WILEY

Interactive Displays:
Natural Human-Interface Technologies

实感交互

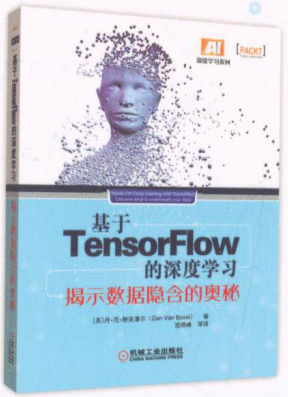
人工智能下的人机交互技术

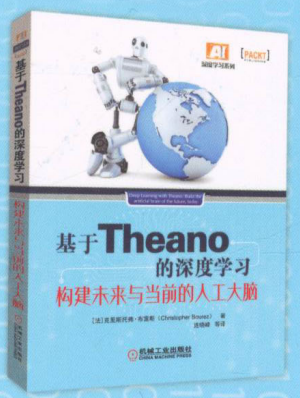
[美] 阿钦蒂亚·K. 鲍米克（Achintya K.Bhowmik） 主编
温秀颖 董冀卿 胡 冰 译

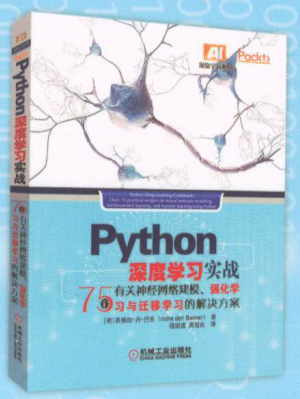
机械工业出版社
CHINA MACHINE PRESS

系列图书如下

基于H2O的机器学习实用方法：
一种强大的可扩展的人工智能和深度学习技术







实感交互：人工智能下的人机交互技术

［美］阿钦蒂亚·K. 鲍米克（Achintya K. Bhowmik） **主编**
温秀颖　董冀卿　胡　冰 **译**
王亚楠 **审校**

机 械 工 业 出 版 社

过往的科幻现已成真，在人工智能时代我们与计算机、手机和娱乐设备的交互正在经历革命性的变化，基于触摸、手势、语音和视觉的自然人机交互正在逐渐替代使用键盘、鼠标和游戏手柄等的交互。显示设备也从单纯的显示设备转变为提供更具吸引力和沉浸式体验的双向交互设备。本书将深入讲解基于触摸、手势、语音和视觉等自然人机交互领域的技术、应用和未来趋势。

本书适合从事人机交互领域工作的研究、设计、开发人员，相关专业师生，以及人工智能时代下对人机交互未来发展趋势有浓厚兴趣的人士阅读。

图书在版编目（CIP）数据

实感交互：人工智能下的人机交互技术/（美）阿钦蒂亚·K. 鲍米克（Achintya K. Bhowmik）主编；温秀颖，董冀卿，胡冰译. —北京：机械工业出版社，2018.3

书名原文：Interactive Displays：Natural Human - Interface Technologies

ISBN 978-7-111-59782-7

Ⅰ. ①实… Ⅱ. ①阿…②温…③董…④胡… Ⅲ. ①人 - 机系统 - 系统设计 Ⅳ. ①TP11

中国版本图书馆 CIP 数据核字（2018）第 087425 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）

策划编辑：林 桢 责任编辑：闾洪庆

责任校对：陈 越 责任印制：孙 炜

北京中兴印刷有限公司印刷

2018 年 6 月第 1 版第 1 次印刷

184mm × 240mm · 20.25 印张 · 472 千字

标准书号：ISBN 978-7-111-59782-7

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

电话服务 网络服务

服务咨询热线：010-88361066 机 工 官 网：www.cmpbook.com

读者购书热线：010-68326294 机 工 官 博：weibo.com/cmp1952

010-88379203 金 书 网：www.golden-book.com

封面无防伪标均为盗版 教育服务网：www.cmpedu.com

原书序

人类对机器的交互式使用可以追溯到数千年以前。世界上第一台自动贩卖机可能是古希腊工程师 Philo 在公元前 220 年发明的。只需投入一枚硬币，它就会把称量好的肥皂传送到盥洗台上。这是一台带有擒纵机构的机械装置。这台颇为先进的机器无疑代表了当时最前沿的科技，但要说它对社会行为的发展产生了显著影响，这是值得商榷的。在往后的 2200 多年里，我们发现自己已经处于明显不同的境地。仅仅 50 年前，第一台电容触摸屏问世了；30 年后这项技术发展成熟，已经广泛应用在笔记本电脑、销售点终端这样的高端设备以及掌上消费品市场；又一个 10 年过去了，如今的触控设备，至少是手持装置已经开始无处不在。

这有力地推动了本书的出版。本书由一组不同领域的高级技术专家联合撰写，涵盖了包括触摸、声音和视觉等主要交互形式。前两类互动形式将分两章分别讨论，视觉部分将在后五章阐释，主要关注已经问世或亟待问世的视觉科技特性。随后两章将介绍用来开发多模态交互显示的多种方法。本书结尾探讨如何呈现最为真实的 3D 立体图像。由于当前显示系统仅能保留高强度数据，而想要获得近似于人眼直接观察到的自然效果，就得通过保留已丢失的相位信息来实现了。

自此，读者会得出结论：本书全面回顾了当前新兴科技的发展。其实不尽然，因为我更想指出的是智能交互技术对社会带来的影响。虽然这些影响主要是积极的，但是也可能存在某些消极方面。这些都是公众关注的重要问题，因而值得辩论。积极的方面包括使用便捷；能够通过电脑或手机系统进行直观推理和预测；用简单的声音指令就能对复杂的结果进行讨论和管理；为身体不便的用户带来便利，并使其充分体验当前的各种产品，等等。然而消极的影响是，比起现在，通信系统会更广泛地侵入用户的生活。通话中的手机不仅早已被怀疑用来追踪定位，其未来系统还会更深入探测我们的行为模式。原则上，依靠计量生物数据来辨别身份的安全系统应当比目前的芯片和个人识别码技术更值得信赖，然而一旦前者的安全系统受损，可能泄露的安全数据将比后者要多得多。

未来人机交互体验的丰富程度，除非当前用户亲身使用，否则他们是难以想象的。过往的科幻现已成真。在技术创造的诸多可能性被供应商和用户采纳之前，有些问题必须讨论和解决。本书为此提供了多方面的素材和依据。对参与该主题的科学技术工作者和开发交互产品的参与者来说，这将是一本重要的书；对于有兴趣了解或需要了解交互技术会如何影响未来社会与人际行为的广大读者来说，本书也不容错过。

Anthony Lowe
于英国 Braishfield

原书前言

什么是“人机交互显示”呢？我们将其定义为不仅能够在屏幕上显示可视信息，还能感知和理解人类行为、接收用户直接输入的显示器。能够“感觉”到手指触摸的触摸屏已经十分普遍，尤其是那些装配在移动设备和一体化计算机上的。现在，新增的类人传感与理解识别技术正在推进新型交互式显示器及系统的开发，使其能够在所处的 3D 空间中“看”“听”且“领会”我们的行为。

我们运用多感官和多模态界面模式来理解周围的客观物理世界，并与人们在日常生活中交流。这些都是通过无缝拼接包括触摸、声音、姿势、面部表情和凝视在内的多种交互模式实现的。我们如果想通过人机交互来获取社交交互的丰富内涵，就必须为这些设备装上能够感知与领悟用户的输入与活动的技术。因此，增加多种自然用户界面能够使人类互动的体验更为真实。

我们与计算机交互的方式经历了最近几十年的变革，依靠鼠标和键盘作为输入工具的图形用户界面已取代传统的基于文本输入的命令式界面。而眼下，随着自然用户界面（通过触摸、姿势、语音等模式的人机交互）的兴起，我们正目睹着下一场技术革命的开始。实施人机界面模式的最终目标就是为用户呈现自然、直观、身临其境般的交互体验。虽然当前的技术局限使得设计师和工程师不得不有所妥协，致使部分目标仅能在完成某个特殊产品时实现，但是为了实现最终目标，我们在近几年来不断取得重大进展。

本书聚焦自然用户界面，对快速兴起的人机交互式显示领域内的技术、应用以及发展趋势进行了深度解读。第 1 章主要介绍人类感知和理解过程的基本要素，回顾了以触摸、声音和视觉感应推理为基础的自然界面技术，以及通过该技术实现的人机交互过程；随后各章深入每种输入与交互模态的细节，在实现多感官和多模态交互的目标过程中，对技术的基本原理及其在多种用户界面模式中的结合与应用展开细致的探讨；最后一章总结了基本要求和技术发展现状，展望了未来有望实现的“真实的”3D 交互界面及其带来的真实的、沉浸式的交互体验。

我向编辑 Anthony Lowe 致谢，是他发现了著书探讨交互式显示的必要性。我感谢对本书做出贡献的企业界和学术界专家，感谢 Wiley 出版社的员工对本书的支持。最后，谨以此书献给 Shida、Rohan 和 Ava，没有你们的鼓励和支持我无法开展并完成这个项目。

Achintya K. Bhowmik
于美国加利福尼亚州

目 录

第1章

交互式显示的感知、理解与自然人机界面

Achintya K. Bhowmik
美国英特尔集团

1.1 引言

如今，可视化显示设备已成为丰富多彩的电子产品中不可或缺的一部分。作为人与电脑、通信系统和娱乐系统交互的主要界面，其应用已经融入居家、工作或出行等生活的方方面面。无论是腕上的手表，还是随身装在口袋或钱包里的手机，抑或是用来网上冲浪、获取多媒体信息的平板电脑，再或者是工作的笔记本电脑或台式电脑，还有客厅中心的巨屏电视、商务会议使用的演示投影仪，可视化显示器都是这些设备面向我们用户的“颜面”。

这类显示器频繁应用于各种特定的公共场所，比如机场自助登机手续办理终端，零售店自助付款机、大型购物商场的广告牌以及博物馆的公共展示——用途不计其数。近十年来，巨大的应用潜力和市场需求促进了全球可视化显示技术的研发。从移动显示到巨屏显示，多样化的产品层出不穷[1-5]。

只要扫一眼可视化显示设备的市场规模，我们就能很快领会它给生活带来的影响。来自显示产业分析公司 IHS 的报告说明，近五年来，销往世界各地的平面显示设备总额高达 170 亿美元[6]，年度出货量超过 50% 的增速也说明了这一技术的快速普及率。

总体来说，一台电子设备主要完成三项基本功能：接受用户指示，按照指示及所获信息执行某些处理功能，呈现输出或向用户报告处理结果。比如，当作者在笔记本电脑上进行本章的写作时，他首先用键盘和鼠标输入信息，然后微处理器就会执行文字处理软件，将敲击键盘和点击鼠标发出的命令转换成目标文本和格式，最后，电脑的液晶显示屏就会以可视化的输出实时显示文字。由此可见，设备里的显示子系统已经在向用户呈现信息方面发挥了至关重要的作用。除了某些特例之外，大多数近期生产的电子产品都配备了显示屏幕，唯一的目的就是为了显示视觉信息。

然而近几年来，人机互动和用户界面范式一直在经历着快速的演变和创新。我们与电脑

交流的方式经过几十年的变革已经大不相同。在文本型的老式命令输入界面被淘汰以后，取而代之的是依靠鼠标和键盘输入的图形用户界面。随着更多自然用户界面的出现，下一场变革的帷幕正在我们的眼前拉开。未来，人机交流不但可以通过触摸、肢体动作、声音、表情和视线来实现，甚至还可以通过我们的思想！

我们正在不断研发高级传感器、系统、运算规则以及应用程序，以实现更为生动自然的互动体验。在这个过程中，运算装置除了能够把握交流意图之外，还能理解用户的表达与情感。这些兴起的界面技术和接踵而至的新型应用产品为显示技术乃至整个电子消费产业创造了振奋人心的机遇。随着自然用户界面的不断整合，显示设备也从以往视觉内容的单向显示转变成了可以接收用户输入的双向互动，这就推动了交互应用程序的开发和沉浸式体验的实现。触摸屏和触控优化界面以及各类应用产品的激增又把这场变革蔓延到了移动显示设备，自然界面技术由于其交互性的强化而不断延展，必然会重新定义整个显示技术和显示系统的维度。

本书全面解析了促使高度交互显示与显示系统兴起的自然人机界面技术与应用。那么什么是“人机交互式显示”呢？我们将其定义为不仅可以在屏幕上显示可视信息，还可以感知和理解人类行为并接收用户的直接输入。一旦装配上类似自然人的感知和理解技术，一个“真实”的人机交互式显示器就能“感受”并探测到我们的触摸，“听到”并回应我们的声音，“看到”并辨识出我们的面貌和表情，“理解”并阐释通过移动手指或其他身体部位发出的肢体指令，甚至能够根据语境推理出我们的意图。

虽然这些目标看起来非常远大，但是正如图 1.1 所示，依靠简单直观的自然人机界面，多种形态因素和应用系统加之自然用户交互技术已经对市场带来了巨大的影响。本书的讨论也在不断揭示这种影响，我们在自然感知、推理技术、系统整合和应用发展方面取得的重大进步将为人机交互的全面创新打下坚实的基础。

图 1.1　各种形态的交互显示器与应用系统已经占据了大片市场，如前面例子所述。除了传统意义上对用户显示视觉信息之外，许多系统内的显示器在直接人机界面设备中发挥着新的作用

图 1.2 描述了交互显示系统的通用功能模块及其流程。用户和显示系统的互动是受各个界面发出的指令支配的，也就是在开始和结束部分显示的输入和输出模块。输入模块由一组

传感器组成，能够把用户输入的物理刺激转换成电子信号。而输出模块则以物理刺激的形式，让用户感知并理解系统反向回应用户的行为。中间的模块处理必要的信号并执行运算功能以促进交流。

