Research
Robotics—Article

Engineering 2015, 1(1): 79–84
DOI 10.15302/J-ENG-2015024

通过行为足迹学习人类习惯的个性化服务机器人

李坤 [1], Max Q.-H. Meng[2*]

摘要：对家用的私人机器人来说，个性化服务和预先设计的任务同样重要，因为机器人需要根据操作者的习惯调整住宅状况。为了学习由诱因、行为和回报构成的操作者习惯，本文介绍了行为足迹，以描述操作者在家中的行为，并运用逆向增强学习技巧提取用回报函数代表的操作者习惯。本文用一个移动机器人调节室内温度，来实施这个方法，并把该方法和记录操作者所有诱因和行为的基准办法相比较。结果显示，提出的方法可以使机器人准确揭示操作者习惯，并相应地调节环境状况。

关键词：个性化机器人，习惯学习，行为足迹

1 前言

传统上，私人机器人被设计来提供不同情境下的标准服务。例如，通过联合房门识别和操纵算法，机器人可以用完全一样的方法打开不同房屋的各种门。结合操作者的命令，这个策略使机器人始终能在不同环境完成每个任务。当机器人用于固定、重复情境时，这些表现是令人满意的，但这个策略不足以满足操作者个性化服务要求。

在智能化家庭中，个性化要求尤为明显，这里的机器人需要智能地监测并调节房屋状况。例如，机器人可能需要把门开到不同的程度，一些操作员想要全开，另一些人也许更喜欢半开。如果是离线设计，这种状态调整需要大量的手工作业。为解决这个问题，机器人必须通过学习操作者习惯进行个性化服务，以便根据每个操作者习惯进行自身调整。

为了学习一个习惯，机器人需要观察环境，并提取与习惯相关的信息。习惯由 3 个因素决定：诱因、行为和回报 [1]。在充分经历过 3 个因素后，操作者在看到诱因后会不由自主地发生行为，而不是刻意表演，以收集最大回报。机器人要理解操作者的习惯，就会从观察中收集所有诱因和行为，以便引导今后的行动，或根据观察学习决定今后行动的回报。第一种方法很简单，因为机器人在面对诱因时可以重复记忆，找到最佳匹配行为，但不足以应对新出现的诱因；第二种方法需要一个额外的学习过程，这样习得的回报会引导机器人在新诱因出现时行动。本文中，第一种方法作为基准办法实施，重点探讨第二种方法。

根据观察，笔者以逆向增强学习为框架，提出了学习操作者习惯的方法，同时用行为引发的环境状况变化描述行为，即行为足迹。同时，机器人根据操作者和物体间的接触观察诱因，在屋内学习基于操作者行为的习惯作为回报函数。然后，使用回报函数引导其日后行动，以便自主地服务操作者。用自主调节室内温度的案例研究实现该方法。本文的贡献包括结合代表操作者行为的行为足迹和提出根据操作者习惯使机器人个性化。

2 相关研究

私人机器人的传统研究重点是设计普遍适用的硬件和软件。例如，文献 [2] 开发了能打开门并自己充电的机器人；文献 [3] 开发了带视觉传感器的感知系统，以引导机器人在不同环境下的运动；文献 [4] 使用多个传感器建

[1]California Institute of Technology, Pasadena, CA 91125, USA; 2 The Chinese University of Hong Kong, Hong Kong, China
* Correspondence author. E-mail: max@ee.cuhk.edu.hk
Received 26 March 2015; received in revised form 29 March 2015; accepted 30 March 2015

引用本文: Kun Li and Max Q.-H. Meng. Personalizing a Service Robot by Learning Human Habits from Behavioral Footprints. *Engineering*, DOI 10.15302/J-ENG-2015024

立人类和机器人互动框架；文献 [5] 开发安全、实用的私人机器人；文献 [6] 开发了在家庭环境中服务于小孩和成人的私人漫游机器人。

很多文献都探讨过家用私人机器人。例如，文献 [7] 使用脑电图信号控制远距机器人，并协助靠电机运动的残疾人；文献 [8] 设计远距机器人帮助老年人进行人际交流；文献 [9] 设计远程医疗系统监测老年人的健康和行动；为了使家庭监护包含机器人动作，文献 [10] 中的服务机器人使用传感器网络和射频识别引导其动作。

