HZ BOOKS
华章教育

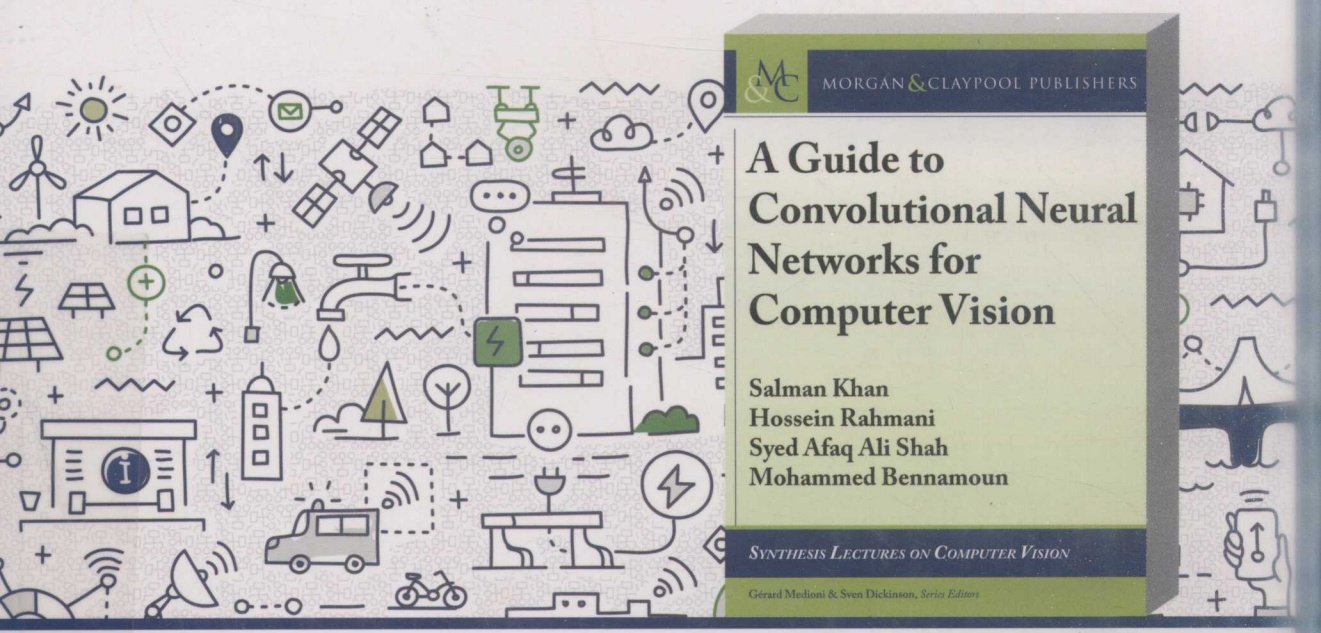
智能科学与技术丛书

A Guide to Convolutional Neural Networks for Computer Vision

卷积神经网络与计算机视觉

[澳] 萨尔曼·汗（Salman Khan）
侯赛因·拉哈马尼（Hossein Rahmani）
赛义德·阿法克·阿里·沙（Syed Afaq Ali Shah）
穆罕默德·本纳努恩（Mohammed Bennamoun） ◎著