本章首先综述了人类感知和理解的基本原则，特别关注了我们在与物理世界的日常互动中部署的机制和流程。以此为基础，我们随后概述了运用自然界面技术（包括触摸、声音、视觉感知和互动）的人机互动过程，并简要梳理了史上最为成功的界面技术。接下来，我们将深入到每类输入与互动的模态细节，对技术原理及其在自然人机界面模式的应用，以及综合互动技术在实现直观的多感观、多模态互动方面的作用进行深入的探讨。本书最后一章总结了基本要求和技术发展现状，展望了未来有望实现的“真实”的 3D 交互式显示及其带来的真实的、沉浸式的互动体验。

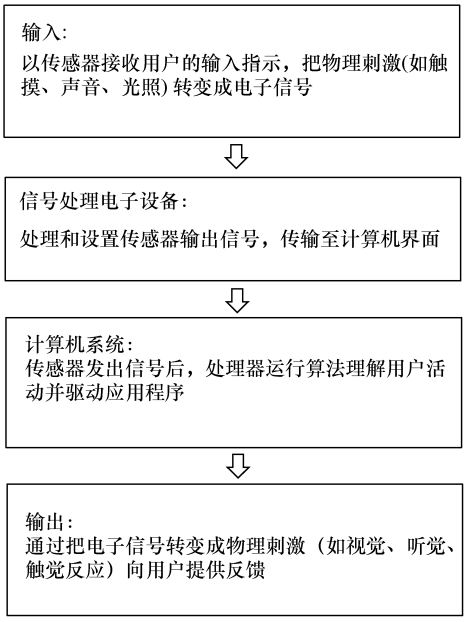


图 1.2　交互显示系统的功能框图。输入模块与输出模块指挥用户与显示器的互动，而信号处理与运算法则促进了这些互动的开展

1.2　人类感知和理解

本书伊始就提出实施人机交互界面方案的最终目标是让用户获得自然、本真和沉浸式的互动体验。虽然目前技术的局限性让设计师和工程师不得不做出妥协，仅能实现某些特定产品的部分目标，但是我们一直在实现总体目标的方向上取得进步。

这里需要进一步阐明一下。所谓“自然”，意思在于运用我们的自然机能与机器实现交流和互动。我们运用多感官、多模态的界面方案来理解周围环境和相互交流，将包括声音、表情、凝视、手势和肢体语言、触觉、嗅觉和味觉等在内的多模态互动无缝衔接。如此，创建自然界面就能使真实的生活体验融入人机互动之中。

所谓“本真”，意指该界面依靠我们多年养成的社交习惯而设计，仅要求用户使用最少的（理想是不需要任何）学习成本就能与机器进行交流。

所谓“沉浸式”，是一种真实世界与虚拟世界边界模糊化的体验，其中电脑或机器成为我们身体与大脑的延续，帮助我们完成任务。这是个很高的要求，需要几十年的持续研发才能接近这些目标。我们努力了解生动逼真的人机界面和交互方案，就能使我们以史为镜，了解人类——毕竟我们是“人机互动”这个词组的第一个字！

我们人类已经进化成了高等交际物种，受助于一个精干的大脑和一系列复杂的感知器官，包括丰富的视觉感知系统、听觉能力、接触敏感的皮肤和触觉感知，还要算上经过鼻腔

和舌头传感的气味和味道的化学感知。超过一半的人类大脑致力于处理感知信号，让我们能够认识太空、生命和周围的物体，也让我们在自然、本真的感知情境中彼此互动。

让我们深入探讨一下我们的感知传感和推理过程，即眼睛和视觉感知过程，耳朵和听觉感知过程，皮肤和触觉感知过程。仅仅专注于这三种感知模态的一个原因是我们与物理世界交互的实质过程主要运用到这些机制，而且我们也将看到，这些机制的功能能够依靠高新技术在电子设备中加以模仿，以便设计和制造高级互动显示器和系统。在人机交互中实现嗅觉和味觉机能当然最好，不过还得等技术进一步发展。

让我们从神经生理学角度探讨自然人机界面与交互显示系统，如图 1.1 所示。这个交互过程可以分解为三个主要过程：感知，理解和辨识，以及行为。从人的视角看，感知过程包括：搜集显示器视觉产出——通过光波介入人眼；说话人听觉产出——以声波形式介入人耳；感觉屏幕的表面——通过用指尖碰触。这些感知传感器将物理刺激通过传导过程转换成神经信号，后被传递到大脑皮层，也就是我们能够理解到“看”“听”和“触”的发生，随后辨识与思考相继启动。

根据感知和辨识过程的结果，我们将指令我们的身体行为。比如，我们把视线聚焦到显示器上想关注的元素上，指引手指触摸并启动屏幕上的具体内容，调整我们对声音产出的听觉注意力，摆出一个合适的面部表情，甚至用我们的手指和手来做一个动作。

我们首先综述一下视觉感知过程。我们仅关注与随后讨论密切相关的操作交互显示器的内容，并把其他更为详细介绍人类感知[7,8]的读物介绍给有兴趣的读者。人眼是人类进化的奇迹，特别体现在其构造上的极端复杂性，功能的有效性及其在连接感知世界与大脑枕叶视觉皮层方面所发挥的核心作用。如图 1.3 所示，人眼和相机的某些核心结构十分相似，都是

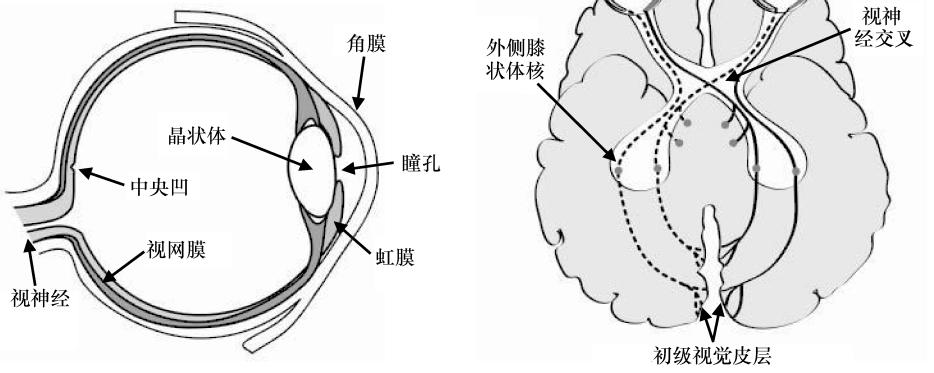


图 1.3　左：人眼解剖图。右：人的视觉系统使用双目成像方式。左视野是由双目的右侧感知到的，并映射到视觉皮层内的主要接收区域的右半部，右视野则经由另一条路线。观测物与双眼的距离是由双目视差察觉的，其他的视觉线索还包括动态视差、视差映射和焦距等

通过透镜系统把外景光源聚焦在眼部后方的视网膜上成像；视网膜周围含有称为感光器的感光细胞。眼部有两种类型的感光体，即有色觉的视锥细胞和无色觉的视杆细胞，后者能把光转换为神经信号。

这台相机的分辨率以及和处理器通信的带宽如何呢？视网膜包含大量的感光器——大约每只眼睛有 800 万个视锥细胞和 12000 万个视杆细胞——然而视觉体系却能够巧妙地发出景物在空间和时间上变化的信号，而不是由感光器探测到的绝对光强，以保持眼睛和大脑的通信带宽降到实际水平上。

当我们把目光投向一个物体且图像形成于视轴周围的一个相对较小的区域时，中心视觉的视敏度是最高的。这是因为视锥感光器最集中地分布于视网膜内的一个小区域——中央凹，这些感光器映射到视觉皮层内的一个比视网膜其他部分要大的区域。另一个相机的重要特质是光敏的动态范围，人眼的视觉跨径可达 10 个数量级，远远超过了现代数码相机的能力。

每只眼睛都是一部优秀的相机，像这样的相机我们拥有两部。人类的视觉系统包括 3D 和深度理解能力，有着双目成像方式以及其他诸如动态视差、视差映射和焦距等视觉线索，这些能让我们在 3D 空间内十分轻松地找到方向并于各种物像交互。双目成像已经普遍演化成大多数生物系统的特征。近期的化石研究论证其早在 5 亿多年前节肢动物生活的早寒武纪时代就已经存在[9]。强大的视觉系统的出现被认为是引发寒武纪大爆炸变革的导火线[10]。部分重叠的横向位移视野导致了“双目视差”，也就是由单眼捕捉到物体相对于另一只眼睛发生了横向位移。我们随后将会了解到，双目视差与观测物到观察人的距离成反比。

有这样的视觉系统帮助理解距离，猎物就更容易发现逼近的猎人而逃生，猎人也有更好的时机三角测距猎物的位置并实施捕猎。双目视觉因此被推定为生物进化成功的推动力，也是最早的哺乳动物的特质之一。时至现代，我们运用我们复杂的双目视觉系统来与 3D 世界互动。图 1.3 也简化地展示了将眼睛连接到视觉皮质的感觉传导路径。

接着，我们来思考一下听力感知的重要元素，包括耳朵和各个听辨过程。恰如眼睛，人的耳朵也有着精致的构造以及像声音传感器这样令人惊叹的功能。我们天然的麦克风——耳朵——能够感知超过 12 个数量级的声音强度以及 3 个数量级的音频（20 ~ 20000Hz）！

如图 1.4 所示，耳廓决定了气流携带声音信号进入含有耳鼓膜的耳道的方向。压力振荡经由中耳组织——锤耳、砧骨和镫骨得以放大，这些部位是人体拥有的最小骨头，英文中可分别用意为锤子（hammer）、铁砧（anvil）和马镫（stirrup）的单词表示，暗指它们是如何放大并向内耳部分传递声音信号的。最后，振荡声波被转经由神经冲动转换成神经信号，更具体地说是由位于呈收敛螺旋状的耳廓部位的听毛细胞转换的。这些神经信号随后发射到位于颞叶的大脑听觉皮层并被处理成能够感知的信号。

正如人眼一样，我们还有一对能在频率信号之外启动双声道感知方案的天然“麦克风”，它可以在 3D 空间内准确定位声音的来源。双耳 3D 感知以及极高的声压灵敏度对我们的进化过程十分重要，在日常生活中，它也对帮助我们在 3D 物理世界的穿梭和交流起到了不可或缺的作用。图 1.4 简单展现了位于人耳与大脑听觉皮层之间的神经分布路径。

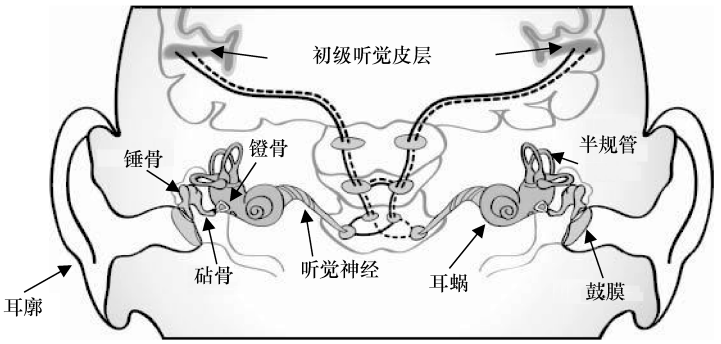
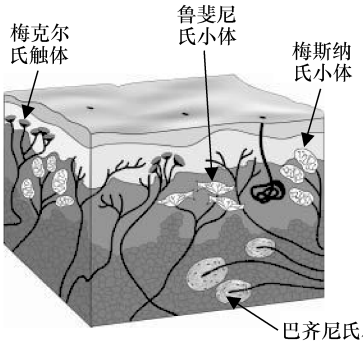


图 1.4　人耳和双声道构造的解剖图，简化地描绘了位于内耳耳蜗和大脑颞叶内的听觉皮层之间的神经分布路径。双耳信号以及频率信号被用来定位声音信号的来源

最后，我们再看看触敏性和触觉感知过程。触觉的感知过程又称皮肤感知，开始于皮肤内的机械性感受器，它们能够在相应的皮肤区域感受到因接触而产生的机械压力。图 1.5 描绘了 4 种主要的机械性感受器。视觉（眼睛）和听觉（耳朵）感知器官位于颅骨内，具有离大脑皮层相对较短的神经生理路径，而触觉感知器官（皮肤）却覆盖了整个身体。因此，来自触觉接收器的信号常常需要经过较长的距离（比如从手指到头部）。脊髓对触觉感受器来说就起到了“信息高速公路”的作用，把从接收器获得的信号传递到顶叶内的大脑体觉皮层——这部分大脑位于处理触觉过程的头部顶端区域。

神经外科医生 Wilder Penfield 在 20 世纪 50 年代关于触觉敏感的重大发现已经证明了人体邻近部位对大脑皮层邻近区域的映射[11]。更有意思的是，这项映射研究确立了作用于身体各个部分的大脑体觉皮层的相对比例。图 1.5 所示的“皮层矮人”（cortical homunculus）



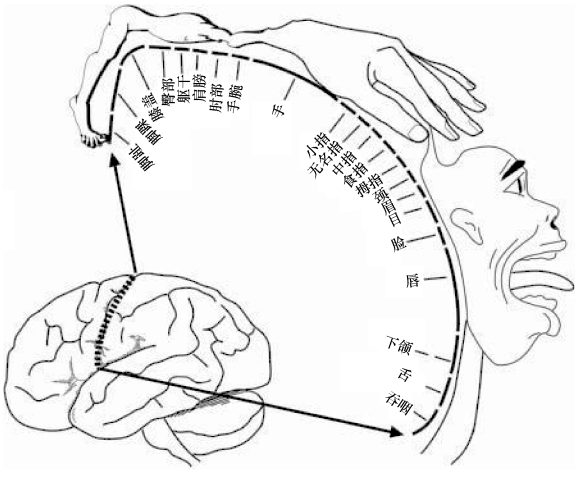


图 1.5　左：人体皮肤的解剖图。所示的主要四种类型的机械性感受器将触碰刺激产生的机械压力转变成神经信号。右：矮人皮层理论，由 Wilder Penfield 首次提出，后续学者陆续完善，揭示了体觉皮层处理来自身体各个部分的触觉信号的位置和相对比例。来源：http://www.intropsych.com/ch02_human_nervous_system/homunculus.html，转载获得 R. Dewey 的许可

