中国科学: 信息科学 2016 年 第 46 卷 第 8 期: 969–981
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
信息科学与技术若干前沿问题评述专刊

面向真实世界的智能感知与交互

陈熙霖[①*], 胡事民[②], 孙立峰[②]

① 中国科学院计算技术研究所, 北京 100190
② 清华大学计算机科学与技术系, 北京 100084
* 通信作者. E-mail: xlchen@ict.ac.cn

收稿日期: 2016–03–30; 接受日期: 2016–05–30
国家自然科学基金 (批准号: 61390510) 资助项目

摘要 感知与交互既是智能机器重要且基本的能力和组成部分, 也是智能机器从现实中学习和获得知识的重要甚至唯一手段. 过去 20 年中感知与交互系统的研究往往是面向限定域的, 并且在若干领域取得了显著的进步. 随着服务机器人、无人驾驶车等的发展, 感知与交互系统将不可避免地需要面对真实世界的挑战. 本文简要地回顾了感知与交互领域的发展历程, 提出了 8 个面向真实世界感知与交互的挑战. 解决这些问题, 有助于将现有智能机器应对外界的能力从特定领域的 "专家" 提升到在一般意义上 "常人" 级水平的感知与交互.

关键词 感知 交互 计算机视觉 机器人

1 引言

经历近 70 年的发展, 计算机已经从过去单纯为科学计算服务渗透到人类生活的方方面面. 在这一过程中, 计算能力的提升是必不可少的因素. 站在人类用户的角度, 在过去 70 年中, 我们关心如何通过人机交互使计算机变得更加易用. 在这方面取得了天翻地覆的变化, 一个标志就是计算机从巨大的体积化身为无处不在的设备, 同时用户从需要长期专业训练到无需手册直接按照直觉使用. 交互手段的改变极大地改善了易用性, 使得非专业人员使用计算机成为可能. 人与人的沟通方式逐渐渗入人与计算机的交互之中.

与此同时, 从作为计算机载体的其他设备/装置的角度来看, 计算机也不再是单一呆板的控制台形象. 小到手表、大到各种航行器, 计算机在其中扮演着不可获缺的重要角色. 因此除了需要和人类用户交互之外, 计算机还需要和自然环境进行交互, 因而感知和交互的对象不仅仅是人类, 还需要和各种真实世界进行交互. 例如对机器人而言, 需要和其拥有者以及之外的其他人和环境进行交互, 对当今的大型飞机而言, 更像是一台会飞行的计算机, 在自动飞行以及盲降过程中都需要和环境进行有效的交互. 这就要求计算系统能够具备感知外界的能力, 实现类似人类的感知功能.

近年来, 人工智能也已经进入了新一轮的快速发展时期, 有一点是明确的, 感知和交互除了是智能行为自身最好的体现之外, 同时还是智能能力发育的基础和支撑. 如同人类只有探索世界才能获得

引用格式: 陈熙霖, 胡事民, 孙立峰. 面向真实世界的智能感知与交互. 中国科学: 信息科学, 2016, 46: 969–981, doi: 10.1360/N112016-00072

© 2016《中国科学》杂志社

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

知识的拓展一样 (也许纯数学不在此列), 智能只能在感知外界和与外界交互后才能得到提升. 就在本文成文的最后几天中, Google 的 AlphaGo [1] 击败了韩国的李世石, 似乎给公众的印象是人工智能相关的技术进了一大步, 但这其实只是封闭世界中若干已经成功的人工智能故事的续集, 距离处理真实世界的问题还差得很远.

从 20 世纪 50 年代中期开始, 人类对构建具有智能的装置寄予了极大的期望. 智能本身是一个涉及面很广的问题, 即便对于人工智能也是如此, 本文主要集中在其前端部分即智能的感知与交互上. 对于生物体而言, 其最基本的能力首先是发展感知和与环境交互的能力, 这是生存与探索世界的基础. 因此对一些低等生物而言, 首先发展的是感知和与外界互动的能力, 之后的高等动物才逐步发展出更为复杂的能力如语言、推理等等. 而一定程度上, 人工智能研究似乎与生物进化有着某种相反的趋势, 我们已经能够在一些高级的智力游戏方面取得成功, 如在国际象棋中深蓝战胜 Garry Kasparov [1)]、在有限域问答 Watson 战胜 Ken Jennings 和 Brad Rutter [2)], 以及在围棋上 AlphaGo 战胜李世石等, 但在基本的探索环境、理解环境和与环境交互等方面却远远落后. 对于智能装置而言, 自动感知能力是其中不可或缺的部分. 经过近半个世纪的探索, 在某些方面的感知能力已经取得了巨大的进步, 典型的如 Boston Dynamics 的 BigDog [3)], 但挑战依旧巨大. 本文在回顾过去发展历史的基础上对未来在处理真实世界中需要面临的挑战性问题进行了展望.

2 计算机感知研究回顾

对外界环境的感知是人类认识自然和适应自然的基本前提. 同样自动感知外界并和外界交互也是人工智能系统必不可少的部分. 在所有的感知手段中, 远程非接触感知能力是极其重要的, 因此以计算机视觉为基础的感知在过去 50 多年中得到了迅速的发展.

由于元器件和传感器的发展, 获取信息的手段不断提升, 特别是近 30 年, 包括光场成像、各种深度成像、可见光与红外成像、雷达成像等, 这些手段为计算机感知提供了强有力的原始信息获取手段, 使得计算机的感知能力得以在某些方面甚至超过人类的感知能力.