黄智濒 戴志涛 ◎译



机械工业出版社
China Machine Press

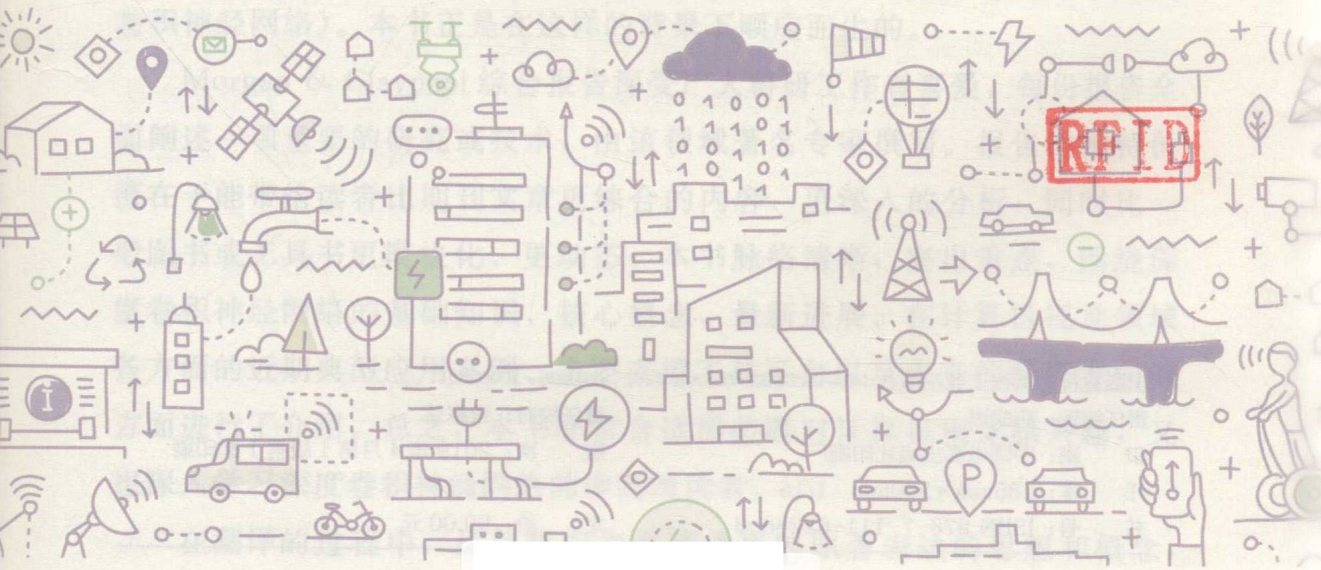
智能科学与技术丛书

A Guide to Convolutional Neural Networks for Computer Vision

卷积神经网络与计算机视觉

[澳] 萨尔曼·汗（Salman Khan）
侯赛因·拉哈马尼（Hossein Rahmani）
赛义德·阿法克·阿里·沙（Syed Afaq Ali Shah）
穆罕默德·本纳努恩（Mohammed Bennamoun） ◎著

黄智濒 戴志涛 ◎译



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

卷积神经网络与计算机视觉 /（澳）萨尔曼·汗（Salman Khan）等著；黄智濒，戴志涛译．
—北京：机械工业出版社，2019.4
（智能科学与技术丛书）
书名原文：A Guide to Convolutional Neural Networks for Computer Vision

ISBN 978-7-111-62288-8

I. 卷… II. ①萨… ②黄… ③戴… III. 计算机视觉－研究 IV. TP302.7

中国版本图书馆 CIP 数据核字（2019）第 051731 号

本书版权登记号：图字 01-2018-5308

本书既全面介绍了卷积神经网络（CNN）的原理，又提供了将 CNN 应用于计算机视觉的一手经验。书中首先讲解神经网络的基本概念（训练、正则化和优化），然后讨论各种各样的损失函数、网络层和流行的 CNN 架构，回顾了评估 CNN 的不同技术，并介绍了一些常用的 CNN 工具和库。此外，本书还分析了 CNN 在计算机视觉中的应用案例，包括图像分类、目标检测、语义分割、场景理解和图像生成。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：唐晓琳　　责任校对：张惠兰
印　　刷：中国电影出版社印刷厂　　版　　次：2019 年 4 月第 1 版第 1 次印刷
开　　本：185mm×260mm　1/16　　印　　张：12.25
书　　号：ISBN 978-7-111-62288-8　　定　　价：99.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88379833　　投稿热线：（010）88379604
购书热线：（010）68326294　　读者信箱：hzjsj@hzbook.com

1998年，Yann LeCun教授提出了第一个真正意义上的卷积神经网络LeNet，并将它应用到手写数字识别上。然而这个模型在后来的一段时间并未流行起来，主要原因是卷积神经网络虽然可以有效处理噪声信号，提取输入数据的特征，但需要较大的计算量，受限于当时的计算能力，识别的错误率比同时代的支持向量机要高很多。支持向量机精巧地在容量调节上选择了更合适的平衡点，从而获得了较大的成功，但其噪声信号处理能力较差。随着计算机性能沿摩尔定律的提升，多核处理器、通用图形处理器以及各类高性能分布式计算模式的出现，计算能力有了突飞猛进的发展。终于，在2012年ImageNet大规模视觉识别挑战赛（ILSVRC）上，AlexNet的卓越表现，使得人们认识到卷积神经网络的巨大潜力，各类深度网络结构层出不穷，各种卷积神经网络的应用如雨后春笋般冒出来。其中计算机视觉领域的应用最有代表性，并获得了巨大成功，同时也促进了人们学习并深入研究深度神经网络（特别是深度卷积神经网络）。本书正是在这样的背景下顺应而生的。