的概念就是该理论的集中体现。别错以为这只是幅随意的讽刺漫画，其实这个矮人图呈现了一个人体各部分所占大脑体觉皮层的相对空间的比例模型。如图所示，该皮层组织致力于处理来自手指碰触的信号远超过其处理来自整个手臂和手腕，这恰恰证明了触摸屏用户界面设计师期望大量运用手指来实现触控式人机交互的合理性！

正如前面讨论过的，神经生理学对我们的感知过程有一个普遍的解释。感知系统的设计十分巧妙，绝大部分的大脑皮层组织与感知接收器最重要的部分是相连的。比如，视网膜中央凹与中心视力，耳蜗听毛细胞与听觉，手指尖与触觉等。虽然我们也拥有其他感觉机制，但是在与周围物理世界的交互中，我们更主要依靠的是看、听和碰触。因此本书主要关注眼睛、耳朵和触感作为自然人与显示器设备交流的主要模态。

相比起生物系统，当今大多数的计算和娱乐设备具有非常初级的感知和处理能力。就手机、平板电脑和笔记本电脑来看，它们是典型的“单眼”工作（仅有一个相机），就像希腊神话中的独眼巨人库克罗普斯一样。此外，它们大多数是单耳结构（仅有一个麦克风），还有许多尚未实现触敏（触摸屏），尤其是笔记本电脑。

但随着技术在多方面的迅猛发展，这一情况将有望在不远的未来得到改善。向自然和人类世界学习，工程师和设计师现在已经开始对计算和通信设备加入“类人”的感触和感知属性，让它们能够“看”“听”和“理解”人类行为和指示，并发挥这些功能以促进自然的、本真的互动。这些发展保证了人机交互实现超越键盘、鼠标、操纵杆和远程遥控的突破，并允许基于碰触、视觉与言语感知和识别技术的自然交互的使用。

尽管现实中我们每时每刻感知和洞察周围世界是那样的自然和随意，但是只有我们尝试在机器中实施这些感知功能的时候才能理解这些任务的复杂性。在下一节，我们将综述人机界面与电子设备的重要技术，包括最近几十年广泛采用的技术先例以及新近实现的与显示器和系统交互的自然本真模态。

1.3　人机界面技术

1.3.1　过往的输入装置

在深入讨论最新自然界面技术和由其推动的应用与用户体验之前，很有必要回顾一下历史，思考一下最成功的用户输出技术的发展。最近几十年间，随着我们在生活中的实践和接触，该技术已经成为人机交互技术的核心支柱。我们无意对人机界面技术及其相关的历史发展进行完整记述，想要在有限的章节内完成这项任务也不可能。我们现在能做的是把过去出版的许多综述文献介绍给有兴趣的读者[13－15]。

以下我们简述几项已被大众采纳的创新发明和主流产品，它们定义了时至当代的人机交互的主要方式。回首过往，我们感谢它们取得成功的重要因素——简易的技术应用，以最适度的价格使用最优的现有技术元素，尤其是应用某项发明来满足用户对丰富个人生活和各项活动的需求。

首先，无处不在的遥控装置可以说定义了我们与电视荧屏的交互关系并塑造了我们的内容浏览行为。尽管远程遥控的概念早在 1898 年就由 Nikola Tesla[16] 提出，第一个电视机遥控器却是由 Zenith 无线电公司于 1950 年开发并投入市场的，并形象地将遥控器命名为“懒骨头”[17]。

电视机自 20 世纪 20 年代起就已经风靡市场，但那时人们需要走到它跟前来调整控制旋钮，尽管自然的观看姿势就是坐在它面前的沙发上。因此，当时的环境对发明远程遥控来说是成熟的；需求很明确，而且技术水准也已经到位。Zenith 的“懒骨头”手持遥控器与电视机之间是有长线相连的。虽然这解决了人们的合理需求，使人们不离开沙发就能换频道，但避免不小心被电线绊倒的需求还是指向了无线遥控的发明。

到了 1955 年，也是由 Zenith 生产的“闪光助手”（Flash – matic）问世。通过光束指向分布在电视机屏幕四角的感应器，用户就能使用这款遥控器实现无线控制。这种激动的心情可以从宣传当日的杂志广告中体会：“不得不让你眼见为实!”虽然兴奋难抑，但是这款光控设备并不能在光亮的房间内很好地发挥作用，因为外界的光线会偶尔改变设置。Zenith 把下一代的设备更换成了超声波作为远程通信媒介，并命名其为“太空司令”，这才解决了问题。一则 1957 年的广告（见图 1.6）振振有词地宣扬了这款“坐享舒适沙发，无声遥控电视”的神奇体验。

图 1.6　1957 年的一则 Zenith“太空司令”遥控装置的广告。来源：www. tvhistory. tv（已取得引用许可）

此后的现代遥控技术发展更是日新月异，产品不仅融入了各种时尚精巧的形状元素，还安装了红外光以遥控娱乐装置。近几年还不断新添了动作感应和声音控制技术等特征。

接下来，我们再回到 Douglas Engelbart 发明电脑鼠标的 1963 年，这是标志着人机交互新纪元的开始之年。在发明和装配鼠标之前，早期电脑输入局限于基于文本的键盘敲击指令。图 1.7 展现了第一个由 Engelbart 和 Bill English 构建的鼠标原型。鼠标由两个在互为直角方向上滚动的滑轮组成，随着鼠标在平面拖动，两个滑轮能跟踪鼠标在 2D 平面上的位置[18]。Engelbart 在 1961 年设想到该装置时正在参加一个电脑作图会议，思考着如何能构建一个能与电脑绘图对象简易高效互动的系统[19]。

值得注意的是，鼠标仅仅是 Engelbart 和他的斯坦福研究学院团队发明的众多电脑输入

设备之一，不过它却是最成功的一项。虽然现在的版本以镭射光替代了滚轮，以无线传输替代了传统数据线，但我们还是习惯地称之为“鼠标”，因为最初命名时想到它连接电脑的线与老鼠的尾巴一样。鼠标连同图形用户界面使用户操作计算系统更便捷，这使其在近几十年与快速推广的个人电脑一样变得无处不在。

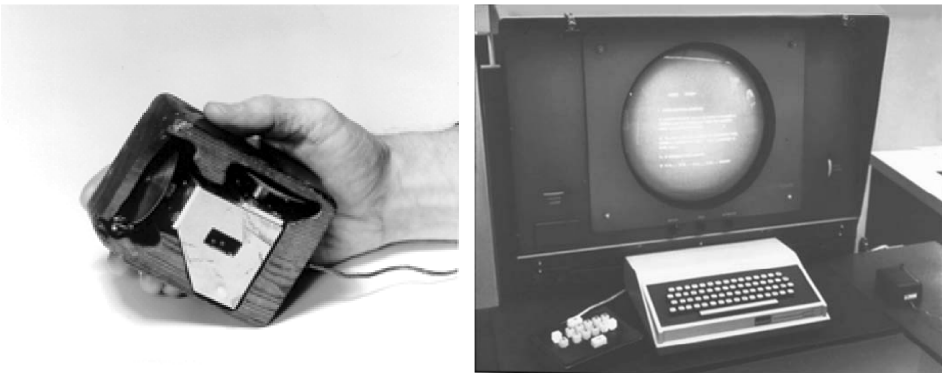


图 1.7　左：第一只由 Douglas Engelbart 和 Bill English 于 1963 年开发的电脑鼠标。右：为实现和用户互动而设计的一台配备有原始鼠标的电脑工作站。来源：SRI 国际（已取得引用许可）

这些早期人机界面设备的“遗产”对其相应的主控系统的发展影响巨大。随着电视机成为全世界家庭娱乐的核心设备，个人电脑成为提高生产力和获取信息的首要工具，遥控器和鼠标也随之成为了我们必不可少的伴侣。但是，虽然它们在近几十年内使用广泛，人机界面和互动的格局仍然十分有限，是时候要展望未来了。接下来我们会谈论到，最近在新传感器技术、推理演算法、计算资源以及系统整合等领域的发展让我们感受到了通过自然人机界面与电子装置和系统互动的可能。现在我们就来看看用户基于触碰、声音、视觉和多模态交互技术实现的自然界面输入。

1.3.2　触控式交互技术

显示器从仅向用户输出可视信息到成为一种交互界面装置主要归因于触控功能与显示器的一体化模式，尤其是其在移动通信装置上的使用。从 1965 年第一份由 Johnson 撰写的电容触摸屏报告[20,21]至今，该技术及其应用已经经历了几十年的发展，并成为了全球主流消费品。

“触摸屏”由 Johnson 发明，是一台由电容覆盖的阴极射线管显示器，如图 1.8 所示。Johnson 是一名英格兰皇家雷达研究所（Royal Radar Establishment）的工程师，该项技术主要用于航空交通控制系统。他所做的论文摘要介绍到：“该‘触摸屏’装置提供了高效的人机联

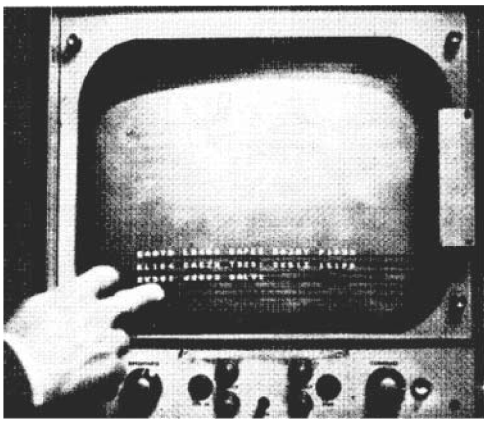


图 1.8　第一款运用电容技术感应的“触控显示器”的存照，由 E. A. Johnson 于 1965 年报告。来源：www. history – computer. com ，转载已经获得许可

结。”尽管这番陈述只是针对自己的发明，他还是深刻地预见到了这些设备几十年后的未来，即将被大众广泛使用。

虽然鼠标和遥控装置风靡多年，但使用这些设备与显示屏互动仍然是一个“间接”的操纵体验。另一方面，触敏显示的引进能让人们仅通过触摸就能“直接”与屏幕内容互动，让人们不用或仅用很少的训练就能有更为方便和本真的体验。近些年来，由于触摸屏手机、平板电脑、超级本、多合一桌面电脑以及各种形式的信息资讯站的普及，触摸屏技术及其商业化部署一直在迅速发展。事实上，触摸屏技术和触控便捷软件界面的无缝衔接已经催生了新一批高级交互应用设备，这使得用户的使用体验发生了巨大的变化。

有很多不同的技术方法能够探测对显示设备的触碰[22]。在第 2 章，Walker 对各类交互显示触控技术进行了深度的综述。下一章我们会详细阐述，显示屏表面感触方位的方法可以主要分为电容式、电阻式、声学和光学技术。

运用电容技术方面，图像显示屏的表面或内部有一个用来存储电荷的夹层。其中一种应用叫交互电容法，是指用户在碰触显示屏任意位置时，有一部分电荷转移到了用户身上，导致了原来位置的电荷存量减少。另一种应用叫自电容法，指人体部位碰触显示屏时增加了相对于地面的单一电极电容。触控夹层中的用来侦测这些变化的电路能够识别碰触部位并向软件操作系统、应用程序及用户界面提供该信息。

电阻式触控方面，当用户通过触碰屏幕上的某个位置而施加一定的机械力时，两层间隔距离较小的光透传导表面受压后逐渐靠近。该位置的坐标数据由电压测量值决定，并传输给软件进行处理。

声学和光学技术分别包括测量由于用户触屏而产生的超声波和红外光波的变化值。具体的系统实现大不相同，全世界很多公司都在开发研制这类产品。

本章涵盖的具体技术包括投射电容、模拟电阻、表面电容、表面声波、红外线、摄像光学、内嵌式整合、弯曲波、压力传感、平面散射探测、视觉传感、电磁共振和这些技术的综合。Walker 论述了这些技术的运行原则、关联系统结构和整合方法、各方法的优缺点、历史发展、产业动态，以及上述触控技术的未来趋势，包括对显示屏内不同层级触控功能集成的阐述。如今市场上已经在显示模块运用了触摸屏的设备随处可见，但近期问世的商业产品还是证明了无需独立触摸屏接入的触敏集成显示面板更能够减少设备的厚度、重量、集成复杂性和成本。第 2 章就详细介绍了这样的“嵌入式触控”技术。

在交互显示和系统中引入触碰输入模式给市场带来了深远的影响。让我们快速浏览一下市场规模，品味一下触摸屏技术留下的痕迹：整个产业每年产出超过 10 亿件触摸屏产品。尽管这些大多数是移动装置，触摸屏技术已经广泛地应用到各种形态的设备中。主流显示器安装触摸屏输入功能不过是时间的问题而已，特别是那些需要与用户实现近距离互动的设备。

1.3.3　声控交互

可以说人与人之间最有效也是最普遍的交互形式是有声语言。要了解这一点，只需要做一个“思考实验”。假想你是一个特立独行的世界探险家，突然发现自己不但无法理解所到

之处的人际交谈，自己的话语也让别人感到不知所云！有声语言交流始终都是现代人类文明发展和社会交融的根本动力。有证据显示，学术界和企业界均对使用声音输入、处理和输出的人机界面的发展有着浓厚的兴趣并付出了巨大的努力[23]。