屋内安装不同种类的传感器，可以分层描述环境状况，其变化也可以用分层隐马尔可夫模型进行描述 [11]，多层隐马尔可夫模型堆叠，以描述分层状态转变，以及分层隐马尔可夫模型 [12]，每个状态的更高层合并低一层的隐马尔可夫模型。

为使机器人服务个性化，机器人需要通过观察学习操作者的习惯。为结合机器人动作和环境状态模拟，提出了很多以逆向增强学习为框架的方法 [13]。此外，通过示范学习 [14] 使机器人能够模仿操作者，并学习不同行为。在我们的应用中，机器人可以观察操作者行为，并用逆向增强学习 [15] 为操作者习惯编码。

本文使用逆向增强学习让机器人学习操作者习惯的回报函数。学习期间，操作者的行为由行为足迹代表，收集一组这些行为观察后，机器人尝试学习操作者习惯。

3 方法

3.1 行为足迹

习惯由三部分决定：诱因、行为和回报。为了学习操作者习惯，机器人必须观察环境，以获取诱因并观察操作者行为。这样，机器人就可以学习回报函数来描述操作者习惯。为此，机器人需要精确地表示环境。本文使用室内物体描述住宅状态：

$$E = (C_1, ..., C_n)$$

式中，E 代表环境状态，$C_i\ (i = 1, ..., n)$ 代表环境中的第 i 个物体，如图 1 所示。

笔者使用行为足迹表示操作者的行为 A，行为足迹由操作者行动导致的物体状态变化来定义，因为这种表示可以更有意义地描述不同种类行为，那些不会改变环境状态的行为除外：

$$A = (E_i, E_j)$$

式中，E_i 和 E_j 分别代表由操作者行为引发的住宅状态转变，如图 2 所示。

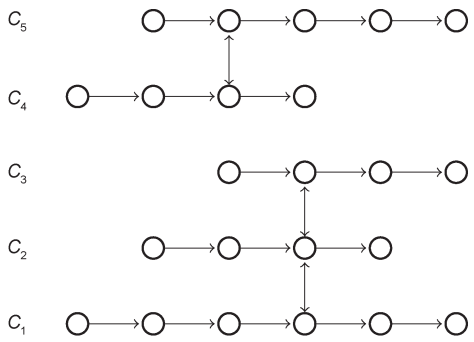

图 1. 环境状态由多个物体状态链组成，这些物体状态可能是异步且不相关的。

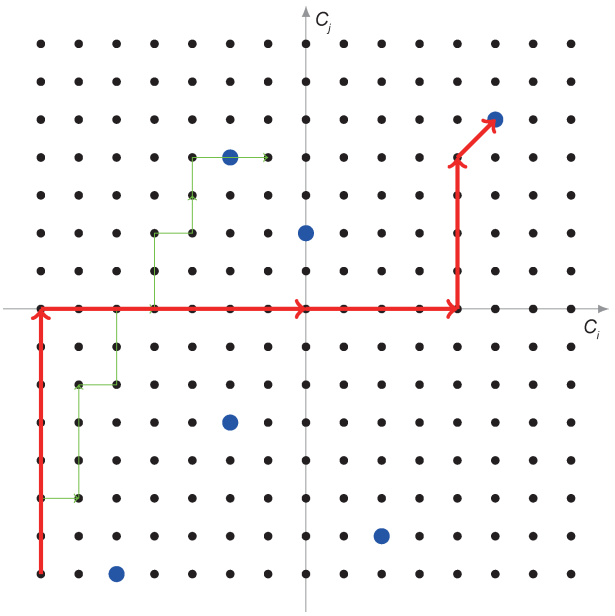

图 2. 住宅状态可由状态矢量空间中的点表示，它会自己变化，如绿色细线轨迹所示；但是，操作者可以手动改变它为任何想要的状态，由蓝色粗点表示，这些动作导致了由红色粗体轨迹表示的住宅状态变化。