与此同时, 在感知外界的模型上, 早期试图以简单的几何构件来建模和理解世界 [2], 认识到这类简单模型存在的问题, 70 年代以 Marr 等为代表的学者提出了一整套以重构为核心的计算理论 [3], 并且引入广义锥体表达空间对象. 尽管重构看起来非常完善, 但不论是计算代价还是恢复的精度都面临很大的挑战. 90 年代初, 一些学者对此进行了深刻的反思 [4], 提出了包括主动视觉、定性视觉等方法, 试图解决这些问题, 从研究机器人角度出发的学者甚至认为所谓的感知与智能只需要有足够好的刺激响应机制就够了 [5]. 与此同时, 在感知对象的识别方面, 近 20 年在一些特定类别的识别上取得了长足的进步, 特别是在感知人和车辆等方面已经在很大程度上能够满足一些特定应用的需求.

尽管如此, 在自动感知外界方面, 对类别的识别能力还是存在极大的限制, 而且识别和感知的能力受到成像条件的严重制约; 同时, 识别对象往往只能是极少数的几类. 在过去十多年中, 研究者努力将这一领域加以拓展, 并取得了一些突破.

在利用多种传感器建模方面, 自动驾驶提供了很好的例证, 2005 年斯坦福大学的自动驾驶车通过采用 LIDAR、可见光相机以及 GPS 建模了覆盖不同距离范围的环境, 从而保证了能够完成 200 多公

1) https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer).
2) https://en.wikipedia.org/wiki/Watson_(computer).
3) https://en.wikipedia.org/wiki/BigDog.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 1 (网络版彩图) 完成 DARPA Grand Challenge 2005 后的 Stanley

Figure 1 (Color online) Stanley after the 2005 DARPA Grand Challenge (源自: https://commons.wikimedia.org/w/index.php?curid=2807954)

里的越野测试[4]. 图 1 是完成这一测试后满是尘土的 Stanley.

利用非传统成像直接获取深度或其等价表示与可见光成像的结合在交互领域同样也带来了显著的进步, Kinect [6] 及其后续的传感器如 RealSense 等的出现, 将室内环境等的交互和建模大大地向前推进了一步 [7,8], 极大地改善甚至是省略了以往充满不确定性的前处理过程 [9].

与此同时, 在向更加一般化对象的识别方面, 从单类对象不断向更多类扩展. 从在近乎理想条件下的 8 大类物体的集合 ETH-80 [10] 到简单背景的超过 100 类物体的 Caltech-101 [11], 研究者们在这些数据集合上取得的结果为后面应对更加复杂和具有挑战性的场景展现了可能, 同时处理的对象也不断逼近从真实环境中直接获取的结果. 在此情形下, 出现了如 PASCAL VOC [12] 和 ImageNet [13] 等更加接近真实情况的挑战集合. 随着在 ImageNet 上结果的不断改进, 似乎机器在这一集合上已经能够和人的能力相当. 与此同时, 两个更大规模的集合 Visual Genome [14] 和 YFCC100M [15] 将成为下一阶段的新目标.

在不断逼近复杂场景的过程中, 一些新的模型和方法也不断得到检验和发展, 从 Adaboost [16]、Bag of Word [17] 到 Deep Learning [18]. 从 90 年代中期对这一领域的反思 [4] 开始, 这一领域在提出挑战性的问题/场景, 推进方法, 逐步逼近解决问题, 再提出更加挑战的问题/场景的循环中不断向前发展.

3 智能系统中的感知与交互

关于人工智能已经有无数的讨论, 但其中最重要的就是到底我们需要的是僵化的人造智能还是划定疆界后的能够在限定空间中生长的人造智能. 如果是后者, 那么感知和交互除了作为低级的智能行为之外, 更加重要的是能够通过感知获得关于外界的知识从而达到生长的目的. 这里生长的含义就如同人类社会中对人的专业技能的学习、训练和积累, 甚至对于以往没有尝试过的任务, 通过探索掌握技能, 而且这种技能能够通过一个个体的学习向其他个体传播.

感知为智能机器提供了被动的信息获取能力. 相对于感知而言, 交互则可以提供一定程度上的主

4) https://en.wikipedia.org/wiki/Stanley_(vehicle).

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 2 (网络版彩图) 多尺度成像与建模

Figure 2 (Color online) Imaging and modeling in multiscale

动能力, 这种交互能力为感知增加了从多维度主动获取信息的能力, 如同所有动物的感知系统, 脱离交互的感知其获取信息的能力是非常有限的. 从孤立问题进行研究的角度, 感知和交互看似是相互独立的, 但实际上这两部分却常常是紧密相关的 —— 感知为交互的探索提供了关于交互对象、交互强度、交互效果的信息. 与此同时, 交互则为感知提供了多维度的感知能力.

从更高的认知层面出发, 感知和交互能够为作为智能主体的智能机器提供探索外界和从外界自动学习的能力. 这里的学习的含义区别于现在机器学习中仅仅调整参数的模型适应过程, 包括了自动模型的调整, 甚至模型的生成与演化. 这方面已经有一些初始的探索, Lake 等关于概率归纳学习的努力提供了一个有益的尝试 [19]. 尽管这一工作看起来还非常初级, 但提供了一种可能的途径. 如果把以往在单一智能行为中的学习看作是单一的在给定函数上的参数学习, 那么这种通过感知与交互的探索则可以提供对函数类的学习能力.