虽然我们可以毫不费力地表达和理解他人的话语（大多数情况下），但是让一台计算机具备人类拥有的对有声语言的理解能力绝非易事——我们为了这个目标已经奋斗了一个世纪。扫一眼图 1.9 就可以迅速了解到这个挑战。图中特别展现了发声短语“mining a year of speech（挖掘一年的语音数据）”的语言波形记录。我们在说话的时候会不均匀地断句或使用短间隔，这会生成一连串没有间隔的听觉信号，或者我们根本就察觉不到声音中的间隔在哪！我们也经常在对话中使用一些不完整的句子，把并无意义的词语安插在句段之间，为断开的意群“搭上桥”。概括地说，语音识别算法的任务就是要把有声话语转换成一系列文本，并摘取该文本表达的含义。声控交互界面发挥了这些功能来在用户和设备间构建一个声音互动方案。

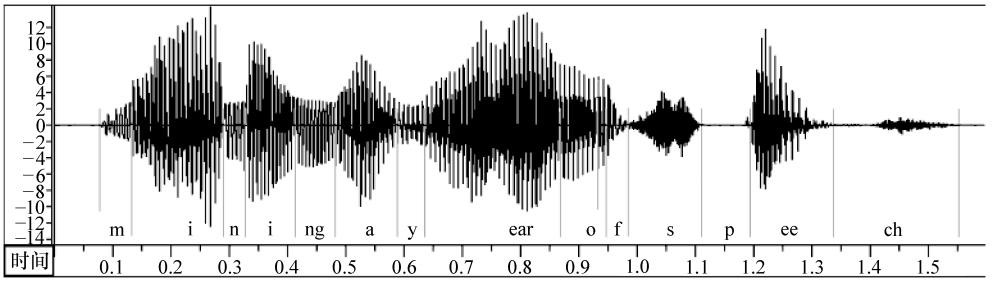


图 1.9　发声短语“mining a year of speech”的声波波形，指出了在话语和非直观声波信号间隙中非均匀分布的短时空间。横轴标注的是以 s 为单位的时间，纵轴显示的是任意单位的信号强度。来源：http：//www. phon. ox. ac. uk/mining_ speech/，转载获得 John Coleman 的许可

自从 20 世纪二三十年代由 Harvey Fletcher 和 Homer Dudley 在贝尔实验室进行的人类语音建模与合成的开创性尝试开始[24,25]，自动语音识别研究一直在过去的几十年间稳步发展，特别是在 20 世纪 80 年代语音建模的统计算法的创立，以及近期在自然语言理解方面所取得的进步尤为引人注目。在 1968 年问世的史诗科幻电影《太空漫游》（A Space Odyssey）中，编剧 Stanley Kubrick 和 Arthur C. Clarke 预见的 HAL9000——一台将在 20 世纪 90 年代诞生的电脑——可以流畅自如地和人类进行语音对话。虽然我们尚未实现 HAL 所具备的所有神奇功能，最近在声音界面和交互领域的突破还是创造了很大的商业价值，并催生出越来越多的在移动装置、计算机工作站和汽车中使用的应用程序。

在多数情况下，计算机装置的声控界面会生成简单自然的人机互动。比如，一句简单的语言命令“播放［歌名］这首歌”就能让装置迅速地从服务器存储的许多歌曲中挑选出来并开始播放。同样的，一道命令“把这张照片放到我的脸谱网（Facebook）主页上”就能够马上上传用户使用智能手机拍摄到的画面，或是从之前存储好的相册中选出来。“播放昨晚保存的温布尔登网球赛”则能在媒体存储器中找到相应的网球比赛，并开始在电视机荧

屏上播放。说一句“指给我去 SFO 的方向”就能显示出前往圣弗朗西斯科国际机场的路线图以及驾驶方向。

用传统的界面完成这些任务需要浏览大量的命令窗口、输入文本以及敲击许多按键。然而通过语音命令完成相同的任务则会从根本上变得更简洁、更迅速，只要该设备能够按要求准确地理解并处理在真实场景中使用的语音命令和指示。

拥有自然语言理解和语音合成功能的语音识别技术保证了电脑、通信、娱乐和许多其他电子设备以及系统的大规模推广。当我们的双手和双眼忙于类似烹饪、驾驶、购物、园艺和锻炼等事务时，不妨使用语音识别来操作相关设备。它还有可能使许多残障人士能够进行电脑操作。

下面将会阐述，声控交互在与其他交互途径联用时功效特别强大，比如手势或凝视追踪。未来的计算机将无孔不入，那时的感知和推理技术将全方位融入我们的生活——衣食住行、工作娱乐等方方面面，而基于有声语言的互动将起到至关重要的作用。目前，从交互显示的观点来说，声音界面似乎可以使我们更简单地与各种形态的显示器交互，从而获得更为本真的体验。

第 3 章中，Breen 等人深度综述了声控用户界面的基本原理和发展。几位学者就语音界面的重要元素展开讨论，包括语音识别、自然和对话语言理解技术、对话管理、语音合成、高效语音处理的硬件及系统结构优化，以及应用众多的交互设备和系统的程序等。

1.3.4　视控交互

我们在 1.2 节已经讨论过，视觉感知，更确切地说是目测和理解 3D 环境的能力，是一种能够使我们在物理世界中畅行、与他人交流的必备素质。2D 相机和成像应用现在已经是计算和娱乐设备中必不可少的组成部分，特别是在手机、平板电脑和笔记本电脑中，该技术还越来越多地应用在一体化的桌面电脑和高端巨屏电视机中。

目前，集成在手机里的 2D 相机的主要应用是拍摄数码静止照片和录像，而那些在大型设备和显示器里的相机则主要用于视频会议应用。电脑视觉研究人员已经开发了能够探测、追踪和识别面部和表情、理解动作和简单手势的 2D 图像处理算法[26-29]。

传统 2D 相机拍下 3D 世界的影像并将其投射在 2D 平面图中，舍弃了许多置身 3D 空间的视觉信息细节。

科学家已经花费了巨大的科研精力研究如何把单一的 2D 图像复原成 3D 信息的过程，以更好地理解人类动作。从 2D 投射中重构 3D 空间信息是一个有着内在歧解的病态问题，即便对架起一个已知的结构（如人体）来说也是一个挑战，很多有前景的研究结果只是非常有限地使用在了实践中[30-32]。这些方法总的来说需要电脑的密切配合和人工输入，因此对需要实时独立分析 3D 环境和人体动作的交互应用程序来说并不适合。

相比之下，人类视觉系统的 3D 成像工艺流程可以捕捉并使用 3D 视觉信息，推进高效稳健的认知和互动。增加实时 3D 视觉传感功能可以实现真正交互式的、理解用户的系统显示和丰富的自然用户交互。这些功能包括在显示器前使用实时 3D 图像传感技术来拍摄 3D

景象；在 3D 空间内用电脑视觉算法来理解 3D 图像和实时用户活动；调试用户界面，使其能够本能地执行人类任务、指示智能系统和回应命令。

视控姿势识别是全世界正在兴起的一个研究和开发领域，学术界和业界的实验报告都反映了该领域快速发展的技术，揭示了基于人类动作行为研究的多层次交互过程的分类和实践发展[29,33-35]。第 4 章是对视控交互方法的综述，包括 3D 传感和肢体动作识别技术，说明了在人机交互应用中使用这些技术的现状和对未来的展望。

基于 3D 传感装置的系统和应用已经在市场上出现，较传统 2D 成像技术，它们为用户带来了更为丰富和稳健的互动体验[36,37]。这些初期的市场成功有力地推动了 3D 视觉技术在未来更多设备系统中的使用，也使得 3D 用户交互更为普及。实时 3D 图像技术在电子设备中的应用实现了显示器前微观用户交互和 3D 空间内的目标操纵。

实现 3D 实时传感的方法各式各样，总的来说都是要输出一个除了彩色图像之外的等深图，使成像的 3D 物体和景象得以重建。其中三个最为突出的方法是，结构光 3D 传感技术、立体 3D 成像和飞行时间法范围成像技术[37]。第 5～7 章将深入到每个具体的 3D 成像方法，为 3D 交互应用的使用打下基础。

运用上述技术实时获取 3D 视觉信息，我们就能通过 3D 图像识别推理技术实现的非触屏互动来启动丰富的人机交互方案。图 1.10 显示了一些在显示屏前依靠 3D 手势而获取的自然体验，而并非使用传统的 2D 输入技术，如鼠标或触摸屏[37]。左图显示了这样一个场景：用户希望伸手“抓住”门把手，“转动”，然后从显示平面中“拉拽”以“打开”那扇门。右图展示了一个“弹弓”应用程序：用户用手指“拉伸”弹力绳，“瞄准”3D 空间中的目标，并“释放”弹力绳，以击中目标并打破 3D 结构的元素。这些动作与使用鼠标、键盘乃至触摸屏都有很明显的不同，后者并非用户的本真体验。但是，使用实时的 3D 图像捕捉以及 3D 电脑视觉算法来实现 3D 手势交互可以产生更为自然和本真的用户体验。

图 1.10　交互应用和基于实时 3D 传感推理技术的体验，包括在显示器前的 3D 空间内操纵物体[37]

除了 3D 空间内的手势互动和物体操纵以外，实时 3D 成像还能变革照相方法、视频会议、远程协作和录像博客等应用程序。比如，通过使用 3D 成像装置生成的等深图，用户可以更轻易准确地从图像中被分离出来，然后从背景中抽出或放入另一个定制的背景中。图 1.11 呈现了这个技术。

虽然图像处理技术可以用在传统的 2D 图像上来达成这种效果，但 3D 传感设备能使分隔更为清晰，还能使实时应用程序使用 3D 景象信息。比如，人们可以通过视频会议程序在

图 1.11　使用深度传感成像设备的 3D 分隔技术能让你轻易地操控背景。在这组图片中，左边的男孩出现在原始的背景前，右边的他则出现在另一个经过处理后的不同背景前。要注意的是，通过分析右图的非连续的明暗底纹可以发现右边的背景并非原始背景。深度传感成像设备通过使用 3D 景象信息能够实现实时分隔，可以用在需要常规背景的视频会议或博客的应用程序中。来源：www. cambridgeincolour. com，获得 Sean McHugh 的许可

家里舒适地参加商务会议，但是在屏幕上显示的却是参会人在自己办公室的背景！

另一个能够显著改善的应用类别是增强现实程序，即把 3D 图像内容加至捕捉的图像序列中。不同于使用 2D 相机的传统增强现实程序，3D 成像可以用 3D 物体和反映真实视觉的景物模型来增强影像内容，并使用户能够与增强现实的元素进行交互。想象一下能够让你虚拟地站在装有 3D 成像设备的交互显示器前试穿衣服或试戴首饰的应用，或是选择合适的家具来虚拟地装饰你的房间。

除了追踪和识别手势和肢体动作之外，在侦测凝视方向和确定用户在显示器上的视线方位方面，3D 科技也有了重大的发展和突破。目光凝视在人际交往方面发挥了显著的作用。凝视是注意力的重要体现指标。图 1.12 就显示了某个人在观赏一幅画时的兴趣点分布。

图 1.12　以凝视方向标注的视觉注意力示例。左：呈现给观众的图像；右：图像上的兴趣点分布。来源：cambridgeincolour. com，获得 Sean McHugh 的许可

神经生理学研究已经显示了凝视在与物理世界进行持续交流方面的重要性[38,39]。尽管眼睛的主要功能是捕捉景物的视觉信息—— 作为部分视觉感知过程，但我们在交流的时候同样也把凝视和语音、手势进行紧密协同。举一个例子，当你说“请给我那个红球”并注

视椅子上的那个红球的时候，看着你的人就会明确地意识到你并不是要那个此时放在地上的红球。这个人只需要简单地跟随你双目凝视的方向就能理解你的意思，即使你并未用手指指向那个在椅子上的球。

研究人员长久以来致力于把强大的交互机制并入含计算系统的用户界面，特别是和其他相关的交互模态一起。比如，我们只需瞧一眼笔记本电脑上的图标，说上一句“打开它”或者“启动”，无须伸手触摸荧屏或使用鼠标对准点击就能将文件打开，甚至还可以在自由空间内打一个手势。在第 8 章，Drewes 详细综述了凝视追踪技术、系统及其应用，包括当前人际交互方案中凝视追踪的局限性和应对这些挑战的可能途径。

1.3.5　多模态交互

人类感知和交互常常是多模态的——我们使用所有的感官，结合由其生成的神经信号来理解周围物理世界并与之交互。比如，我们用双耳声频信号和频率提示来定位声音的来源，随后用眼内的聚合和调节系统把双目视线指向该声源，并把物体反射出的光线聚焦于我们的视网膜上，以实现视听同步。同样在其他的场合中，我们的听觉感知也可能跟踪视觉感知并使其增强。例如在逛公园的时候，我们也许先看到一只鸟，然后注意到它的叫声。在真实环境中，我们运用多模态互动相互交互。根据意图和情境，我们用碰触、手势、声音、眼神、面部表情和感情的集合来本能地与人类同胞交流。

1976 年，McGurk 和 MacDonald 发表原创论文并形象地命名其为《听唇看音》（Hearing Lips and Seeing Voices）。文中他们叙述了偶然发现的视觉和听觉的互动，也就是后人称为的“麦格克效应”[40]。该研究显示，当我们听到说话人发出的声音伴随着和其他不同的声音一致的视觉信号时（相当于配音过程），会导致我们感知到另一种声音的存在。我们感知过程中的视听一体的情形在表演腹语口技时也非常明显，同样的效果还体现在剧院，我们产生了演员在屏幕上说话的幻觉，其实不过是装置在场所其他方位的扬声器发出声音。神经生理学证据已经显示，当我们使用多重感官系统来理解周围的环境时，来自一个感知传感器的神经信号可以促进、覆盖或修改来自另一个传感器的信号。不同的传感区域在大脑中互相作用，为连接脑内视觉、听觉和触觉的接收区域提供了实验依据[41]。