Morgan & Claypool综合报告深受广大科研工作者喜爱，每份报告全面阐述一项重要的研究或技术，由该领域著名专家撰写。报告的独特价值在于能带给读者比期刊文章更综合的内容、更深入的分析，同时比一般图书或工具书更模块化、更动态。本书脉络清晰，突出重点，围绕深度卷积神经网络的基础知识、核心概念、最新进展、在计算机视觉领域各方面的近期典型应用案例、主要支撑工具平台以及未来的研究方向等方面进行了介绍。总之，本书非常合适那些既对计算机视觉感兴趣，又想深入学习深度卷积神经网络的中高级读者。

在翻译的过程中，虽然我们力求准确反映原著表达的思想和概念，但由于本书内容大部分来自于近期有影响力的国际期刊和会议论文，有很多新名词尚没有标准的中文译名，因此只能通过查询互联网来选择被广泛接受的中文译名。我们将这些译名整理成术语表附在本书最后，希

望本书的出版能推动这些中文译名的统一化，便于国内的研究学习和交流。由于译者水平有限，翻译中难免有错漏之处，恳请读者和同行批评指正。

最后，感谢家人和朋友的支持和帮助。同时，要感谢在本书翻译过程中做出贡献的人，特别是北京交通大学附属中学的韩乐铮，北京邮电大学的董丹阳和赵达菲；还要感谢北京邮电大学计算机学院的大力支持。

北京邮电大学计算机学院

智能通信软件与多媒体北京市重点实验室

黄智濒　戴志涛

2018 年 11 月于北京

本书的主旨是从计算机视觉的角度全面深入地介绍卷积神经网络（Convolutional Neural Network，CNN）的主题，覆盖了与理论和实践方面相关的初级、中级和高级主题。

本书共分为9章。第1章介绍计算机视觉和机器学习主题，并介绍与它们高度相关的应用领域。第1章的后半部分提出本书的主题“深度学习”。第2章介绍背景知识，展示流行的手工提取的特征和分类器，这些特征和分类器在过去二十年间仍然在计算机视觉中很受欢迎。其中包括的特征描述符有尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）、方向梯度直方图（Histogram of Oriented Gradients，HOG）、加速健壮特征（Speeded-Up Robust Features，SURF），涵盖的分类器有支持向量机（Support Vector Machine，SVM）和随机决策森林（Random Decision Forest，RDF）等。

第3章描述神经网络，并涵盖与其架构、基本构建块和学习算法相关的初步概念。第4章以此为基础，全面介绍CNN架构。该章介绍各种CNN层，包括基本层（例如，子采样、卷积）以及更高级的层（例如，金字塔池化、空间变换）。第5章全面介绍学习和调整CNN参数的技巧，还提供可视化和理解学习参数的工具。

第6章及其后的内容更侧重于CNN的实践方面。具体来说，第6章介绍目前的CNN架构，它们在许多视觉任务中表现出色。该章还深入分析并讨论它们的相对优缺点。第7章进一步深入探讨CNN在核心视觉问题中的应用。对于每项任务，该章都会讨论一组使用CNN的代表性工作，并介绍其成功的关键因素。第8章介绍深度学习的流行软件库，如Theano、TensorFlow、Caffe和Torch。最后，第9章介绍深度学习的开放性问题和挑战，并简要总结本书内容。

本书的目的不是提供关于CNN在计算机视觉中的应用的文献综述。相反，它简洁地涵盖了关键概念，并提供了当前为解决计算机视觉的实际问题而设计的模型的鸟瞰图。

我们要感谢 Gerard Medioni 和 Sven Dickinson，他们是计算机视觉系列综合图书的编辑，让我们有机会为这个系列做出贡献。非常感谢 Morgan & Claypool 执行编辑 Diane Cerra 的帮助和支持，他负责管理整个的图书准备流程。感谢在我们的职业生涯中遇到的同事、学生、合作者和合著者，他们的激励使我们对这一主题始终兴趣不减。我们也深深感谢各类相关研究团体，他们的工作带来了计算机视觉和机器学习方面的重大进步，本书将介绍其中的一部分。更重要的是，我们想对那些允许我们在本书的某些部分使用他们的数据或表格的人表示感谢。本书极大地受益于同行评审的建设性评论和赞赏，这有助于我们改进所呈现的内容。最后，没有家人的帮助和支持，这项成果是不可能实现的。