要达到使智能系统超越预先设定的限定域, 解决开放域的问题, 具备面向真实世界的感知与交互能力是其中的必要前提. 为实现这一目标, 必须要解决以下一些挑战性问题.

4 真实世界感知与交互的挑战性问题

虽然限定场景下的感知能够满足部分应用的需求, 但大规模的媒体内容搜索、服务机器人、自动驾驶的汽车和飞机等都需要面对真实世界随时发生的千变万化, 感知其中的态势. 这些都是现在方法和系统难以应对的. 我们不仅需要在安静环境下特定人的语音识别, 也需要在噪声环境下的非特定人语音识别; 不仅需要识别正面的人脸, 还需要能够从体态、行为习惯等识别人的身份; 不仅需要能够驾驶车辆在路况良好的路面行驶, 还需要理解拥堵、理解行人以及各种非机动车的行动意图. 对于智能系统而言, 其智能不仅来自于系统构造之前和构造过程中的学习, 还需要能够通过场景自动感知学习, 实现从感知到认知的过渡和连接. 应对真实环境的不确定性, 就需要解决以下一些挑战性问题.

4.1 挑战 1: 真实世界的多粒度建模

随着考察对象尺度的不同, 在感知外界环境时需要考虑的尺度也是变化的. 对航天飞行器而言, 地球可以简化为一个球体, 但对一般飞机而言, 高山、峡谷就是不能忽略的对象, 至于自动驾驶的车辆, 就需要考虑十厘米量级的障碍. 类似于图 2 的情形, 真实世界需要考虑的是多粒度的建模.

以往在很多任务中, 对外界的建模往往是作为一个正问题来解决, 如 Google Maps 等是以地理坐标为基础结合 LIDAR 和图像数据进行建模. 而对需要自主感知环境的智能机器而言, 这一正问题的假设常常是得不到保证的, 例如对于陌生室内环境甚至外星球环境的建模.

近年来相关工作在一些特定的问题上取得了较好的进展, 典型的如以下几方面.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1. 在城市场景的建模方面利用多传感器的建模已经得到了一定程度的应用, 这类建模一般依赖于包括激光测距仪、全景相机等多种传感器集成的获取装置. 这类装置可以是车载甚至是背负式的. 尽管这种环境下可以完成对空间场景的建模 [20, 21], 但这类建模往往是与后续的应用相分离的, 不能支持现场的建模应用, 因而其应用范围仍然是有限的. 与此同时在利用零散的图像进行城市建模方面, 也已经有一些探索性的结果, 文献 [22] 中利用未经配准的网络照片实现了大规模城市场景的恢复与建模. 尽管这类工作在精度、灵活性、计算复杂度等方面还有很大的改进空间, 但已经为实现灵活的多粒度建模提供了一种可行的途径.

2. 在室内有限空间的建模方面, 通过融合多传感器, 已经能够实现较为精确的对象建模, 典型的如 Leap Motion 和 Kinect 等传感器. 通过融合多种感知源并在时空上对齐所获取的不同信息实现有效的高精度三维建模. 利用 Leap Motion 可以在 60 cm × 60 cm × 60 cm 的空间内对 10 个手指达到 0.01 mm 的定位精度. 这为人机交互提供了可靠的交互手段. 利用以 Kinect 为代表的消费级深度传感器, 我们可以方便地实现对小范围室内场景的实时建模 [23, 24]. 虽然由于成本限制, 现阶段的深度传感器的采集范围和精度都很有限, 但是随着互联网上免费三维数据的数量逐渐增长和越来越成熟的机器学习技术, 数据驱动的建模方法 [25, 26] 也在一定程度上弥补了采集设备的缺陷. 这些相关技术的发展为虚拟现实、增强现实等后续应用提供了有力的支持.

3. 3D 电影的兴起, 将三维场景的重建与数字视频紧密地联系在一起. 除了新产生的 3D 电影之外, 将一些经典的电影转换为 3D 电影的需求也在不断增长. 通过利用运动、聚焦等信息的三维自动重建也引起了学术和产业的兴趣, 借助自动生成和交互修改, 已经有很多电影因此衍生出新的 3D 版本. 这一过程中除了采用自动的三维建模之外, 还不得不耗用大量的人工交互过程, 以电影《泰坦尼克》为例, 据称使用 300 位工程师工作超过 60 周. 通过技术的改进 [27, 28], 这类工作的效率有望得到进一步提高.

尽管这些建模方法对于特定领域的建模发挥了越来越重要的作用, 面对网络上千变万化的多媒体信息, 现有的建模手段还远远不能满足其要求. 即使是对于处于复杂环境中的服务机器人、无人机等而言, 虽然可以有多种感知手段, 要实现多粒度的建模, 就必须解决不同尺度、不同模态甚至不同时空感知信息的多维度对齐 (registration) 问题. 对于这些非同源信息, 由于不可避免地存在信息缺失、精度差异、视角和形态变化等问题, 因此对齐过程中必须要考虑利用多种辅助手段消除其中的不确定性.