因此，自然、本真的人机交互方案必须是多模态的。结合语音识别与位置感知的早期研究结果在 Bolt 于 1980 年发表的论文中有所记录。他指出了人机自然交谈的可行性，比如“放在那里”“变成一颗蓝色的大钻石”“称……作日历”等等[42]。Quek 写道：“为了让人机交互能够达到人际交流的透明水平，我们必须明白对话互动的现象学和其他能够帮助我们理解的可抽取的种种特征。”作者还论述了使用语音和手势作为交际的共同表达形式[34]。

第 9 章里，LaViola 等人评述了人机交互的多模态感知界面，探索了合并多种输入模态以构建自然交流的可能性。该章研究了主导交互类型，各层次多模态集合的可用性，以及调试这些模态的途径以期达到逼真的自然交互。解决多模态界面方案的人为因素问题往往决定了内置多模态交互功能的新设备、新系统能否取得商业上的成功。除了之前章节提到的输入模态（如触摸、手势、语音、凝视和面部表情）之外，本章还发起了关于通过脑电图学和

肌电图学来侦测肌肉活动的讨论，以期实现整合新兴的人机界面技术。

科幻小说作者一直在幻想着一个人能用脑电波控制电脑、机器和系统的未来世界，在那里人们只需要“心想”就能“事成”！尽管那样的未来还尚未实现，但是最近在人脑界面技术的发展已经显示了人们具有通过思考生成大脑信号来控制和操纵显示内容的能力。该领域的研究一直在持续，力争可以创造出前所未有的交互方案和应用，以进一步丰富未来交互显示系统[43]。LaViola 等人在第 9 章讨论了这种在多模态交互方案内的人机界面整合。

除了与屏幕内容进行多模态交互，在面控和声控用户识别方面的突破也有望用自然的多模态生物计量验证取代原有的密码身份验证。在日常的社交生活中，我们使用面部、声音和基于自然人辨识方案的行为特征来建构与我们交流的人群的身份。然而，电脑识别其用户的能力却仍然很大程度上限制于密码或口令牌。随着计算系统的普及与不断融入我们的社会生活，这种认证方式将不再充分适用。

Poh 等人在第 10 章综述了多模态生物计量，探讨了包括技术设计和可用性的问题以及该领域的近期发展。作为另一个多模态感知的范例，我们常常在相互交流的时候使用面部表情的线索来理解口头话语。同样的字，以不同的面部表达方式道出可能会指代完全不同的事物。面部表情可以通过具体的脸部姿态下意识地补充某种交流需要，或是自然而然地显露某种内心的感觉和情绪。其他观察者对说话人的面部表情的揣测往往取决于当时的语境[44]。

150 多年以前，Duchenne 以研究肌肉运动如何产生多种面部表情为目的对受试人进行了实验。图 1.13 是他的研究成果的一个例子，表现了通过电导探针诱导脸部肌肉收缩而产生的一系列面部表情。这是使用了新发明的相机设备记录下来的[45]。近几十年来，数码相机、高级图像处理技术和计算机资源的普及使学者有机会对自然化的面部表情开展研究。就在最近，3D 传感和处理技术越来越多地用于更为高级的自动化面部表情识别。关于视控表情识别技术的发展在第 4 章探讨视觉传感和肢体动作交互的部分将会提及。

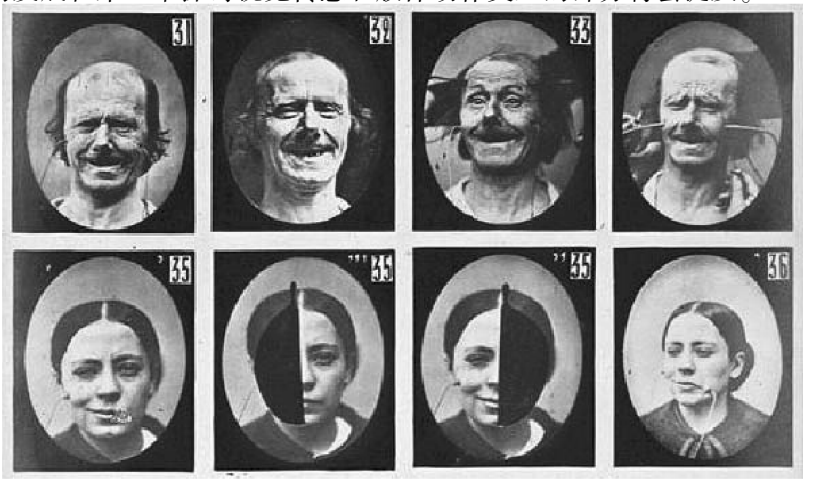


图 1.13　Duchenne 的原创作品，出版于 1862 年。介绍了通过应用电子探针激活肌肉收缩，从而诱导不同的面部表情。上排的图片显示了脸两边相同的表情，下排的图片则显示了左右两边脸的不同的表情。改编自 Duchenne[45]。来源：转载获得 www. zspace. com 许可

1.4　“真实”3D 交互显示探索

虽然图像显示器已经无处不在，并且成为了我们生活中不可或缺的一部分，但是目前绝大部分显示器主要显示的是单眼视觉信息（2D），而无法重构通过人类感官获得的 3 D 真实世界的重要视觉信号。所以，近年来，3D 立体显示技术开始获得市场关注。目前 3D 商用显示器的主要关注点一直在于通过启动我们视觉体系中的双眼合像来提供实体视觉线索。该过程中，不同的视觉图像呈现到用户的左右眼以获取深度感知。许多书籍都介绍了各种重构 2D 和 3D 图像的显示技术的运行原理[1-5]。

我们最终的目标是构建“真实”3D 交互显示系统，为用户提供栩栩如生和身临其境的视觉和交互体验。这样的显示系统的发展需要更为仔细的研究，包括考察人类视觉感知系统和重构与视觉线索一致的信号流程，这样，我们才能利用传感技术来感知我们日常生活中的 3D 世界。那么，我们应该如何以我们的视觉和感知处理系统来理解 3D 技术呢？除了立体观测，我们对现实世界的 3D 感知利用了一些重要的 3D 视觉线索。这些线索包括：①运动视差效应，即当我们在移动的时候，总是感觉离我们近的目标相较于离我们远的目标运动速度更快；②聚合效应，即眼球会向内侧或外侧转动以聚焦在一个离得近或远的物体；③调节效应，眼部晶状体的形状会因为聚焦某个物体而自行调节；④遮挡效应，即较近的目标部分遮挡了较远的目标；⑤线性透视效应，平行线会在视野上的远点汇聚；⑥纹理梯度效应，间隔均匀的目标从远处观察会显得更密集；⑦与目标的 3D 位置和照明环境一致的投影；⑧以及其他来源于我们已有的知识线索，比如熟悉的大小和朦胧的环境等。这些重要的 3D 视觉线索有助于我们的 3D 感知，如图 1. 14 的所示。

已经证明，在显示器上实施的运动视差效应：投射在视网膜上的图像与观众的头眼移动变化一致。这为用户提供了除了立体观测外更逼真的视觉体验。该产品的一个例子是来自 zSpace 的一个显示器，如图 1. 15 所示。该 3D 显示系统通过红外相机传感器来跟踪用户的头部移动，并根据用户的特定位置创建立体图像对，从而提供实时运动视差的视觉线索[46]。系统还包括一个手写笔来操控 3D 空间的虚拟物体。

传统 3D 立体显示也受到不一致的焦点线索的影响，这是由于双眼聚合的目标和晶状体调节双眼聚焦射入光线的不匹配而造成的，该冲突是导致人类视觉疲劳的原因[47]，最近有人提出了通过使用电调节镜片来应对这一问题[48]。

在本章前面的部分，我们已经讨论了触控传感器和相关用户界面的增加（特别是在移动显示器上）正逐步将传统显示器变为双向沟通的交互设备。我们也观察到，除了 2D 平面显示器有限的触摸输入，最近 3D 成像和行为识别技术的进展越来越多地允许用户在显示器前实现与系统的 3D 交互。我们预计，这两个领域发展的结合将会构造一个点对点的 3D 用户交互界面系统，并同时可以显示 3D 视觉内容、理解用户的输入。

显然，使用 2D 用户输入方案（如触碰或点击鼠标）以操控 3D 显示器上显示的内容无法实现自然或本真的用户体验，此时使用 3D 交互方案可能更合适。例如，研究用户主观体

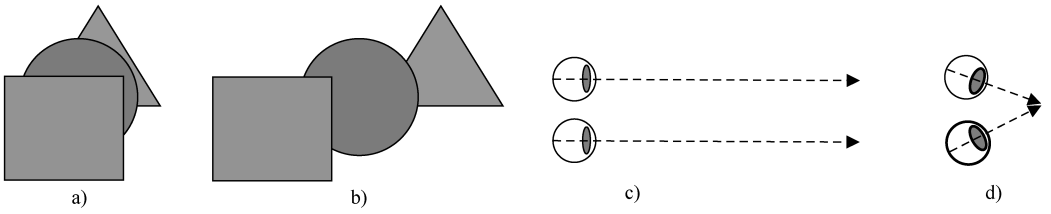


图 1.14　突出的视觉线索的描述有助于丰富我们对周围环境的 3D 感知和把握未来“真实”3D 显示器生产的方向：需要为用户提供身临其境的 3D 视觉体验。上方图内的叠加图形描述了实体视觉线索的双眼差异：a）是用户左眼看到的图像，b）是右眼所看到的图像。它还展示了遮挡线索，它作为一个单一视图足以暗示正方形更接近观察人，而三角形则更远。这也解释了运动视差效应：随着眼睛位置在视野中从左向右移动，正方形在视野中左移的距离比圆形左移的距离要更远，因为圆形距离观察人更远。下方图示则解释了聚合线索和调节线索：c）展示了当看到一个遥远的物体时，眼睛的光学轴是几乎互相平行的；d）展示了当双眼聚焦于一个近处的物体时，晶状体的形状会调整，以使图像聚焦在视网膜上。除此之外，还有其他的 3D 视觉线索将会在正文中解释

图 1.15　由 zSpace 生产的结合了运动视差效果和立体观测的交互式显示器图示。系统跟踪用户的头部运动并向用户展示了根据用户所处的位置而创建的立体图像对。手写笔被用于与显示器的虚拟对象进行实时交互。来源：www. zspace. com，转载获得许可

验的数据表明，用平面触摸方式在 3D 立体显示器上进行 3D 视觉交互存在重大问题[49]，而用户更趋向于用手势与 3D 虚拟对象进行交互[50]。为了直接操纵 3D 显示器上的内容，研发直接的 3D 交互方案引起了人们越来越多的兴趣[51-54]，不过实现这一方案的实践应用仍然需要进一步的发展。

未来“真实”的 3D 交互显示将需要呈现动态的 3D 视觉内容，为用户提供一致的立体观测、视差和焦点线索，实现除了获取单眼 3D 线索外的深度感知；同时研发 3D 传感和推理技术以实现与 3D 空间内重构的物体进行沉浸式交互。第 11 章是对实现这一目标的需求

和进程的深入分析。这一章首先详细列出了利用光场原则和人类视觉感知的基础知识重建“真实”3D 视觉信息的内容。然后综述了能够提供所有重要视觉线索和逼真 3D 感知的“真实”3D 图像显示器的技术发展。最后，我们提出了集成人机交互/3D 视觉内容和系统的建议，包括人为因素问题和潜在的解决方案。

1.5　结语

可视信息显示设备现在已经无处不在了。它们是所有类型的电脑运算、通信、娱乐和其他电子设备系统的表情。近几十年来，科学技术的突飞猛进为实现高质量的可视化效果打下了基础，大批各种型号的、轻薄的、低电耗且价格合理的显示设备已经发挥了绝妙的视觉功效。消费者快速地接受各种形态的可视装置导致商业出货规模猛增，这些产品从可穿戴装置、手持智能手机到平板电脑、笔记本电脑，再到巨屏电视机和信息咨询站等。如今，显示设备正在从单向视觉信息迈向双向交互发展的新纪元。

人机交互方案同样也在经历变革，传统的键盘和鼠标界面正在被更为直接、自然的触碰、声音或手势取代或改善。受益于触敏技术的快速发展，手机用户获得了前所未有的新界面、新应用的使用体验。内置实时 3D 图像捕捉技术和推理算法的 3D 可视电脑有望通过开启 3D 空间内的各类人机互动得到进一步发展。此外，在语音界面、凝视侦测、脑机界面方面的研究已经取得了重大进步。基于多感官感知方案的多模态互动及其集成的各式输入途径有望让人们的互动体验充满真实的精彩。

本书聚焦自然、沉浸式的用户界面这一话题，对当下蓬勃发展的互动显示领域进行了综述，包括领域内的技术、应用和发展趋势。本章概述了与交互显示意义和发展有关的感知与理解过程，并审视了自然人机界面技术。就好像几十年前发明的鼠标和图形用户界面催生了无数新的电脑应用，还有近几年间由触碰界面的普及带来的其他诸如智能手机和平板电脑等新的应用一样，基于 3D 多模态感知和推理技术的自然、本真用户界面也必将引发新一轮的交互应用热潮。显示系统的未来是交互的，而这样的未来已经开启！