3.2 诱因和行为

机器人可以用行为足迹观察操作者的行为，并触发这些行为的诱因。

行为是由操作者接触对物体状态的改变表示的。但是，有些行为是随机的，需要排除这些不符合操作者习惯的行为。为了评估操作者行为的规律性，本文使用以下测量：

$$r = r(A)$$

式中，r 测量引发行为 A 的诱因的标准偏差。根据测量的规律性和基于实验选择的界限值，机器人只保留规律行为的样本。

习惯的另一个重要因素是诱因，由行为发生时的环境状态所定义。这是在操作者出现之前，通过收集数据样本完成的。

$$S = [D_1(S_{t1}, ..., S_{tj}), ..., D_m(S_{ti}, ..., S_{tn})]$$

式中，D_i ($i = 1, ..., m$) 表示操作者出现的时刻，每个 (S_{t_1}, ..., S_{t_n}) 表示操作者出现后的一系列住宅状态。

诱因有两种，一种是令人愉悦的，操作者不改变环境状态；另一种是令人不悦的，这时操作者手动改变了一些物体状态。根据观察，这些样本被赋予了二元指标：

$$R = [R_1, ..., R_n]$$

3.3 回报

机器人通过操作者的常规行为样本和环境约定性的二元指标推断操作者的习惯。这个问题用公式表示为逆向增强学习，机器人通过观察操作者的动作学习回报函数 [16]：

$$\max \sum_{s \in S} \min_{\alpha \in \{A\}} \{p(E_{s' \in P_{s\alpha_1}}[V^{\pi}(s')] - E_{s' \in P_{s\alpha}}[V^{\pi}(s')])\}$$
$$\text{s.t.} |\alpha_i| \leqslant 1, i = 1,..., d \quad (1)$$

式（1）中，

$$V^{\pi}(s_0) = E_{\pi}(\sum_{t=0}^{\infty} \gamma^t R(s_t)) \quad (2)$$

表示某种规则下预期的折算回报，如图 3 所示。

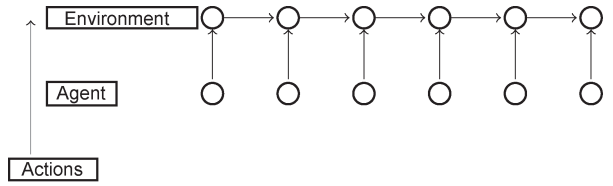

图 3. 逆向增强学习旨在展现基于最佳行动策略的回报函数。

最优化使操作者动作和学习操作者习惯的其他动作之间的差异最大化。

根据环境状态约定性的二元指数，式 (1) 的最大化简化为

$$\max \min \{E_{s' \in P_{s\alpha_1}}[V^{\pi}(s')] - E_{s' = P_{s\alpha_2}}[V^{\pi}(s')]\}$$
$$\text{s.t.} |\alpha_i| \leqslant 1, i = 1,..., d \quad (3)$$

式 (3) 中，α_1 表示适合操作者习惯的动作；α_2 表示不符合操作者习惯的动作。用二元指标测量约定性。

通过式 (3)，机器人学习回报函数，也是环境状态函数：

$$R_t = \phi(S_t)$$

回报函数的学习基于文献 [15] 的公式化，回报函数是一系列预设计基函数的线性组合：

$$R_t = \omega_1 \phi_1(S_t) + ... + \omega_n \phi_P(S_t) \quad (4)$$

式 (4) 中，ϕ_i 是基函数。

在个性化环境中，回报函数必须为因环境状态里物体的出现和消失而产生的潜在变化编码。有了行为足迹，利用不同尺寸及距离之间的相关性，这个问题可以通过把状态矢量空间尺寸聚集到多个抽象尺寸中来解决：

$$(cst_1, ..., cst_n) = partition\ (S, RLT)$$

聚集不仅排除了因物体状态相关性导致的冗余信息，也展现了无形的状态转换。而且，它还避免了物体序号变化时基函数的重构。这是因为只有和现存所有尺寸不相关的物体才需要重构基函数。此外，聚集可以使机器人用一个动作改变所有相关物体状态。

