最大化D和θ的联合概率：

$$\mathcal{L}(\theta) = \log P(D, \theta | R^*) = \underbrace{\log P(D|R^*)}_{\mathcal{L}_D} + \underbrace{\log P(\theta)}_{\mathcal{L}_\theta}, \qquad (3)$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_D}{\partial \theta} + \frac{\partial \mathcal{L}_\theta}{\partial \theta}.$$

$$\frac{\partial \mathcal{L}_D}{\partial \theta} = \frac{\partial \mathcal{L}_D}{\partial R^*} \cdot \frac{\partial R^*}{\partial \theta}, \quad = \underbrace{(\mu_D - \mathbb{E}[\mu])}_{\text{State Visitation Matching}} \cdot \underbrace{\frac{\partial}{\partial \theta} g(\phi, \theta)}_{\text{Back-propagation}}, \qquad (6)$$

其中$R^* = g(\phi, \theta)$，μD为专家策略下状态访问频率，E[μ]为访问次数。

Algorithm 1 Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL)

Input μ_D^a,ϕ,S,A,T,γ
Output Optimal weights θ^*

θ = initialize_weights()
for m = 1:M **do**
  R^{*m} = g(ϕ,θ^m)
  MDP solution with current reward function
  π^m = approx_value_iteration(R^m,S,A,T,γ)
  $\mathbb{E}[\mu^m]$ = propagate_policy(π^m,S,A,T)
  Maximum entropy loss calculation and gradients
  $\mathcal{L}_D^m$=log(π^m)×μ_D^a
  $\frac{\partial \mathcal{L}_D}{\partial R^{*m}}$ = μ_D-$\mathbb{E}[\mu^m]$
  Compute network gradients
  $\frac{\partial \mathcal{L}_D^m}{\partial \theta_D^m}$ = nn_backprop($\phi, \theta^m, \frac{\partial \mathcal{L}_D}{\partial R^{*m}}$)
  θ^{m+1} = update_weights($\theta^m, \frac{\partial \mathcal{L}_D^m}{\partial \theta_D^m}$)
end for

known. The state transition probabilities denoted by T, are assumed to be one for our case as we consider the pedestrian motion agent to be a deterministic system. The discount

2、特征向量

特征向量可以从行人i的空间位置s_i=[x,y]及周围人的空间位置和速度中提取。所提取的特征主要为：

- Social Affinity Map Feature（SAM）
  social affinity可以被定义为周围人间的motion affinity。

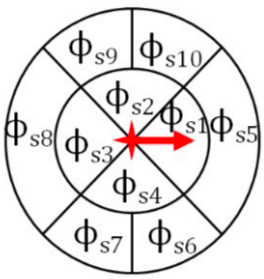


Fig. 3: The SAM feature represented by spatial location bin disks around the pedestrian of interest (i.e., the center). The pedestrian location is shown by the red cross and its motion direction is shown by the red arrow.

以行人为中心，半径分别为r和r'划分两个同心圆，内部的圆四等分，两圆之间的圆环先四等分，然后对行人左右两侧的圆环部分再进行二等分。 行人的移动方向用红箭头表示。 这含有10个元素（对应10个子区域）

的特征向量可以表示为 $\phi_S = [\phi_{S1}, \phi_{S2}, \ldots, \phi_{S10}]$，如果有人在则对应的元素为1否则为0。

每一个子区域的特征向量可以表示为 $\phi_{SV} = [\phi_{o1}, \phi_{o2}, \phi_{o3}, \phi_{v1}, \phi_{v2}, \phi_{v3}]$，我们分别计算每一个子区域中所有人的平均速度（速率l和方向α），并根据阈值表得到0或1值。

TABLE I: SAM: Velocity feature thresholds.

Feature	Thresholds
ϕ_{o1}	$\alpha \in (-\frac{3\pi}{4}, \frac{3\pi}{4}]$
ϕ_{o2}	$\alpha \in [\frac{\pi}{4}, \frac{3\pi}{4}) \bigcup [-\frac{3\pi}{4}, -\frac{\pi}{4})$
ϕ_{o3}	$\alpha \in (-\frac{\pi}{4}, \frac{\pi}{4}]$
ϕ_{v1}	$l \in [0,0.5)$ m/s
ϕ_{v2}	$l \in [0.5,1.0)$ m/s
ϕ_{v3}	$l \in [1.0,\infty)$ m/s

包含所有子区域的完整特征向量为 $\phi_{SC} = [\phi_{SV1}, \phi_{SV2}, \ldots, \phi_{SV10}]$。

并得到包含φ_S和φ_{SC}的完整SAM特征向量：$\varphi_{SAM}=[\varphi_S,\varphi_{SC}]$

- Density Feature

  $\phi_d = [\phi_{d1}, \phi_{d2}, \phi_{d3}]$.