4.2 挑战 2: 非均匀的感知与处理

高等生物的视觉系统几乎都是非均匀的. 这种非均匀性同时表现在感知空间精度和对刺激的响应上. 以人类视觉系统为例, 在人眼中大约有 600~700 万个视锥细胞, 9000 万个视杆细胞, 如果只按照中央凹的密度考虑, 人眼在水平 120°, 垂直 60° 的范围内大约相当于 6 亿左右的像素; 可另一方面, 同时具有精细观察能力的等价像素只是中央凹部分的 600~700 万像素, 视杆细胞所产生的边缘视觉分辨率则极低, 正是这种机制加上注意选择保证了生物视觉系统能够在精度、反应速度、视野之间做到很好的平衡.

以往图像/视频的采集、记录是以服务人类视觉系统的再观察为首要目标的, 因此均匀采样、线性量化都是与这一目标相符的. 但对感知目的而言, 就必须兼顾分辨率、视野和传输以及处理能力, 针对非均匀采样和非线性量化获取信号就需要探索新的感知与处理方法. 在这些方面, 尽管已经有一些尝试, 如仿枝节动物眼睛成像 [29]、Catadioptric 成像 [30]、非线性量化 [31] 等, 但在后续的处理上通常还是通过将这些非均匀的成像映射回均匀的几何模型下进行处理, 因而失去其优势. 这方面需要发展的理论和方法如下所述.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1. 新型成像模型. 传统的透视投影模型都是基于均匀采样的, 非均匀成像意味着需要有相应的成像模型. 同时要考虑从中央到边缘的均匀过渡和物体在不同区域表达的连续性与不变性.

2. 非均匀的特征获取理论与方法. 以往的均匀量化成像虽然给显示带来了很大的方便, 但在特征表达方面, 其实并非是最为有效的方式. 一个典型的例子就是边缘获取往往是基于差分的, 同样的边缘结构, 在照度改变之后差分的结果会有很大差别. 因此, 均匀量化对于很多特征提取方法而言并非是最适合的数字化方法, 探索新的非均匀量化方法, 如对数量化等, 这不仅有利于提升感光范围, 而且可以有效支持稳定特征的获取.

3. 分辨率依赖的控制与任务调度. 对于非均匀采样的系统, 如同人类中央视觉和边缘视觉的分工一样, 需要有相应的任务分工; 同时为了保证系统能够持续处理某一类任务, 如跟踪, 需要能够对采集进行连续的主动控制, 这可以体现出主动视觉的优点, 同时大大降低后续处理的代价.

4.3 挑战 3: 对象的识别与理解

对动物而言, 不论是发现捕食对象还是逃避天敌, 识别对象和环境都是感知系统不可或缺的功能. 这也是计算机视觉系统的重要目标, 但是早期的视觉系统在识别能力上远远不能满足需求. 近年来在特定类别对象的识别上所取得的进展为一般意义上的识别提供了很好的基础. 对于某些单一类别问题, 如人脸、车辆, 视觉系统的识别能力已经达到了在某些场合可以替代人工的水平. 特别是近年来深度学习引入到语言、图像识别上后, 将识别系统的性能提升了一大步 [18, 32]. 但需要注意的是在这类任务中, 对象还是相对简单和孤立的, 这与真实环境中的物体处于复杂背景中是不同的. 这方面需要进一步解决的问题如下所述.

1. 复杂背景下对象的分割. 这在计算机视觉领域是一个经典的问题, 由于成像过程中的信息丢失, 因而这是一个典型的病态问题. 在只有可见光图像的情况下, 分割与识别将构成一个假设检验的过程, 以期获得正确的结果. 考虑到一些信息传感器如光场相机、深度相机等的使用, 分割问题有望得到更好的解决.

2. 非刚体变形对象的识别. 在所有的物体识别问题中非刚体变形以及复杂的类内变化是识别必须解决的问题, 典型的情形如图 3 中品种不同、形态各异的狗. 因此需要发展有效的方法解决这类问题. 尽管深度学习对这类问题似乎可以取得较好的结果, 但前提是需要大量的训练数据支撑. 发展小数据下的方法是未来需要重点研究的问题之一.

3. 对成像条件不敏感的对象识别方法. 尽管深度学习 + 大数据似乎提供了强大的识别框架, 但从成像物理和几何关系出发, 研究对成像条件不敏感的识别方法可以将现有方法在数据和计算上的复杂度大大降低.

4. 新型传感器下的识别方法. 随着新型传感器如深度、光场以及 HDR 相机等的普及, 需要研究相应的特征表示和识别方法, 使之能够最大程度地发挥这些传感信号所带来的优势. 例如消费级双目相机、红外深度传感器等设备的普及, RGB-D (彩色深度) 图像的获取变得极为方便. RGB-D 图像更符合人眼对世界的感知模式, 经常作为 “眼睛” 应用于机器人上. 因此基于 RGB-D 图像、甚至三维模型的对象识别与理解也变得极为重要. 传统的基于深度图像的特征与方法 (如文献 [33, 34]) 在处理杂乱无章的真实世界数据的时候遇到了极大的挑战. 以深度学习为代表的模型为这方面的研究带来了极大的契机, 相关数据集也在逐渐丰富 [35]5).

5) http://shapenet.cs.stanford.edu.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

图 3 (网络版彩图) 类别识别中需要面对的挑战 —— 形态各异的同类对象

Figure 3 (Color online) Challenge in categorization – appearance variations

4.4 挑战 4: 人类动作行为的理解与协同

感知与交互的重要对象是人, 手势、动作和体态是除了语言之外最为重要的交互手段, 因此对真实世界的交互就不可避免地要解决人类动作行为的识别与理解. 关于人的感知是最近 20 年中计算机视觉研究的最重要主题之一, 尽管在身份识别等方面取得了很好的进展, 并且在很多领域得到了非常成功的应用, 但动作行为的理解却始终是一个巨大的挑战. 在动作行为的理解方面分为以下几类.