我们还要感谢澳大利亚研究理事会（ARC），其资金和支持对本书的一些内容至关重要。

Salman Khan，Hossein Rahmani，
Syed Afaq Ali Shah，
Mohammed Bennamoun
2018 年 1 月

萨尔曼·汗（Salman Khan） 2012 年以特优成绩获得美国国家科学技术大学（NUST）电气工程专业工学学士学位，2016 年获得西澳大利亚大学（UWA）博士学位。他的博士学位论文获得了院长荣誉榜的特别奖励。2015 年，他是澳大利亚国家信息通信技术堪培拉研究实验室的访问研究员。他目前是联邦科学与工业研究组织（CSIRO）Data61 部门的研究科学家，自 2016 年起担任澳大利亚国立大学（ANU）的兼职讲师。他获得了多项著名奖学金，如博士生国际研究生研究奖学金（IPRS）和硕士生富布赖特奖学金。他曾担任多个领先的计算机视觉和机器人会议的程序委员会成员，如 IEEE CVPR、ICCV、ICRA、WACV 和 ACCV。他的研究兴趣包括计算机视觉、模式识别和机器学习。

侯赛因·拉哈马尼（Hossein Rahmani） 2004 年在伊朗伊斯法罕技术大学计算机软件工程专业获得理学学士学位。2010 年在伊朗德黑兰沙希德贝赫什迪大学软件工程专业获得理学硕士学位。他于 2016 年在西澳大利亚大学学习并获得博士学位。他曾在 CVPR、ICCV、ECCV 和 TPAMI 等顶级会议和期刊上发表过多篇论文。他目前是西澳大利亚大学计算机科学与软件工程学院的研究员。他曾担任多个领先的计算机视觉会议和期刊（如 IEEE TPAMI 和 CVPR）的审稿人。他的研究兴趣包括计算机视觉、动作识别、3D 形状分析和机器学习。

赛义德·阿法克·阿里·沙（Syed Afaq Ali Shah） 分别于 2003 年和 2010 年在巴基斯坦白沙瓦工程技术大学（UET）电气工程专业获得理学学士和理学硕士学位。他于 2016 年在西澳大利亚大学计算机视觉和机器学习领域获得了博士学位。他目前在西澳大利亚大学计算机科学与软件工程学院担任副研究员。他获得了由澳大利亚研究理事会资助的 3D 面部分析项目“Start Something Prize for Research Impact through Enterprise”。他曾担任 ACIVS 2017 的程序委员会成员。他的研究兴趣包括深度学习、

计算机视觉和模式识别。

穆罕默德·本纳努恩（Mohammed Bennamoun） 在加拿大金斯敦女王大学获得控制理论领域的硕士学位，在澳大利亚布里斯班昆士兰科技大学获得计算机视觉领域的博士学位。他在女王大学讲授机器人学，之后于1993年加入昆士兰科技大学担任助理讲师。他目前是西澳大利亚大学的温思罗普教授，并曾担任西澳大利亚大学计算机科学与软件工程学院院长五年（2007年2月至2012年3月）。在1998～2002年他曾担任昆士兰科技大学卫星导航空间中心主任。

他曾在2013～2015年期间担任澳大利亚研究理事会（ARC）专家委员会委员。他于2006年在爱丁堡大学担任伊拉斯莫斯学者和客座教授。他还曾担任CNRS（国家科学研究中心）客座教授，2009年法国里尔高等电信工程师学院客座教授，2006年赫尔辛基理工大学客座教授以及2002～2003年法国勃艮第大学和法国巴黎13的客座教授。他是《Object Recognition：Fundamentals and Case Studies》（施普林格出版社，2001）一书的合著者，也是2011年出版的《Ontology Learning and Knowledge Discovery Using the Web》一书的合著者。

穆罕默德已经发表了100多篇期刊论文和250多篇会议论文，并从ARC、政府和其他资助机构获得了极具竞争力的国家拨款。其中一些赠款是与行业合作伙伴（通过ARC联动项目计划）携手解决行业的实际研究问题，这些合作伙伴包括澳洲游泳协会、西澳大利亚体育学院、纺织公司（Beaulieu Pacific）和AAMGeoScan。他致力于研究问题，并与来自不同学科（包括动物生物学、语音处理、生物力学、眼科学、牙科学、语言学、机器人学、摄影测量学和放射学）的研究人员（通过联合出版物、资助和指导博士生）合作。他与澳大利亚境内（如CSIRO）的研究人员以及国际（如德国、法国、芬兰、美国）的研究人员合作。他曾多次获奖，包括1998年昆士兰科技大学年度最佳导师奖、2016年卓越教学奖（指导学生奖）和校长奖（研究导师奖）。2008年他还获得了西澳大利