  半径r内：①少于等于2人，：$\varphi_{d1}=1,\varphi_{d2}=0,\varphi_{d3}=0$;

  ②大于2人小于5人：$\varphi_{d1}=0$，$\varphi_{d2}=1$，$\varphi_{d3}=0$;

  ③大于等于5人：$\varphi_{d1}=0$，$\varphi_{d2}=0,\varphi_{d3}=1$。
- Distance Feature

  φ_{dis}衡量行人当前位置到目标位置的距离
- Default Cost Features

  φ_{def}用于平衡上述三个特征因子，置1。

完整的特征因子为：

$$\phi = [\phi_d, \phi_{SAM}, \phi_{dis}, \phi_{def}].$$

具体的实验操作：

使用专家数据集（http://www.irc.atr.jp/crest2010 HRI/ATC dataset/）训练得到回报函数R*，再进行轨迹的预测。

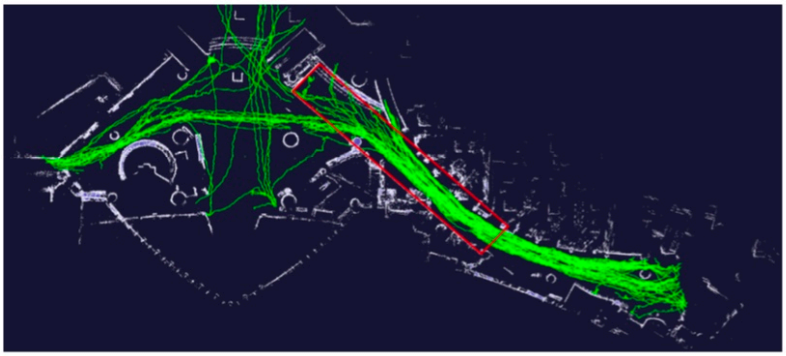


Fig. 4: The map of the mall with pedestrian trajectory data shown in green. Static obstacles are shown in white color. The red box denotes the area from which pedestrian trajectories were selected for learning R^*.

C. Evaluation Experiments

Using the learned reward function R^* from the previous step, feature vectors, possibles states and actions, and discount factor, the pedestrian navigation problem is solved as an MDP sequential decision making process using approximate value iteration algorithm detailed in Algorithm 2. Here V_s is the optimal value for each state and Q is the Q-value.

In the next section, we present performance evaluation results of the proposed method.

Algorithm 2 Trajectory Evaluation

Input R^*,θ,S,A,T,ϕ,γ
Output Motion policy π

1: Load DNN parameters θ, that were trained by Algorithm 1
2: V_s=-∞
3: **repeat**
4: V_{t_s}=V_s;$V_{s_{goal}}$=0
5: $Q_{s,a} = R^*_{s,a}+\gamma \sum_{s' \in S} T_{s,a,s'} V_{s'}$
6: $V_s = softmax_a Q_{s,a}$
7: **until** $max_s(V_s - V_{t_s})<\epsilon$
8: $\pi^*(a|s) = e^{Q_{s,a}-V_s}$

measured即为实际轨迹，**simulated**即为预测轨迹

- 片段轨迹预测
  图中红色区域，正确率96.6%
- 整体轨迹预测

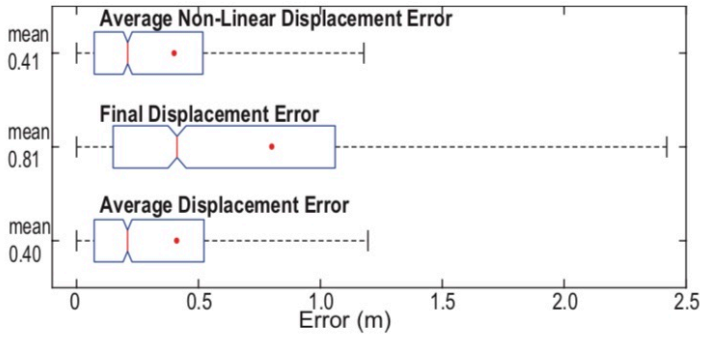


Fig. 5: Box plot of the errors between the simulated and measured pedestrian trajectories. The mean error value for each metric is shown by a red circle and the mean value is shown in the label along y-axis.