参 考 文 献

1. Bhowmik, A.K., Bos, P.J., Li, Z. (Eds.) (2008). *Mobile Displays: Technology & Applications*. John Wiley & Sons, Ltd.
2. Lee, J.H., Liu, D.N., Wu, S.T. (2008). *Introduction to Flat Panel Displays*. John Wiley & Sons, Ltd.
3. Brennesholtz, M.S., Stupp, E.H. (2008). *Projection Displays*. John Wiley & Sons, Ltd.
4. Tsujimura, T. (2012). *OLED Displays*. John Wiley & Sons, Ltd.
5. Lueder, E. (2012). *3D Displays*. John Wiley & Sons, Ltd.
6. IHS Displays Report Portfolio, www.ihs.com 2009–2013.
7. Goldstein, EB. (2013). *Sensation and Perception*. Cengage Learning.
8. Snowden, R., Thompson P., Troscianko T. (2006). *Basic Vision: An Introduction to Visual Perception*. Oxford University Press.
9. Lee, M., Jago, J., García-Bellido, D.C., Edgecombe, G.D., Gehling, J.G., Paterson, J.R. (2011). Modern optics in

exceptionally preserved eyes of Early Cambrian arthropods from Australia. *Nature* **474**, 631–634.
10. Parker, A. (2011). On the origin of optics. *Optics & Laser Technology* **43**, 323–329.
11. Penfield, W., Rasmussen, T. (1950). *The Cerebral Cortex of Man: A Clinical Study of Localization of Function*. MacMillan.
12. Dewey, R. *Psychology: An Introduction*. www.intropsych.com.
13. Jacko, J.A. (Ed) (2012). *Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. CRC Press.
14. Milner, N.P. (1988). A review of human performance and preferences with different input devices to computer systems. In: Jones DM, Winder R. (Eds). *Proceedings of the Fourth Conference of the British Computer Society on People and computers IV*, pp. 341–362. Cambridge University Press.
15. Buxton, W. *Human Input to Computer Systems: Theories, Techniques and Technology*. http://www.billbuxton.com/inputManuscript.html.
16. Tesla, N. (1898). *Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles*. US Patent 613,809.
17. A Brief History of the Remote Control. (1999). *DigiPoints: The Digital Knowledge Handbook* **3**, 4.
18. English, W.K., Engelbart, D.C., Berman, M.L. (1967). Display-Selection Techniques for Text Manipulation. *IEEE Transactions on Human Factors in Electronics* HFE-8, 5–15.
19. *Father of the Mouse*. http://dougengelbart.org/firsts/mouse.html.
20. Johnson, E.A. (1965). Touch Display – A novel input/output device for computers. *Electronics Letters* **1**, 219–220.
21. Johnson, E.A. (1967). Touch Displays: A Programmed Man-Machine Interface. *Ergonomics* **10**, 271–277.
22. Bhalla, M., Bhalla, A. (2010). Comparative study of various touchscreen technologies. *International Journal of Computer Applications* **6**(8), 975–8887.
23. Pieraccini, R., Rabiner L. (2012). *The Voice in the Machine: Building Computers That Understand Speech.* The MIT Press.
24. Fletcher, H. (1922). The Nature of Speech and its Interpretations. *Bell Systems Technology Journal* **1**, 129–144.
25. Dudley, H. (1939). The Vocoder. *Bell Labs Record* **17**, 122–126.
26. Yang, M., Kriegman, D., Ahuja, N. (2002). Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(1), 34–58.
27. Tolba, A, El-Baz, A, El-Harby, A. (2006). Face Recognition: A Literature Review. *International Journal of Signal Processing* **2**(2), 88–103.
28. Fasel, B., Luttin, J. (2003). Automatic Facial Expression Analysis: a survey. *Pattern Recognition* **36**(1), 259–275.
29. Mitra, S., Acharya, T. (2007). Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews* **37**, 311–324.
30. Lee, H., Chen, Z. (1985). Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing* **30**, 148–168.
31. Guan, P., Weiss, A., Bălan, A., Black, M. (2009). Estimating Human Shape and Pose from a Single Image. *Int. Conf. on Computer Vision* 1381–1388.
32. Ramakrishna, V., Kanade, T., Sheikh, Y. (2012). Reconstructing 3D human pose from 2D image landmarks. *Proceedings of the European Conference on Computer Vision, Part IV* 573–586.
33. Pavlovic, V., Sharma, R., Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence* **19**(7), 677–695.
34. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., McCullough, K.E., Ansari, R. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* **9**, 171–193.
35. Wexelblat, A. (1995). An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction* **2**, 179–200.
36. Han, J., Shao, L., Xu, D., Shotton, J. (2013). Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* **43**(5), 1318–1334.
37. Bhowmik, A. (2013). Natural and Intuitive User Interfaces with Perceptual Computing Technologies. *Inf. Display* **29**, 6.
38. Pelphrey, K.A., Morris, J.P., McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain* **128**, 1038–1048.
39. Klin, A. Jones, W., Schultz, R., Volkmar F. (2003). The enactive mind, or from actions to cognition: Lessons from autism. *Philosophical Transactions of the Royal Society of London B* **358**, 345–360.
40. McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, 5588.
41. Murray, M., Spierer, L. (2011). Multisensory integration: What you see is where you hear. *Current Biology* **21**,

R229–R231.
42. Bolt, R. (1980). Put-That-There: Voice and Gesture at the Graphics Interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* 262–270.
43. Tan, D.S., Nijholt, A. (2010). *Brain-Computer Interfaces and Human-Computer Interaction*. Springer-Verlag.
44. Carroll, J., Russell, J. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology* **70**, 205–218.
45. Duchenne, G. (1862). The Mechanism of Human Physiognomy (Mecanisme de la physionomie Humaine).
46. Flynn, M., Tu, J. (2013). Stereoscopic Display System with Tracking and Integrated Motion Parallax. *International Display Workshops* **3**, D1–3.
47. Hoffman, D., Girshick, A., Akeley, K., Banks, M. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* **8**(3), 1–30.
48. Bos, P.J., Bhowmik, A.K. (2011). Liquid-Crystal Technology Advances toward Future True 3-D Flat-Panel Displays. *Inf. Display* **27**, 6.
49. Pölönen, M., Järvenpää, T., Salmimaa, M. (2012). Interaction with an autostereoscopic touch screen: effect of occlusion on subjective experiences when pointing to targets in planes of different depths. *Journal of the Society for Information Display* **20**(8), 456–464.
50. Grossman, T., Wigdor, D., Balakrishnan, R. (2004). Multi-finger gestural interaction with 3d volumetric displays. *Proceedings of the 17th annual ACM symposium on User interface software and technology,* 61–70.
51. Alpaslan, Z., Sawchuk, E. (2004). Three-dimensional interaction with autostereoscopic displays. *Proceedings of SPIE Vol. 5291A, Stereoscopic Displays and Virtual Reality Systems XI* 227–236.
52. Blundell, B. (2011). *3D Displays and Spatial Interaction: Exploring the Science, Art, Evolution and Use of 3D Technologies*. Walker & Wood Ltd.
53. Bruder, G., Steinicke, F., Stuerzlinger, W. (2013). Effects of Visual Conflicts on 3D Selection Task Performance in Stereoscopic Display Environments. *Proceedings of IEEE Symposium on 3D User Interfaces 3DUI*. IEEE Press.
54. Zhang, J., Xu, X., Liu, J., Li, L., Wang, Q. (2013). Three-dimensional interaction and autostereoscopic display system using gesture recognition. *Journal of the Society for Information Display* **21**(5), 203–208.

第2章

触觉感知

Geoff Walker
美国英特尔集团

2.1 引言

本章试图为应用于人机交互界面的触控技术提供一个明确的定义。本章的目的在于让读者对 18 种不同的触控技术的操作、功能、应用、优缺点、局限性等方面有深入广泛的认识。这对用户了解如何与机器互动很有帮助，因为随着触控与其他输入模态的不断结合，用户的选择也日趋广泛，如第 1 章和第 9 章所述。

本章讨论的范围仅限于接触显示屏的触控技术，不包括笔尖输入和手指“悬空”输入这两种与显示屏有 1cm 距离的输入方式。非透明表面（非显示屏）接触、近距离感应以及手势（3D）输入也不在本章讨论范围内。本章同样不涉及触摸屏生产制造方面的具体内容。

在触控技术和集成系统的一系列命题中，我们主要探讨各类技术的特点而非专注于某一项技术，因此在内容上我们注重广度而不追求深度。在本章（乃至整个触控产业）中，“触摸屏”和“触控面板”是同义词，前者多用于西方国家，后者则在亚洲比较常见。两种说法都指向同一种包含了由触控传感器、触控控制器和计算机界面构成的触控模块。

本章将全部触控技术划分为六个大类，每个大类又依次划分为若干小类（用圆括号表示，一共有 18 种）如下：电容式触控技术（2 种）、电阻式触控技术（3 种）、声学触控技术（3 种）、光学触控技术（5 种）、嵌入式触控技术（4 种）及其他（1 种）。文中“嵌入式”指的是触控功能已在制造过程中被显示器制造者完全集成到显示器中，与此相对应的“分离式”，指的是触控功能被触摸屏制造者添加到显示器中的技术。

触控产业具有高度的保密性，在该领域至少 200 个以上的公司，甚至包括一些大公司都是私有企业。由此产生的结果就是很少有触觉传感技术发明者、开发者或者供应商发表论文或出版图书，这也是本章有别于其他章之处。本章的参考资料范围特别广泛，包括网络、杂志、时事通讯、白皮书、专利权、会议演讲材料、新闻稿和用户指南，甚至包括博客文章。

也正是因为缺乏有关触觉传感技术的学术论文和图书，本章中追溯触觉传感技术历史的部分更侧重于其商业化的时间而不是研发时间。

2.2 触控技术简介

从 CRT 显示屏到 OLED 显示屏，显示屏很早就被用作输出设备。只是最近因为将触觉感知功能外加或者集成到显示屏技术的兴起，显示屏才被大量作为交互式的输入设备而使用。在 1965 年，Johnson 第一次以书面的形式记录了电容式触摸屏的使用之后[1]，大约过了 30 年后触摸屏才充分广泛地应用于商家使用的产品（即卖方应用领域）中，例如销售终端和机场的登机系统[2]。触摸屏第一次广泛而明确地应用在消费者产品（即买方应用领域）中是 20 世纪 90 年代中期开发的电子记事簿。第一台掌上电子记事簿是 1993 年苹果公司出品的 Newton 电子记事簿，紧接着在 1997 年又出现了更为有名的 Jeff Hawkins 掌上电脑。

最终导致当下“触摸无处不在”的浪潮的大事件则是 2007 年苹果公司 iPhone 手机的发明。苹果公司创新性地启用了一项之前默默无闻但是使用起来异常简单的触控技术（投射电容式触控技术），让用户在使用手机时有一种身临其境的体验，从而点燃了人们使用触摸屏的热情，并使这种热情持续攀升（见图 2.1 和图 2.2）[4]。苹果公司这一发明也彻底改变了触摸屏产业的格局，传统的占主导地位的模拟电阻式技术很快被发展迅速的投射电容式技术（p－cap）所取代（见图 2.3）[4]。

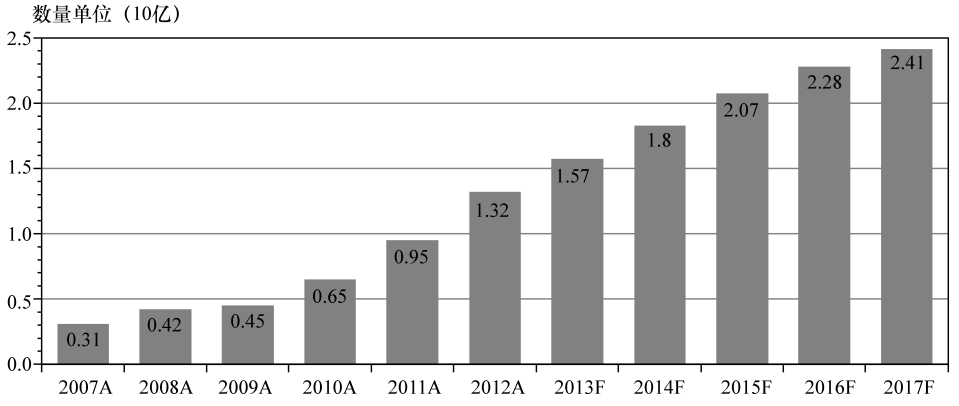


图 2.1 触控模块的出货数量图（单位：10 亿），2007～2012 年的数据为实际数，2013～2017 年的数据为估计数。数据来源参考文献［4］

2009 年 7 月，微软公司 Windows 7 系统投放市场，标志着一体式（AiO）家庭桌面电脑初现雏形。第二年，苹果公司推出 iPad（2010 年 4 月），这是第一部百分之百触摸操作的消费电子产品（所有的平板电脑都具有触摸功能，但是并不是所有的手机都有触摸功能）。微软公司 Windows 8 在 2012 年的 8 月上市，标志着 Windows 系统从桌面操作系统（OS）到“触摸优先”操作系统的转变。而本书成稿时，该转变产生的影响依旧存在于整个个人电脑