1. 人为定义的命令式动作, 如交警的手势、旗语等. 这类动作由于相对规范, 同时词汇集合较小, 因此能够较好地解决. 这方面的研究在包括新型概念车、手术辅助机器人等场合取得了一定的进展 [36].

2. 部分人为定义的手语. 这类动作具有一定的约束, 但由于地区、个体等原因, 以及词汇量较大带来的问题, 在自动识别理解方面还存在很多问题需要解决. 随着深度传感器等的使用, 在过去几年中可自动识别的词汇集合规模有了显著的提升 [37].

3. 日常生活中的一般动作, 这类动作由于缺乏规范性, 同时与其他多种环境因素等有关, 因而其识别理解具有极大的挑战性. 在日常行为中, 这类动作往往还会和其他无意识的动作混在一起, 因而理解有意义的行为就更加需要依赖上下文的联系, 但要达到能够实际应用还面临诸多挑战.

4. 微动作的感知与理解. 微动作一般是指由潜意识导致的具有很短持续时间的动作行为, 这类行为虽然持续时间很短, 但对于理解用户的交互意愿和心理具有极其重要的作用, 这是实现真实自然交互的重要体现. 一个典型的例子是在交谈中, 有人下意识地看一下手表, 从而反映出其对结束谈话的预期或后面还有其他事务需要处理. 关于微动作的感知和理解直到最近才被关注 [38].

上述 4 类问题从难度上依次加大, 对于需要融入人类社会中的智能机器而言, 后两类问题的解决尤为重要, 这可以保证人类用户在不改变交互习惯的情况下实现人机的融合, 使得机器能够理解人类的行为与通过行为所表达的意向.

4.5 挑战 5: 识别、概念与语言

研究表明, 5~6 岁的儿童能够轻易辨别上万种对象, 而这种能力显然不是通过为每一类分别建立一个分类器实现的. 解决海量类别的识别及其关系推理问题对于理解真实世界具有决定性作用. 但是, 面对数以万计的物体类别, 以往在单类识别中的方法显然会很快遇到瓶颈. 从人类认识事物的基本过程出发, 不难看出蕴含在类别背后的概念及其特征 (属性) 是导致类别间差别和联系的根本原因, 认知神经科学关于集群编码的假设也为这一层级化的关系提供了支持. 因此如何自动发现类别、概念以及

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

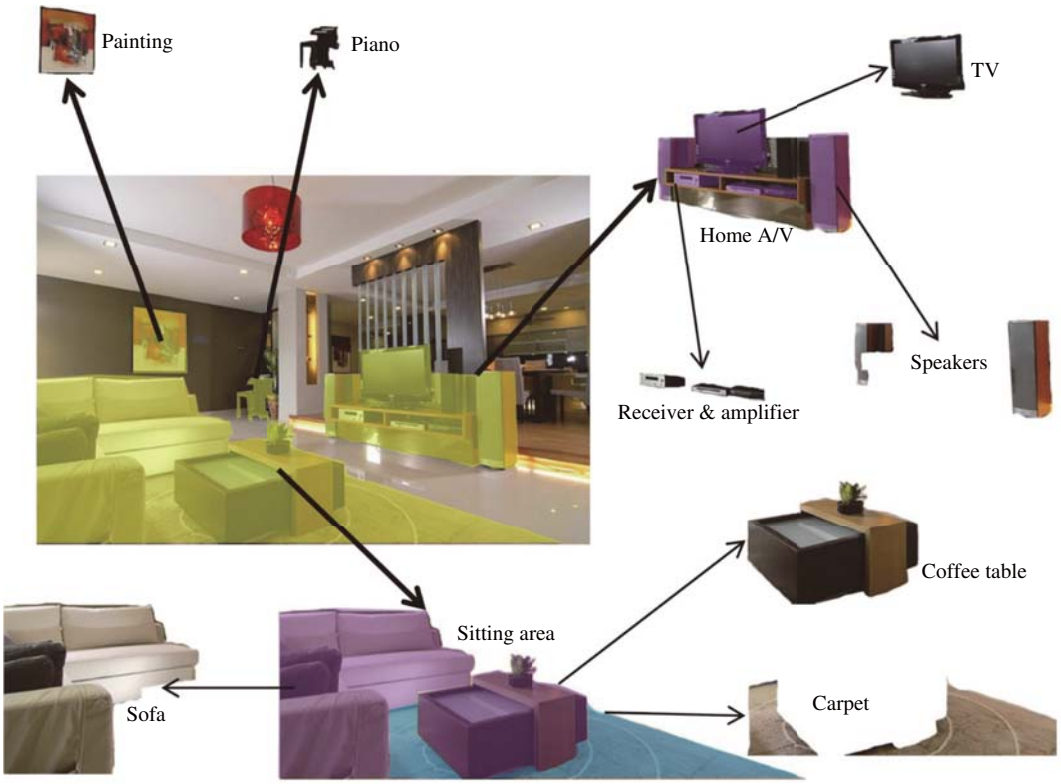


图 4 (网络版彩图) 场景与物体的分层识别与理解

Figure 4 (Color online) Hierarchical object recognition and scene understanding

类别 – 类别间、类别 – 概念间和概念 – 概念间关系就成为不断提升识别、理解能力的基础. 为了应对真实环境中复杂对象的识别和关系表达, 就需要解决分层次的物体识别和理解的问题, 如图 4 所示.