- 个人行为轨迹预测

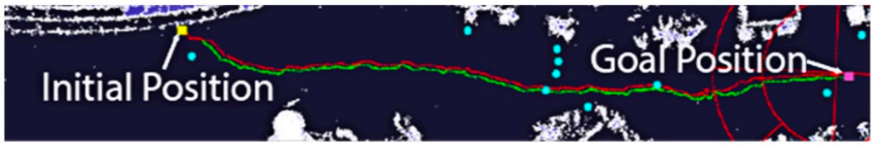


Fig. 6: The measured and simulated trajectory of the pedestrian C shown in green and red respectively. The initial position, goal position and SAM feature bins are also shown. Neighboring pedestrians are shown by turquoise disks.

（白色为静止障碍物，红色圈反映SAM特性）

- 跟随行为

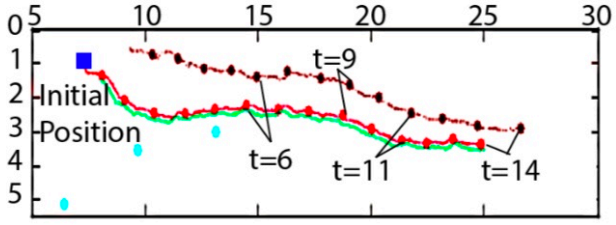


Fig. 7: Visualization of the leader-follower behavior that the pedestrian exhibits. The leader's measured trajectory, the follower's measured and simulated trajectories are shown in brown, green and red, respectively. The figure shows the *simulated* follower trajectory (in red) matches the *measured* follower trajectory (in green), thus validates our proposed algorithm.

- 避障行为

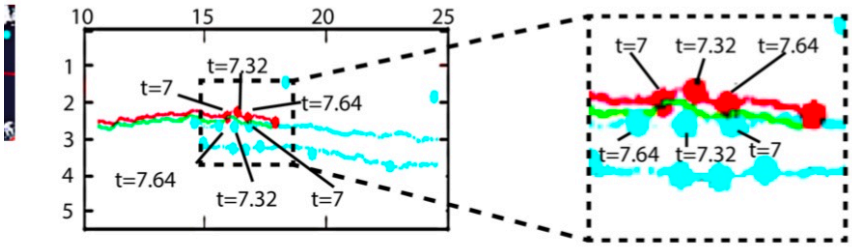
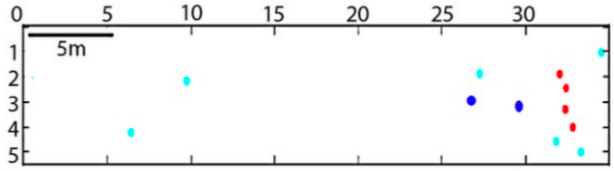
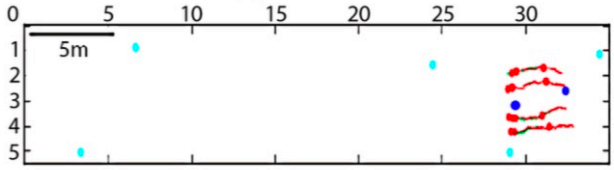


Fig. 8: Collision avoidance behavior of the measured (in green) and simulated (in red) pedestrian trajectories. The pedestrians execute maneuvers from 7 secs to 7.64 secs to avoid collision.

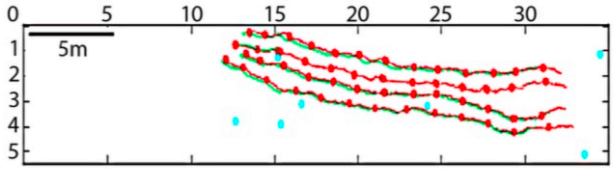
- 群体行为预测（split-and-rejoin）



(a) t = 0 sec



(b) t = 2.12 sec



(c) t = 13 sec

Fig. 9: Group behavior (Split-and-rejoin) shown by a group of pedestrians. Two pedestrians passing between the group are shown in blue in Fig. 9a to Fig. 9b. The group is then shown to keep formation till 13 secs in Fig. 9c.

存在的问题

1、改进为端到端的IRL
2、考虑到行人和静止障碍物之间的交互
3、在实际情况下验证方法

改进的思路