和触摸屏产业。

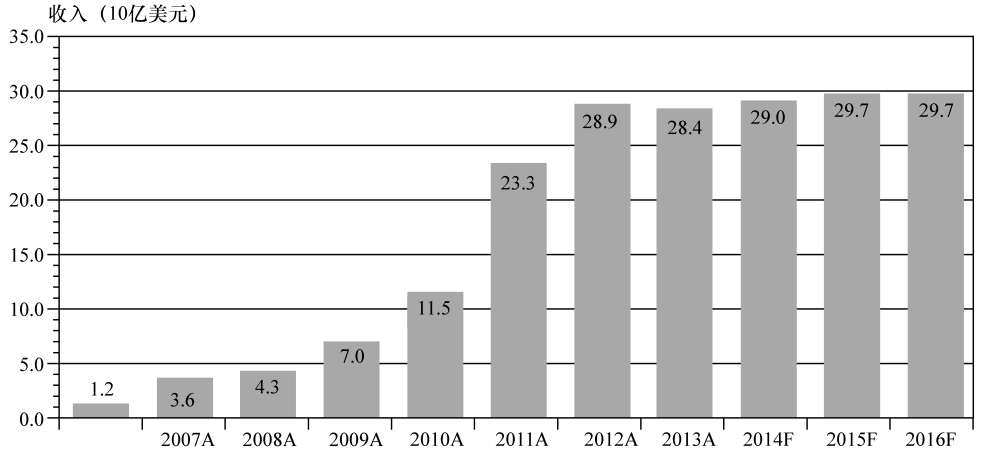


图 2.2　触控模块的收益图（单位：10 亿美元），2007 ~ 2013 年的数据为实际数，2014 ~ 2017 年的数据为估计数

注：本书作者认为 2011 ~ 2012 年 103% 的收益增长是因为市场调研报告撰写者有变，并不是因为市场规模的扩大而造成的。数据来源参考文献［4］。

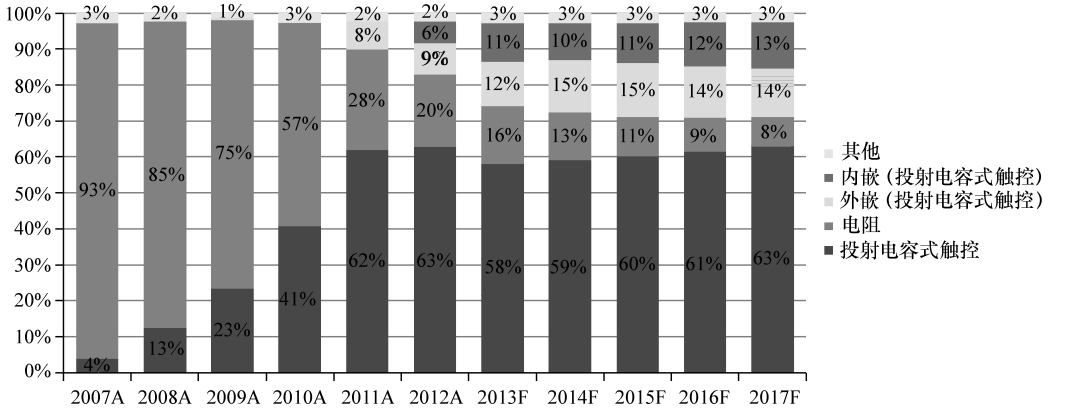


图 2.3　触控技术的出货数量比例图，2007 ~ 2012 年的数据为实际数，2013 ~ 2017 年的数据为估计数。模拟电阻式触控技术在 2007 年占据了 93% 的份额，而在 2012 年就暴跌至 20%；与此相比，投射电容式触控技术的份额（包括分离式和嵌入式）则达到 78%。有研究预测在 2017 年这一比例会上升到 90%，非常接近模拟电阻式技术在 2007 年的水平。数据来源参考文献［4］

2.2.1　触摸屏

从普通用户的角度来看，触摸屏就是一种可以感知并且对触碰到屏幕的物体——手指、输入笔或者信用卡的一个角——做出响应的计算机显示屏。而从技术者的角度来看，显示屏和用来感应触碰物体的元件分属于不同的电子系统，两者必须被区别对待。当两者合为一体

时，这类产品通常被称为“交互式显示屏”或者有时候就被叫作“触摸显示屏”。

在本章中，“触摸屏”一词仅仅用于描述可以感知用户的触摸，并且将这种触摸的信息转化为电脑可以理解和应用的信号的电子系统。对于当下大部分产品而言，这样的系统通常由专研触摸屏技术的公司提供（他们通常被称为触摸元件制造商）。触摸屏和显示屏的集成可以由触摸元件制造商、显示屏制造商、系统集成商，或者原始设计制造商/原始设计制造商来完成（对于消费电子产品而言，原始设备制造商通常指购买其他厂商的产品或者技术后冠注自己商标来销售的厂商，而原始设计制造商则是指设计或者制造设备的厂商）。

除了触控技术，触摸屏包含如下三个要素：传感器、控制器和电脑界面。这三者可以用一个立体图来体现（见图2.4）。对于除嵌入式触摸屏以外的所有触摸屏而言，传感器和保护性的屏幕玻璃盖片属于一个主体。而实际上感应元件有可能安装在玻璃盖片下面、边角上、表面上，或者是直接放在玻璃盖片上方。而对于嵌入式触摸屏而言，感应元件则集成在屏幕的内部，玻璃盖片则仅仅起到一个保护性的作用而已。

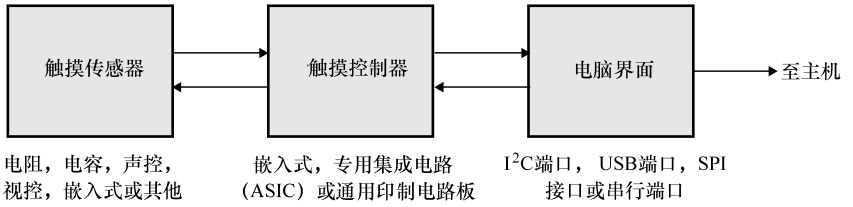


图 2.4 触控技术的要素概念图

2.2.2 按大小和应用对触控技术进行分类

大部分的触控技术都在各自擅长的领域有特殊的应用。就如世界上最著名的触觉技术研究者 Bill Buxton 所说，“每样东西在某一方面是最好的，但是在另一方面又是最坏的。”[5] 表2.1将本章中涉及的18项触控技术按两种标准进行了分类。第一种标准是设备的类型和尺寸，如下：

- 移动设备，例如平板电脑（2～17in[⊖]）。
- 固定的商业设备，例如销售终端机（10～30in）。
- 固定的消费设备，例如一体式电脑（10～30in）。
- 所有的大于30in的设备（通常称为“大画幅”触摸屏）。

在表2.1中，本章涉及的18种触控技术均按照设备的类型、大小和使用状态进行了分类。在表示设备类型的行和表示大小的列的交汇处，我们用如下不同的方式来表示每一种触控技术的使用状态：如果此种技术被广泛使用且被普遍地接受，则用A表示；如果此种技术虽然目前仍在使用但是已经接近被淘汰的状态，则用L表示；如果此种技术正在兴起，刚刚进入市场或者应用，则用E表示；空白则指并不存在此种对应大小和市场类型的触控技术。表中每种触控技术前的编号将沿用于整章中。

⊖ 1in = 0.0254m。——译者注

表 2.1　18 种触控技术分类

编号	名称	移动设备（2～17in）	固定的商业设备，例如销售终端机（10～30in）	固定的消费设备，例如一体式电脑（10～30in）	“大画幅”触摸屏（＞30in）
1	投射电容式触控	A	A	A	A
2	表面电容式触控		L		
3	模拟电阻式触控	A	A	L	
4	数字多点电阻式触控	E			
5	模拟多点电阻式触控	E		L	
6	表面声波		A	L	A
7	声学脉冲识别		A		
8	色散信号			L	
9	传统红外		A		A
10	多点触控红外	E		E	E
11	摄像光学触控			A	A
12	平面散射检测光学触控（玻璃光学平面探测）			E	E
13	视觉光学触控				E
14～16	嵌入式触控（外嵌，内嵌，混合）	A			
17	嵌入式光传感式触控				E
18	力传感式触控		E		

第二种分类的标准则是基于以上四种触控技术的普遍性做出的，如下：

- A：活跃等级，表示此种技术被广泛使用且被普遍地接受。
- L：式微等级，表示此种技术虽然目前仍在使用但是已经接近被淘汰的状态。
- E：新兴等级，表示此种技术正在兴起，刚刚进入市场或者应用。
- （空白）：指并不存在此种对应大小和市场类型的触控技术。

表 2.1 可以竖着看也可以横着看。例如，从移动设备那一列往下看，我们可以知道投射电容式触控技术、模拟电阻式触控技术（单点触控）以及嵌入式触控技术是移动设备生产中最主要的技术类别（A）；多点电阻式触控技术和多点触控红外技术则并没有完全在移动设备制造中普及（E）；除此之外就没有其他的应用于手机设备的触控技术了。同样地，从固定的商业设备那一列往下阅读，我们可以看到有五种触控技术在这一领域广泛使用（A）——相较于其他列是比较多的。这是因为商业性应用已经存在了将近 30 年，并带动了其他配套性触控技术的发展。

从表面电容式触控技术那一行看过去，我们可以发现这项技术仅仅应用在固定商业设备中，并且这种技术最终会消失（因此我们把其归于 L 类）。同样地，我们看到玻璃光学平面探测那一行，可以看到此种技术刚刚兴起，目前仅应用在固定的消费设备（例如一体式家庭电脑）和大型设备（例如信息屏）这两个方面。必须要知道的是，玻璃光学平面探测（第 12 项）只是技术的暂用名。其基础技术起初由触控技术供应商 FlatFrog 命名为“平面散

射检测”，这是投入市场后会采用的更准确的名称。

2.2.3　按材质和结构分类的触控技术

本节将讨论另一种为 18 种触控技术进行分类的方法，即按照材质和结构分类。触摸屏最基本的材质就是透明导体，最具有典型代表性的就是导电玻璃（ITO）。图 2.5 将触控技术按照“使用导电玻璃”（左边 8 个）和“不使用导电玻璃”（右边 10 个）分成了两组，而对“使用导电玻璃”，又会按照是否压制成薄片进行分类。

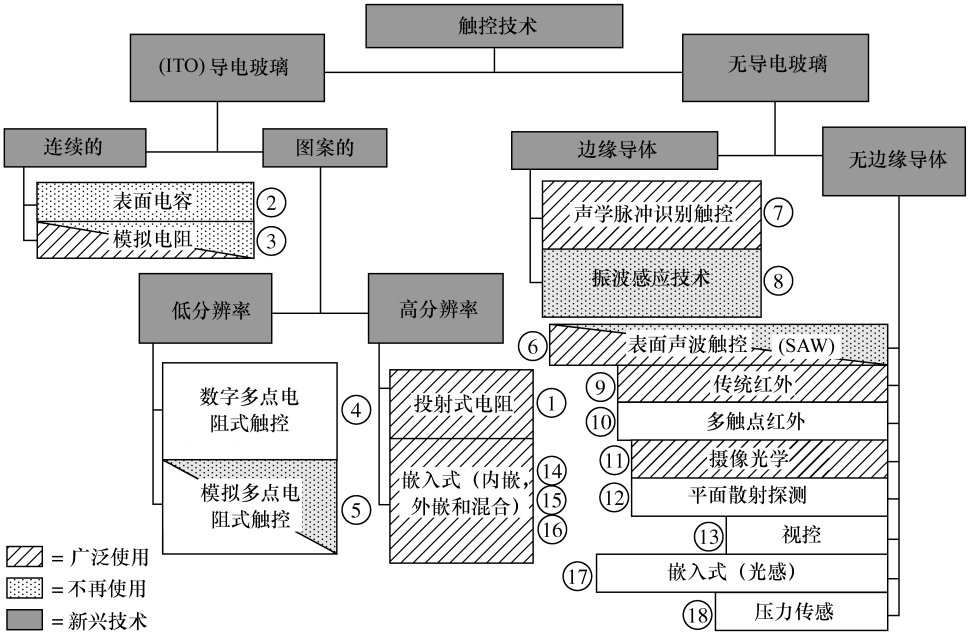


图 2.5　18 种触控技术首先会按照是否使用透明导体材质（典型的是 ITO）进行分类。然后会按是否被压制成薄片对使用 ITO 进行分类，然后再进一步按照分辨率来细分。而对不使用 ITO 的则进一步按是否使用边缘连接器来分类。注意触控技术的数目与表 2.1 相匹配

如果导电玻璃被压制成薄片，还会有低分辨率（毫米）和高分辨率（微米）之分。在不使用导电玻璃的触控技术中，有两种是使用边缘连接器的，另外 8 种则是不使用的。

2.2.4　按检测物理量分类的触控技术

不了解触控技术的人经常会问为什么有那么多种不同的种类。最简单的答案就是，触碰是一种间接的不容易测量的行为。如果你触碰某种东西，并没有一种确定的方法可以确定你触碰在什么地方，使用的力度有多大，以什么物体进行触碰的，甚至是不是你碰的都无法确定。因此有必要以表 2.2 中列出的物理量来描述一种触碰的行为。尽管如此，还是无法单用一种触控技术来明确如上提到的四个方面。这一难题被人们戏称为“不存在任何一种完美的触控技术”。

表 2.2　18 种触控技术可以用 9 种不同的物理量来衡量。为了确定触碰的位置、触碰的力度、触碰的物品，以及特定触碰人这四个方面，需要综合多种触碰技术来进行

编号	技术名称	测量的物理量
1，14～16	投射电容式触控技术，嵌入式电容式触控技术	电容
2	表面电容式触控技术	电流
3～5	电阻式触控技术（所有的形式）	电压
6	表面声波触控技术	超声波振幅
7，8	声学脉冲识别式触控技术	弯曲波
9～12	红外、摄像光学及平面散射探测光学触控技术	光的缺失或减弱程度
13	视觉光学触控技术	图像的移动
17	嵌入式光传感式触控技术	摄入光
18	力传感式触控技术	力量

2.2.5　按感知能力分类的触控技术

2011 年发表的一篇文章关注了深广度两分法对具有不同感知功能的触控软件的分类[6]，多伦多大学的 Daniel Wigdor 在该文中提出了图 2.6 所示的分类方法。在图 2.6 的左半部分，他列出了三种类型的能被感知的对象：触点（触碰次数和用户）、触控笔（支持程度）和影像（仅适用于视控式触碰技术）。在图 2.6 的右半部位，他列出四种可被感知的信息：接触（来自身体的不同部位或者不同的用户）、悬浮（支持程度）、接触数据（关于接触物的信息），以及压力（支持程度）。在本章提到的 18 种触控技术中，每一种都可以按照七种功能来表示。图 2.6 可以用来表示任意一种触控技术的特征，例如用于 iPhone 和 iPad 的 p－cap 技术，可以这样来表示：