根据尺寸聚集，每个基函数记录一个团簇态的组合：

$$F_i = \phi_i(cst_1, ..., cst_n)$$

把基函数带入式 (4)，回报函数为

$$R(S_t) = \omega \cdot \phi(S_t) \quad (5)$$

式 (5) 中 $\omega = [\omega_1, ..., \omega_n]$，且 $\phi = [\phi_1, ..., \phi_p]$。

将式（5）代入式（2）：

$$V_{\pi}(S_0) = \omega \cdot E_{\pi}(\sum_{t=0}^{\infty} \gamma^t \phi(S_t))$$

通过式（2），式（3）简化为

$$\max \min \omega \cdot (\mu_1 - \mu_2)$$
$$\text{s.t.} |\omega_i| \leqslant 1, i = 1,..., d$$

式中，$\mu_i = E_{\pi_i}[\sum_{t=0}^{\infty} \gamma^t \phi(S_t)]$，描述第 i 个行动策略下的预期回报。

受到文献 [15] 研究的启发，笔者把最大化转化为类似于支持向量机的优化：

$$\max_{t, \omega} t$$
$$\text{s.t.} \omega \cdot \mu_1 \geqslant \omega \cdot \mu_2 + t$$
$$\|\omega\|_2 \leqslant 1$$

这个优化通过现有的支持向量机的实现得以解决 [18]。

3.4 机器人动作

通过学习指示操作者习惯的回报函数，机器人可以把它作为正常增强学习问题，引导其动作。

4 试验和结果

4.1 装置

本文使用 Turtlebot 作为个性化机器人，来观察由 4 个室外状态和 4 个室内状态组成的环境中人们的行为。4 个室外状态包括室外温度、湿度、风和雨；4 个室内状态包括温度计、门、空调开关和操作者状态。为了精确观察室内物体，在机器人操作系统中，用 Gmapping 工具

包 [17] 建立一个地图。收集 7 天的状态后，机器人尝试学习习惯，并用习惯引导日后动作。

笔者的机器人没有装配手臂，无法用身体改变物体状态，因此机器人动作是模拟的。

4.2 实验

4.2.1 习惯监测

机器人观察了 4 个从气候网站 (www.weather.com) 上摘录的气候条件，包括温度、湿度、雨和风。这些环境状态收集自香港 7 天的数据，如图 4 所示。

机器人观察了 4 个室内物体，包括温度计、门、空调开关和屋内操作者状态。这些物体的状态是机器人根据其视觉外观测量的，如图 5 所示。

4.2.2 习惯学习

根据观察，机器人收集操作者行为和引发操作者行为的诱因。当操作者接触物体时，收集诱因作为环境状态。例如，当操作者进入屋内，打开空调，此时的环境状态被收集为导致空调开关变化的诱因。

行为被收集为因操作者行为造成的环境状态变化，如开关空调、开门等。

收集 7 天的诱因和行为后，机器人用它们学习操作者习惯，并根据新的观察更新结果。该习惯用回报函数表示。

4.2.3 机器人动作

学习了回报函数，机器人寻找调节环境的最佳行动。本文手动应用这些发生的动作，以评估其效果。

4.3 结果

收集观察和学习操作者习惯 1 周后，机器人得出一组与逐渐增强的样本相符的回报函数。为了评估这些学习的回报函数，使用了两个指数，包括回报函数的准确度 r_A，通过比较机器人对住宅状态约定性评估和操作者提供的真实数值计算得到，以及由机器人和操作者行动不一致比例表示的机器人动作准确性 r_D。