亚大学指导学生奖。

他曾担任几个国际期刊的特刊的客座编辑，如国际模式识别和人工智能期刊（IJPRAI）。他受邀参加欧洲计算机视觉会议（ECCV）、国际声学语音和信号处理会议（IEEE ICASSP）、IEEE 国际计算机视觉会议（CVPR 2016）、Interspeech（2014）以及国际深度学习暑期学校的课程（DeepLearn2017）。他组织了几次专题会议，包括 IEEE 国际图像处理会议（IEEE ICIP）的专题会议。他是许多会议的程序委员会成员，例如 3D 数字成像和建模（3DIM）以及计算机视觉国际会议。他还为许多地方和国际会议的组织做出了贡献。他感兴趣的领域包括控制理论、机器人技术、避障、目标识别、机器/深度学习、信号/图像处理和计算机视觉（特别是 3D）。

目 录

A Guide to Convolutional Neural Networks for Computer Vision

摘　要

由于计算机视觉在智能监视和监控、健康和医药、体育和娱乐、机器人、无人驾驶飞机和自动驾驶汽车等领域的广泛应用，近年来它变得越来越重要和有效。视觉识别任务(例如图像分类、定位和检测)是许多应用的核心构建块，卷积神经网络(Convolutional Neural Network，CNN)的近期发展使得当前的这些识别任务和系统获得了出色的性能。因此，CNN 现在是计算机视觉的深度学习算法的关键。

这本独立的指南对于既想了解 CNN 背后的理论，又想获得有关 CNN 在计算机视觉中应用的实践经验的人将会有所帮助。本书提供了对 CNN 的全面介绍，从神经网络背后的基本概念开始，依次介绍 CNN 的训练、正则化和优化。本书还讨论了各种各样的损失函数、网络层和流行的 CNN 架构，回顾了评估 CNN 的各种技术，并介绍了计算机视觉中一些常用的 CNN 工具和库。此外，本书描述和讨论了与 CNN 在计算机视觉中的应用有关的研究案例，包括图像分类、目标检测、语义分割、场景理解和图像生成。

本书非常适合本科生和研究生，因为理解本书并不需要该领域的背景知识，也适合有兴趣快速了解 CNN 模型的新晋研究人员、开发人员、工程师和从业人员。

关　键　词

深度学习、计算机视觉、卷积神经网络、感知、反向传播、前馈网络、图像分类、行为识别、目标检测、目标跟踪、视频处理、语义分割、场景理解、3D 处理

第 1 章

A Guide to Convolutional Neural Networks for Computer Vision

简　介

在过去十年间，计算机视觉和机器学习在各种基于图像的应用程序开发中起到了决定性作用，例如，由 Google、Facebook、Microsoft、Snapchat 提供的各种服务。在此期间，基于视觉的技术已经从感知模式转变为可以理解现实世界的智能计算系统。因此，掌握计算机视觉和机器学习(例如，深度学习)知识是许多现代创新企业所需的重要技能，并且在不久的将来可能变得更加重要。

1.1　什么是计算机视觉

人类用眼睛和大脑观察和理解周围的 3D 世界。例如，给定如图 1.1a 所示的图像，人类很容易在图像中看到“猫”，从而实现：对图像进行分类(分类任务)；在图像中定位猫(分类加定位任务，如图 1.1b 所示)；定位并标记图像中存在的所有对象(目标检测任务，如图 1.1c 所示)；分割图像中存在的各个对象(实例分割任务，如图 1.1d 所示)。计算机视觉旨在为计算机提供类似(如果不是更好)能力的科学。更确切地说，计算机视觉寻求开发方法以复制人类视觉系统中最令人惊异的能力之一，即纯粹使用从各种物体反射到眼睛的光来推断 3D 真实世界的特征。

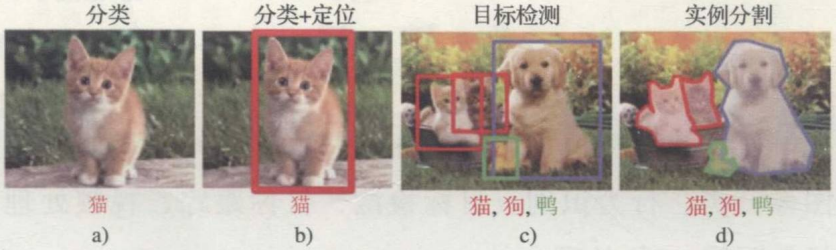


图 1.1　我们希望计算机对图像数据做什么？查看图像并执行分类，分类加定位(即找到图像中主对象(猫)的包围盒并标记它)，定位图像中存在的所有对象(猫，狗，鸭)并标记它们，或者执行语义实例分割，即场景内各个对象的分割(即使它们是相同类型)