对于复杂的关系, 在表达上离不开语言的支持. 关于语言认知的研究表明, 语言是连接类别、概念直至推理的纽带和基础. 因而对于真实世界的识别与深度理解离不开概念和语言的支撑. 这方面需要解决的问题如下所述.

1. 从少量示例中自动进行概念的学习, 并籍此能够获得关于类别和属性的自动获取与分类; 近来关于 Zero-shot 识别的研究逐渐引起了研究者的兴趣, 这为自动识别和处理未知对象提供了一种可能 [39, 40];

2. 从场景图象自动发现类别间、类别属性间以及属性间的关系, 从而为形成属性、类别直至场景的自动理解与推理提供支撑. 人类关于客观世界中对象间关系的建立往往来源于对外界的观察, 对于智能机器同样如此. 在智能机器构造的过程中将所有关于外界的知识全部注入其中是不可能的, 因而需要提供的是通过场景发现并形成关系的能力.

3. 基于对场景中对象的识别和对象间关系的理解, 实现对真实世界的描述与推理. 对上面两个问题的解决有助于进一步完成对场景的理解, 其中最为简单的任务是从场景生成总结式的描述 (caption), 更为复杂的问题包括理解细节和对象间的关系, 并能够回答关于场景的问题. 这些关系需要通过对大量的场景自动学习和文本知识的学习获得.

上述这些问题近年来逐渐引起研究者的重视, 一些初步的结果给这一方向的研究带来了希望 [19, 41, 42], 同时相应的数据也为这类研究提供了可能 [14, 15].

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.6 挑战 6: 感知与主体的结合

感知是连接主体 (智能机器) 与客体之间的桥梁, 虽然研究中常常把感知问题如计算机视觉、语音识别等孤立出来进行研究, 但很多问题是和主体环境密不可分的, 对于这类问题如自动驾驶、机器人导航等, 必须与主体相结合. 这种结合为感知系统和作为主体的智能机器提供了很好的互补信息. 典型的情形如主体的运动可以为视觉系统提供从不同角度观察环境的能力, 从而弥补了视觉系统基线长度受限、以及成像光照和姿态的问题, 同时可以提供从对象已知视角学习未知视角表示的能力; 同样变化头部位置与朝向为听觉系统提供了更加准确的辨认声源的能力. 同时不同感知器官间相互补充, 可以提供对客体的更加完整的表达与理解.

一旦感知系统随主体进入真实环境, 感知系统就不可避免地需要面对开集问题, 以往闭集条件下的一些假设如解的存在性就不再能够得到保障, 因而问题的难度将大大增加. 与此同时, 与主体结合的感知一旦出错, 所付出的代价也会大大增加. 例如对于 Apple Siri 或 MS Cortana 而言, 错误地识别/回答一段问题或命令不会造成致命的错误, 而对于语音控制的自动驾驶而言, 错误理解命令的代价却可能是致命的. 就在本文即将付印之时, Tesla Model S 的事故引起了关于自动驾驶的广泛关注. 因此需要研究感知与主体结合的安全性和鲁棒性问题. 正是这些原因导致现在多数无人机还是协同驾驶的, 需要飞行员在远程进行操作 [43].

与主体的结合一方面增强了感知的能力, 同时对感知系统本身而言也面临着新的挑战. 对感知的实时性要求更高, 同时要和相应的交互响应构成环路. 典型的感知问题, 如在 DARPA 的 Robot Challenge 中为了解决开门、关闭阀门等动作, 除了通过视觉进行定位、识别之外还需要能够解决手臂力量的灵活调节与控制问题, 这在服务机器人、抢险救灾机器人等应用中尤其重要. 因此需要解决如下关键问题.

1. 多种感知手段的融合. 由于不同模态感知手段在作用距离、感知特性等方面的差异, 融合多种感知手段可以显著地提升主体对外界的感知能力. 近来一些典型的融合如融合深度与可见光传感器, 包括激光雷达 + 可见光相机, 超声波测距 + 可见光相机等. 由于多种感知手段所提供的信息有时不仅是不相关的甚至是抵触的, 故而对多模态信息需要提供融合和决策支持, 这与以往单一感知有着重要的差别.

2. 感知与响应系统的交互作用, 不论是在以往的感知还是交互系统中, 都是相对独立的. 一旦通过主体相关联, 感知与响应之间的相互作用将是一个非常重要的问题, 这其中的两个关键是实时性 (从感知到响应) 和感知与响应的环路集成. 前者在如飞行驾驶中的自动着陆等具有高速行为的应用中尤为重要, 后者对如救灾机器人施救力量的控制等则是决定性的.

4.7 挑战 7: 用户融入环境

之前 6 个问题主要关注在真实环境中机器如何能够无缝地融入其中, 并理解环境中的各种对象及其关系. 与智能机器融入真实世界不同的一点是, 有些应用中需要通过虚实结合让人类用户通过对虚拟世界的交互实现对远程真实世界的遥操作. 因而重要的问题是解决交互环境生成的响应速度与真实感问题.