利用环境中不同数量物体进行了两组试验，每组试验中，都实施基准办法和提出的方法，根据 r_A 和 r_D 评估。结果如图 6 和图 7 所示。

结果显示两种方法在评估住宅状态上精确度相似，

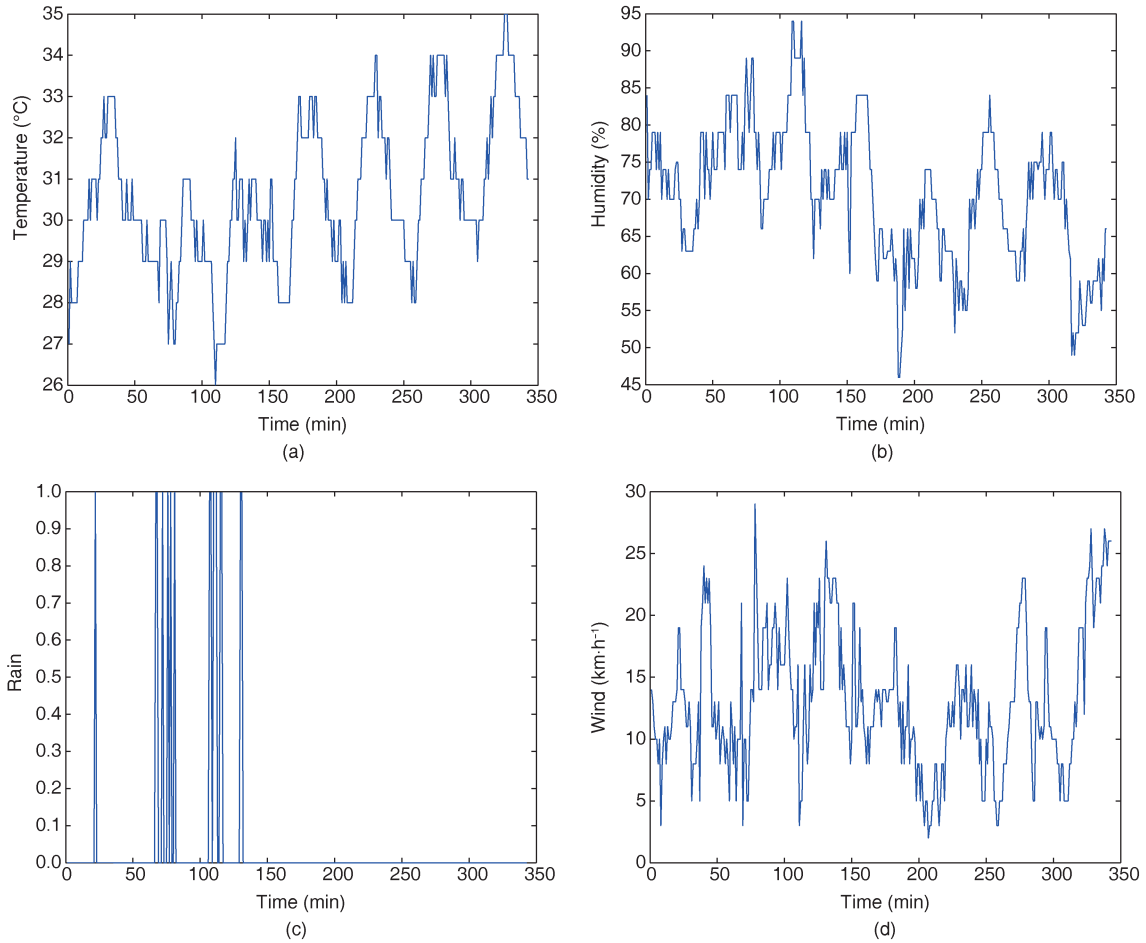

图 4. 天气状况，包括温度、风、湿度和雨。图示为从 7 月 25 日到 7 月 31 日收集的样本，每隔 30 min 收集一次。(a) 室外温度；(b) 室外湿度；(c) 室外雨量；(d) 室外风速。