然而，从由相机捕获的二维图像中恢复和理解世界的3D结构是一项具有挑战性的任务。计算机视觉的研究人员一直在开发数学技术，以从图像中恢复物体/场景的三维形状和外观。例如，给定一个从各种视图捕获的同一对象的足够大的图像集(见图1.2)，计算机视觉算法使用跨多个视图的密集对应，可以重构出对象的一个精确的稠密三维表面模型。然而，尽管取得了所有这些进步，但是达到与人类一样的图像理解水平仍然具有挑战性。

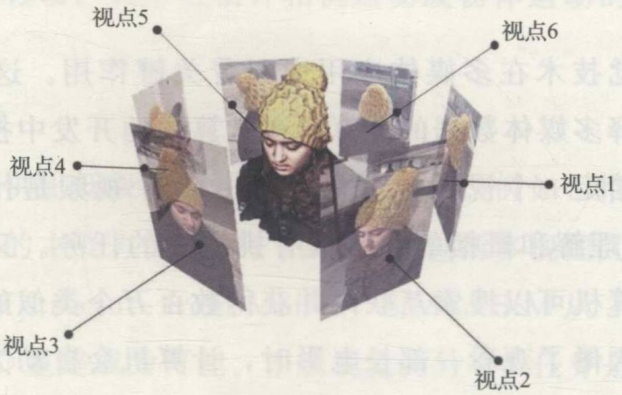


图1.2 给定一组从六个不同视点捕获的对象(例如，人体上半身)的图像，可以使用计算机视觉算法重建对象的密集三维模型

1.1.1 应用案例

由于计算机视觉和视觉传感器技术领域的重大进步，计算机视觉技术如今正在各种各样的现实应用中使用，例如智能人机交互、机器人和多媒体。预计下一代计算机甚至可以与人类同水平地理解人类行为和语言，代表人类执行一些任务，并以智能方式响应人类命令。

1. 人机交互

如今，摄像机广泛用于人机交互和娱乐业。例如，手势可用于手语交流，在嘈杂的环境中传送消息，以及与计算机游戏交互。摄像机提供了一种自然而直观的、人与设备通信的方式。因此，这些相机最重要的一个方面是识别视频中的手势和短暂动作。

2. 机器人

将计算机视觉技术与高性能传感器以及经巧妙设计的硬件集成在一起，产生了新一代机器人，它们可以与人类一起工作，并在不可预测的环境中执行许多不同的任务。例如，一个先进的人形机器人可以以与人类非常相似的方式跳跃、说话、跑步或走楼梯。它还可以识别并与人交互。通常，先进的人形机器人可以执行各种活动，这些活动对人类仅是本能反应，并不需要高智力。

3. 多媒体

计算机视觉技术在多媒体应用中起着关键作用。这导致人们在处理、分析和解释多媒体数据的计算机视觉算法的开发中投入了大量研究工作。例如，给定一个视频，人们会问："这个视频是什么意思?"这是涉及图像/视频理解和概括的非常具有挑战性的任务。又如，给定一段视频剪辑，计算机可以搜索互联网并获得数百万个类似的视频。更有趣的是，当人们厌倦了观看一部长电影时，计算机会自动为他们概述这部电影。

1.1.2 图像处理与计算机视觉

我们可以将图像处理视为计算机视觉的预处理步骤。更确切地说，图像处理的目标是提取基本图像基元，包括边缘和角点、滤波、形态学操作等。这些图像基元通常表示为图像。例如，为了执行语义图像分割(一种计算机视觉任务，见图 1.1)，人们可能需要在该过程中对图像做一些滤波(图像处理任务)。

图像处理主要集中在处理原始图像而不会给出关于这些图像的任何知识反馈，与图像处理不同，计算机视觉产生图像的语义描述。基于输出信息的抽象级别，计算机视觉任务可以分为三个不同的类别，即低级、中级和高级视觉。

1. 低级视觉

基于提取的图像基元，可以在图像/视频上执行低级视觉任务。