通过感知系统为用户提供对远程环境具有沉浸感的刺激是虚拟现实和增强现实系统非常重要的功能. 真实环境中不仅存在作为外界对象的人, 同时还存在作为操作用户与机器共融在同一感知空间的人. 对于这部分用户, 需要通过机器将其融入到远程环境中, 从而为这些用户提供将本地操作映射到远程空间的能力. 典型的如遥控无人驾驶飞机, 由于传输延时等的存在, 这时的操作不同于单纯的

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

对虚拟世界的操作, 随着空间差距的变化, 延迟也会发生变化, 如何提供保证时空实时性的具有融入感的增强环境就成为其中重要的挑战. 对用户而言, 远程的感知来源于各种视觉信号的刺激, 由于不可避免地存在传输延迟, 同时受传输带宽的限制, 因此保证实时性的唯一手段就是高速的获取和高效快速的编码压缩.

同时对于操作者则言, 操作对象的尺度与操作控制尺度之间的比例控制也是极其重要的问题, 这方面不论是遥控机械挖掘机还是手术机器人, 都需要很好地解决从操作尺度到操作对象空间尺度的映射. 这与驾驶汽车中控制转向的机构类似, 除了需要设计合理的尺度关系之外, 对用户的训练和标准化也是必不可少的.

4.8 挑战 8: 计算能力

计算能力一直是推动智能方法/技术发展的重要基础和限制其应用的制约因素. 过去 30 年的历史表明, 这一领域的重要方法往往是随着计算能力提升才逐渐受到重视的. PCA (principal component analysis) 作为一种有效的线性降维手段, 尽管在 60 年代就得到了充分的认识, 但受限于计算能力, 直到 90 年代才逐渐得到广泛应用 [44]. 即使是今天大行其道的深度学习, 其基本思想在 90 年代初就已经形成 [45], GPU 的应用对于推动后来的应用起到了关键的作用. 要满足处理真实世界感知与交互这一任务, 今天的计算装置在处理能力与功耗上仍然存在巨大挑战, 一个例子是即使是对阵欧洲冠军 Fan Hui 时, AlphaGo 也使用了 48 个 CPU 和 8 个 GPUs [1]. 而处理真实环境的感知与交互需要保证计算装置的体积和功耗与笔记本计算机相当甚至更小, 因此至少需要将现有的计算能力提升千倍以上, 同时功耗需要降低为现在同等处理能力的千分之一.

5 结束语

感知与交互是智能机器必不可少的能力之一, 这一方面保证了与外部世界的连接, 同时是智能机器获取外界知识的重要手段. 以往关于感知与交互的研究与作为一大类感知与交互系统的实体如机器人的研究常常是脱节的, 因而感知系统的能力 (如主动感知) 往往受到缺乏实体支撑的限制, 同时也制约了机器人等的发展. 过去 20 年感知与交互系统在解决限定领域的问题上取得了重要的进展, 但和人工智能的其他领域类似, 今后需要重点突破的是处理开放域的问题, 使得未来的智能系统不仅仅是限定域问题的 “专家”, 同时也是能够应对开放域问题的具有学习能力的 “常人”.

参考文献

1 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489

2 Roberts L. Machine perception of three-dimensional solids. In: Optical and Electron-optical Information Processing. Cambridge: MIT Press, 1965. 159–197

3 Marr D. Vision: a Computational Investigation Into the Human Representation and Processing of Visual Information. Cambridge: MIT Press, 2010

4 Jain R C, Binford T O. Ignorance, myopia, and naiveté in computer vision systems. CVGIP: Image Und, 1991, 53: 112–117

5 Brooks R A. Intelligence without reason. In: Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, 1991. 569–595

6 Zhang Z. Microsoft Kinect sensor and its effect. IEEE Multimed, 2012, 19: 4–10

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

7 Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from single depth images. Commun ACM, 2013, 56: 116–124

8 Fankhauser P, Bloesch M, Rodriguez D, et al. Kinect v2 for mobile robot navigation: evaluation and modeling. In: Proceedings of the 17th International Conference on Advanced Robotics, Istanbul, 2015. 388–394

9 Han J, Shao L, Xu D, et al. Enhanced computer vision with Microsoft Kinect sensor: a review. IEEE Trans Cyber, 2013, 43: 1318–1334

10 Leibe B, Schiele B. Analyzing appearance and contour based methods for object categorization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, 2003. 2: 409–415

11 Li F-F, Rob F, Pietro P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Comput Vis Image Und, 2007, 106: 59–70

12 Everingham M, van Gool L, Williams C K I, et al. The PASCAL visual object classes (VOC) challenge. Int J Comput Vision, 2010, 88, 303–338

13 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, 2009. 248–255

14 Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. arXiv:1602.07332

15 Thomee B, Elizalde B, Shamma D, et al. YFCC100M: the new data in multimedia research. Commun ACM, 2016, 59: 64–73

16 Viola P, Jones R. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, 2001. 2: 524–531

17 Fei-Fei L, Perona P. A Bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 2: 524–531

18 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems 25, Lake Tahoe, 2012. 1097–1105

19 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. Science, 2015, 350: 1332–1338

20 Chen J, Chen B. Architectural modeling from sparsely scanned range data. Int J Comput Vision, 2007, 78: 223–236

21 Lin H, Gao J, Zhou Y, et al. Semantic decomposition and reconstruction of residential scenes from LiDAR data. ACM Trans Graphics, 2013, 32: 1–10