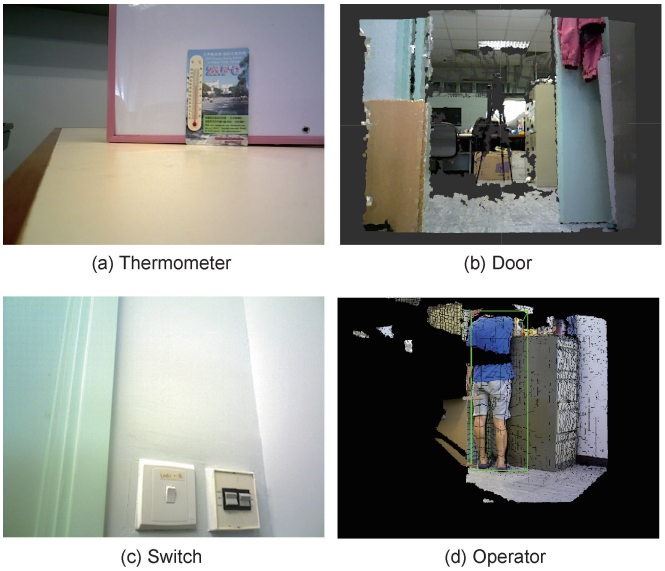

图 5. 从 4 个室内物体的外观检测其状态。机器人定期收集这些物体的状态，以监测住宅状态。

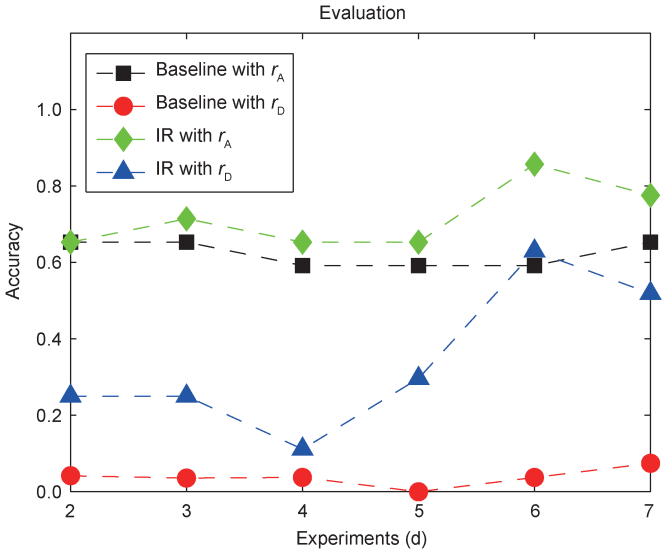

图 6. 机器人观察了气候条件、空调开关和门，学习操作者习惯，以便在开关和门上做动作，来调节环境状态。IR 是逆向增强学习的简称，r_A 表示回报函数的精确度，r_D 表示机器人动作的精确度。

但本文所提方法在引导机器人行动上更加准确。原因是，在新的状态下，基准办法必须在记录中搜索。但是，如果记录里没有行动 – 诱因，基准办法就无法找出正确的对策，而通过学习回报函数的提议方法，可以根据环境状态产生不同的行动。

5 结论

本文提出了以逆向增强学习为框架，让机器人根据观察学习操作者习惯的方法。用行为引发的环境状况变化描述行为，即行为足迹。机器人根据操作者和物体间的接触学习诱因，在屋内学习基于操作者行为回报函数

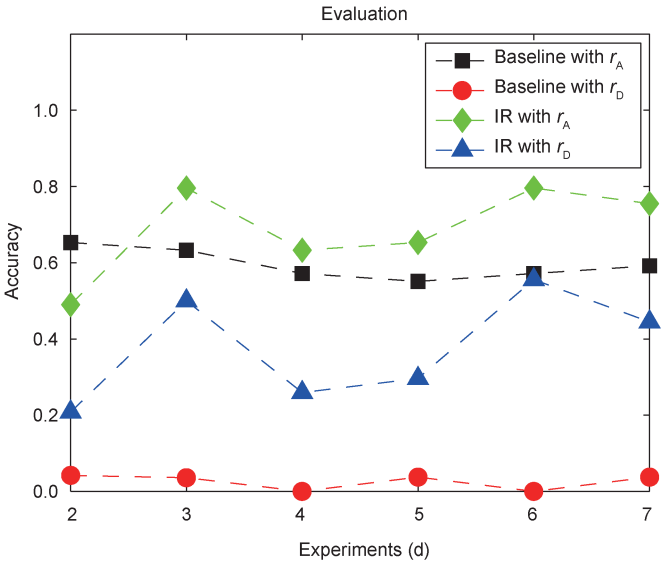

图 7. 机器人观察气候条件、空调开关、门和温度计，学习操作者习惯，以便在开关和门上做动作，调节环境状态。

的习惯。然后，使用回报函数引导其日后行动，以便自主地服务操作者。本文重点是学习调节室内温度，在住宅状态评估和机器人行动选择上比较提出的方法和基准办法。结果显示，本文提出的方法能够更准确地在复杂情景中引导机器人行动。

以后的工作中，该方法还可从多个方面改进。首先，基函数设计应更灵活，以描述并分析环境状态变化。此外，学习方法可以被改进，以覆盖除以一些基本功能为代表的习惯之外其他不同种类习惯。

致谢

该项目部分得到香港研究资助局支持，授予 Max Q.-H. Meng (CUHK14205914 和 CUHK415512)。

Compliance with ethics guidelines

Kun Li and Max Q.-H. Meng declare that they have no conflict of interest or financial conflicts to disclose.

References

1. W. Wood, D. T. Neal. A new look at habits and the habit-goal interface. *Psychol. Rev.*, 2007, 114(4): 843–863
2. W. Meeussen, et al. Autonomous door opening and plugging in with a personal robot. In: *Proceedings of 2010 IEEE International Conference on Robotics and Automation (ICRA)*, 2010: 729–736
3. R. B. Rusu, I. A. Sucan, B. P. Gerkey, S. Chitta, M. Beetz, L. E. Kavraki. Real-time perception-guided motion planning for a personal robot. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009: 4245–4252

4. J. F. Gorostiza, et al. Multimodal human-robot interaction framework for a personal robot. In: *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006: 39–44
5. K. A. Wyrobek, E. H. Berger, H. F. M. Van der Loos, J. K. Salisbury. Towards a personal robotics development platform: Rationale and design of an intrinsically safe personal robot. In: *Proceedings of IEEE International Conference on Robotics and Automation*, 2008: 2165–2170
6. E. Falcone, R. Gockley, E. Porter, I. Nourbakhsh. The personal rover project: The comprehensive design of a domestic personal robot. *Robot. Auton. Syst.*, 2003, 42(3–4): 245–258
7. L. Tonin, T. Carlson, R. Leeb, J. del R. Millán. Brain-controlled telepresence robot by motor-disabled people. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011: 4227–4230
8. T. C. Tsai, Y. L. Hsu, A. I. Ma, T. King, C. H. Wu. Developing a telepresence robot for interpersonal communication with the elderly in a home environment. *Telemed. J. E Health*, 2007, 13(4): 407–424
9. P. R. Liu, M. Q. H. Meng, P. X. Liu, F. F. L. Tong, X. J. Chen. A telemedicine system for remote health and activity monitoring for the elderly. *Telemed. J. E Health*, 2006, 12(6): 622–631
10. M. Baeg, J. H. Park, J. Koh, K. W. Park, M. H. Baeg. Building a smart home environment for service robots based on RFID and sensor networks. In: *Proceedings of International Conference on Control, Automation and Systems*, 2007: 1078–1082
11. N. Oliver, A. Garg, E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, 2004, 96(2): 163–180
12. S. Fine, Y. Singer, N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Mach. Learn.*, 1998, 32(1): 41–62
13. R. S. Sutton, A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998
14. B. D. Argall, S. Chernova, M. Veloso, B. Browning. A survey of robot learning from demonstration. *Robot. Auton. Syst.*, 2009, 57(5): 469–483
15. P. Abbeel, A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In: *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004: 1
16. A. Y. Ng, S. J. Russell. Algorithms for inverse reinforcement learning. In: *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000: 663–670
17. G. Grisetti, C. Stachniss, W. Burgard. Improved techniques for grid mapping with Rao-Blackwellized particle filters. *IEEE Trans. Robot.*, 2007, 23(1): 34–46
18. T. Joachims. Making large-scale SVM learning practical. In: B. Schölkopf, C. J. C. Burges, A. J. Smola, eds. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press. 1999: 169–184