22 Agarwala S, Furukawaa Y, Snavely N, et al. Building Rome in a day. Commun ACM, 2011, 54: 105–112

23 Newcombe R A, Izadi S, Hilliges O, et al. KinectFusion: real-time dense surface mapping and tracking. In: Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, 2011. 127–136

24 Henry P, Krainin M, Herbst E, et al. RGB-D mapping: using depth cameras for dense 3D modeling of indoor environments. In: Proceedings of the International Symposium on Experimental Robotics, New Delhi and Agra, 2010. 22–25

25 Nan L, Xie K, Sharf A. A search-classify approach for cluttered indoor scene understanding. ACM Trans Graphics, 2012, 31: 1–10

26 Chen K. Lai Y-K, Wu Y-X, et al. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. ACM Trans Graphics, 2014, 33: 1–12

27 Zhang L, Vazquez C, Knorr S. 3D-TV content creation: automatic 2D-to-3D video conversion. IEEE Trans Broadcast, 2011, 57: 372–383

28 Karsch K, Liu C, Kang S B. Depth transfer: depth extraction from video using non-parametric sampling. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 2144–2158

29 Song Y M, Xie Y, Malyarchuk Y, et al. Digital cameras with designs inspired by the arthropod eye. Nature, 2013, 497: 95–99

30 Yokoya R, Nayar S K. Extended depth of field catadioptric imaging using focal sweep. In: Proceedings of the 15th IEEE International Conference on Computer Vision, Santiago, 2015. 3505–3513

31 Nayar S, Mitsunaga T. High dynamic range imaging: spatially varying pixel exposures. In: Proceedings of IEEE

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Conference on Computer Vision and Pattern Recognition, Hilton Head, 2000. 472–479
32 Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag, 2012, 29: 82–97
33 Johnson A, Hebert M. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans Pattern Anal Mach Intell, 1999, 21: 433–449
34 Bo L, Ren X, Fox D. Depth kernel descriptors for object recognition. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, 2011. 821–826
35 Xiao J, Owens A, Torralba A. SUN3D: a database of big spaces reconstructed using SfM and object labels. In: Proceedings of the 14th IEEE International Conference on Computer Vision, Sydney, 2013. 1625–1632
36 Jacob M G, Li Y-T, Akingba G A, et al. Collaboration with a robotic scrub nurse. Commun ACM, 2013, 56: 68–75
37 Chai X, Li G, Chen X, et al. VisualComm: a tool to support communication between deaf and hearing persons with the Kinect. In: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, Bellevue, 2013. 76
38 Häuslschmid R, Menrad B, Butz A. Freehand vs. micro gestures in the car: driving performance and user experience. In: Proceedings of IEEE Symposium on 3D User Interfaces (3DUI), Arles, 2015. 159–160
39 Lampert C H, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 453–465
40 Liang K, Chang H, Shan S, et al. A unified multiplicative framework for attribute learning. In: Proceedings of the 15th International Conference on Computer Vision, Santiago, 2015. 2506–2514
41 Malinowski M, Rohrbach M, Fritz M. Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the 15th International Conference on Computer Vision, Santiago, 2015. 1–9
42 Liu H, Wang R, Shan S, et al. Deep supervised hashing for fast image retrieval. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016
43 Ross P. When will software have the right stuff? IEEE Spectrum, 2011, 48: 38–43
44 Kirby M, Sirovich L. Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans Pattern Anal Mach Intell, 1990, 12: 103–108
45 Schmidhuber J. Learning complex extended sequences using the principle of history compression. Neural Comput, 1992, 4: 234–242

Towards real world perception and interaction

Xilin CHEN[1*], Shi-Min HU[2] & Lifeng SUN[2]

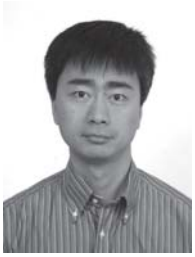
1 *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*;
2 *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China*
*E-mail: xlchen@ict.ac.cn

Abstract Perception and interaction are the most important and essential parts of an intelligent machine. They are crucial and even unique channels by which to learn from the real world. In the past two decades, there has been significant progress in closed world research on perception and/or interaction. With the current rapid developments in the areas of service robots and unmanned vehicles, perception and interaction are confronted with challenges from the real world. This paper briefly reviews the history of computer perception and interaction, and lists eight problems in real world perception and interaction that, if solved, will elevate the perception and interaction capabilities of intelligent machines from a specialist- to human-level in the real world.

Keywords perception, interaction, computer vision, robot

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Xilin CHEN received a Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China in 1994. Currently, he is a professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, and multimodal interfaces. He is an IEEE and CCF Fellow.



Lifeng SUN received a Ph.D. degree in systems engineering from National University of Defense Technology, Changsha, China, in 2000. Currently, he is a professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include video streaming, video coding, video analysis, and multimedia cloud computing. He is a member of ACM, IEEE CAS VSPC TC, and IEEE ComSoc MMC TC.

Shi-Min HU received a Ph.D. degree from Zhejiang University, Hangzhou, China in 1996. He is currently a professor in the Department of Computer Science and Technology, Tsinghua University, Beijing. His research interests include geometry processing, image and video processing, rendering, computer animation, and computer-aided geometric design. He is a member of ACM, a senior member of IEEE, and a CCF Fellow.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net