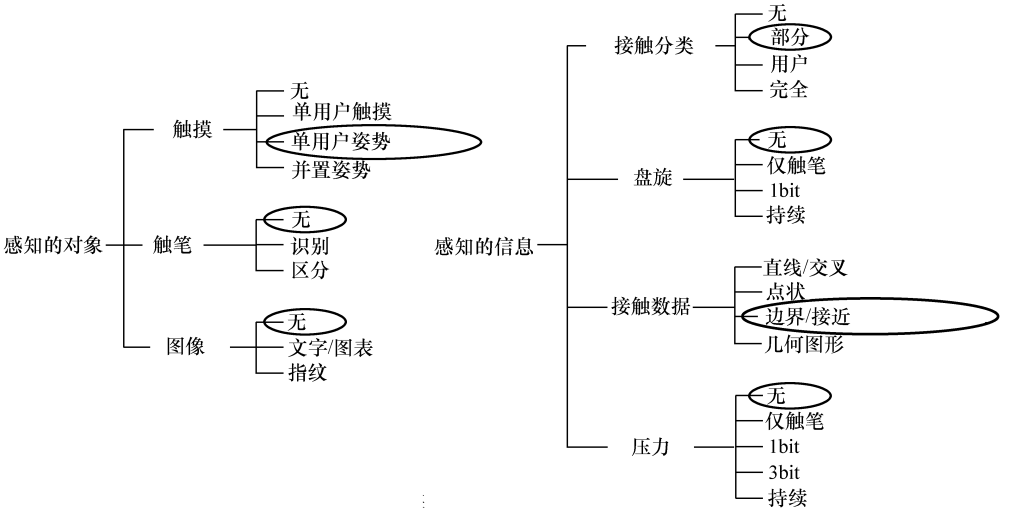


图 2.6　Daniel Wigdor 提出的基于感知触碰物和信息获取类型的触控技术分类法。画圈处代表了诸如在 iPhone、iPad 和类似移动设备中应用的 p－cap 技术的特征。来源：Wigdor，D.，2011。转载获得国际信息显示学会的许可

- 单用户使用，具有手势识别功能。
- 可以识别电容触控笔，但无法区分触碰来自手指还是触控笔。
- 不具有影像识别功能。
- 不具有区分触摸动作来自于身体不同部位或者来自于不同用户的功能。
- 不具有悬浮触控功能。
- 可以以一个矩形方块来粗略估计接触物的尺寸。
- 不具有压力感应功能。

2.2.6 触控技术的未来

尽管触控技术已经存在了半个世纪，并且在最近25年内风靡全球，然而它们依旧不够成熟并且没有充分地商品化。其中一个原因就是我们之前提到的“并不存在任何一种完美的触控技术”。另外一个原因就是触控技术产业需要强大的知识产权创新能力来驱动，竞争尤其激烈。在这一行业，新兴的公司如雨后春笋般冒出，它们不断地推出创新的触控形式（比如新发明更好地满足了消费应用需求的测量弯曲波的方法，或者测试触控力的新途径），不断优化触控过程（比如缩短了触控的反应时间），不断地引入新的制造材料（比如新的导电玻璃可以将投射电容式触控模块的成本降低将近一半）。诸如此类的创新行为不断地提升着触控技术领域的潜在发展空间。以下的几个方面可以帮助我们了解触控技术的未来走向：

- 触控技术的应用范围大大地拓宽了，从1in便携式设备到200in的投影屏幕都可以找到触控技术的用武之地。
- 将触摸行为和触摸物体完全整合。
- 嵌入式触控技术以更低的成本和更高的收益对分离式触控技术形成强有力的竞争。
- 投射电容式触控技术不断完善增强，可以将2号铅笔作为感知的对象。
- 触控技术包括了更多在图2.6中所示的感知功能。
- 能将2D触控、3D触控以及其他交互方式进行无缝对接。
- 使更多非透明物具有触敏的功能，任何物体都可以感知触摸。
- 成本，尤其是大屏幕触控方面的成本更低。
- 不断改善的软件开发环境使得创造更快更简便的用户体验变得可行（即触摸更为稳定流畅，用户完全不需要思考，感觉触摸起来如同行云流水般自然）。

2.3 触控技术的历史

触控技术有着丰富的发展历史，对其有6种基本的触控技术、每种都经历了不同的变化过程，我们并不感到奇怪。表2.3展现了从1965年到现在（将近50年!）触控技术的历史。此表列出了6种基本的技术类型，对于每一种技术的发明或者商业化起到了重要作用的公司或者机构都按年代顺序列出，并且附上了一些简要的说明。

表 2.3　本表在已发行的资料中最全面地记录了 6 种基本触控技术在历史上起到过的重要作用的公司。表中对每个重要的公司都会附上一句话来描述其贡献并注明相应的年份

公司名称	重要贡献	年份
电容式技术		
英国皇家雷达研究院（E. A Johnson）	第一个公开使用透明触摸屏的机构（在空中交通监控终端的显示屏上使用了互电容式技术）[1]	1965
欧洲核子研究组织（Bent Stumpe）	第二个使用了互电容式技术的机构（应用在质子加速器中）[7]	1977
MicroTouch Systems 公司（2001 年被 3M 触控公司收购）	第一家将表面电容式技术投入市场的机构[8]	20 世纪 80 年代中期
Dynapro Thin Films 公司（在 2000 年被 3M 触控公司收购）	第一家将互电容式技术商品化的机构（这项技术后更名为 3M 近场影像技术）	20 世纪 90 年代中期
Zytronic 公司（最先从英国发明家 Ronald Binstead 处获得专利权）	第一次将自电容式大画幅技术和互电容式大画幅技术投入市场[9]	1998；2012
Visual Planet 公司（第二家从 Ronald Binstead 处购买专利权的企业）	第二家将自电容式大画幅技术投入市场的企业[9]	2003
TouchKO 公司（2007 年被 Wacom 收购）	发明了反向斜铺场电容技术（RRFC™）[16]	2004
苹果公司	首先在消费产品中使用互电容式 p－cap 技术（iPhone 手机）[3]	2007
电阻式技术		
西屋电气公司	最先发明了透明模拟电阻式触摸屏（3 线式），但是从未投入市场[20]	1967
Sierracin/Intrex 公司	最先推出了数字化矩阵模拟电阻式技术，也有可能是最先将四线模拟电阻式技术投入市场的企业[21]	1973；1979
Elographics 公司（1986 年被 Raychem 公司收购，后者在 1999 年被 Tyco Electronics 公司收购，而 Tyco Electronics 公司又在 2012 年剥离出一个子公司 Elo Touch Solutions）	最先发明并且商品化五线模拟电阻式技术[18,19]	1977－1982
JazzMutant 公司（2007 年更名为 Stantum）	最先推出了数字化矩阵模拟电阻式技术，也有可能是最先将四线模拟电阻式技术投入市场的企业[21]	2005
JTouch 公司	最先在消费电子产品中使用了多触点电阻式技术	2008
声学触控技术		
Zenith 公司（SAW 专利在 1987 年被 Elographics/Raychem 公司收购，后者在 1999 年被 Tyco Electronics 公司收购，而 Tyco Electronics 又在 2012 年剥离出一个子公司 Elo Touch Solutions）	发明了表面声波（SAW）触控技术（SAW 触控技术的发明者 Robert Adler 在 1956 年发明了电视机超声波远程遥控器）[33,34]	1985
SoundTouch Ltd.（于 2004 年被 Elo Touch Solutions 公司收购）	联合发明了采样弯曲波触控技术（发明者 Tony Bick－Hardie，2006 年此项技术被 Elo Touch Solutions 公司更名为声波脉冲识别（APR）技术）[40]	21 世纪初

（续）

公司名称	重要贡献	年份
声学触控技术		
Sensitive Object 公司（于2010 年被 Elo Touch Solutions 公司收购）	联合发明了采样弯曲波触控技术（原名称为 ReverSys™，后被 Elo Touch Solutions 公司更名为声波脉冲识别（APR）技术）[41]	21 世纪初
NXT PLC 公司（于2003 年将专利许可转让给 3M Touch Systems）	第一家推出运用实时弯曲波触控技术产品的公司（此项技术由 3M 触控公司命名为色散信号技术（DST））[42]	2006
光学触控技术		
伊利诺伊大学	第一次使用了红外触控技术（第五代 PLATO 计算机辅助指令系统）[43]	1972
Sperry Rand 公司	采用 CCD，发明了摄像视觉触控技术	1979
惠普公司	第一次在商品中采用了红外触控技术（HP－150 微型计算机）[44]	1983
Carroll Touch 公司（1984 年被 AMP 公司收购，后者于 1999 年被 Tyco Electronics 公司收购，然后在2012 年剥离出 Elo Touch Solutions 公司）	大范围地在产品中使用红外触控技术	1980～1999
Poa Sana 公司	首先发明了波导红外技术[48]	1997～1999
SMART Technologies 公司	联合发明了运用 CMOS 技术的摄像光学触控技术	2003
NextWindow 公司（2010 年被 SMART 公司收购）	联合发明了运用 CMOS 技术的摄像光学触控技术；并为惠普公司的第一台消费性电脑提供了光学触控技术（TouchSmart 一体式电脑系列）	2003；2007
Perceptive Pixel 公司（由 Jeff Han 创立并在 2012 年被微软公司收购）	联合发明了运用 CMOS 技术的摄像光学触控技术；并为惠普公司的第一台消费电脑提供了光学触控技术（TouchSmart 一体式电脑系列）	2006
微软公司	推出第一台有投影的视觉触控产品（微软 Surface V1.0）	2007
RPO 公司（于2007 年创立；2012 年资产清算）	发明红外波导触控技术的第二家企业[46,47]	2007～2012
PQ Labs 公司	第一家推出采用多点红外触控技术产品的企业[49]	2009
FlatFrog 公司	发明了平面散射检测光学触控技术[55]	2007
Baanto 公司	最先推出运用二极管视觉触控技术的产品[53,54]	2011
MultiTouch 公司	最先推出运用集成相机视觉触控技术的产品[59]	2011
三星公司	第一个将内嵌光感视觉触控技术运用到产品中（SUR40 产品，使用于 Microsoft Surface 2.0，随后在2012 年被命名为 Microsoft PixelSense）[60,77]	2012
嵌入式电容技术		
Planar 公司	第一个发表了内嵌光感技术的论文[66]	2011

（续）

公司名称	重要贡献	年份
嵌入式电容技术		
东芝松下显示器公司	第一个声称发明了内嵌光感技术的公司[74]	2003
三星公司	第一个推出了采用任意形式的内嵌触控技术的产品（在 ST10 数码相机中使用压力电容）[64,65]；第一个推出了外嵌互电容式 p－cap 的产品（S8500 Wave 型号的 OLED 显示屏）	2009；2010
夏普公司	第一个推出了采用内嵌光感技术的产品（PC－NJ70A 上网本）	2009
IDTI 公司	第二个推出了采用内嵌光感技术的产品（21.5in LCD 监视器）[76]	2010
索尼公司（目前属于 Japan Display）	发明了内外嵌混合式互电容技术（最先用在索尼的智能手机 Xperia P™ 和 HTC 的产品 EVO Design 4G™ 中）[71]	2012
新思国际公司	和索尼公司一起开发了混合式电容技术[69,70]	2012
苹果公司	第一个推出了采用内嵌互电容技术的产品（iPhone 5）[72]	2012
其他触控技术		
IBM 公司	第一个推出了采用压力传感触控技术的产品（TouchSelect™ overlay）	1991
MyOrigo 公司（2004 年出售公司管理权；2005 年于芬兰重开，2006 年倒闭并在美国重开，被 TPK 公司在 2009 年收购）	目前为止唯一一个较为成熟的压力传感触控技术供应商（不考虑几个初创公司）[81]	2009
QSI 公司（2008 年从 Vissumo 公司中剥离出来，2009 年倒闭）；2010 年被 Beijer Electronics 公司收购	第一个成功推出了运用压力传感技术的产品（收费站的触摸终端）[79,80]	2008

2.4　电容式触控技术

2.4.1　投射电容式触控技术（编号 1）

投射电容式触控技术的历史对于一般人而言并没有其他触控技术那么清晰，主要是因为苹果公司在 iPhone 手机中对这一技术的创新应用太出名了，以至于模糊了对该技术之前使用的关注。通过电容变化来进行触觉感知的概念其实早在 20 世纪 60 年代就提出了。实际上，英国皇家雷达研究院在 1965 年就发明了透明触摸屏，并将其应用在英国的空中交通运输系统控制终端中，这项技术在现如今就是被我们所熟知的互电容技术[1]。可考据的对互

电容的第二次应用是在1978年欧洲核子研究组织的质子加速器中[7]。表面电容式触控技术（带未图案化的触摸屏）则是在20世纪80年代中期被MicroTouch Systems公司投入市场[8]。在20世纪90年代中期，几个美国公司开发出多层复合薄膜透明电容触摸屏（ITO，是如今投射电容式触摸屏的基本材料）。其中的两个公司Dynapro Thin Films和MicroTouch Systems分别在2000年和2001年被3M公司收购，组成了3M Touch Systems。Dynapro Thin Films公司的投射电容式技术更名为“近场成像触摸屏技术（简称NFI）”，这是3M公司在2001年的第一个投射电容式触摸屏产品。在1994年，英国独立发明家Ronald Peter Binstead发明了以超细微（25μm）电线作感应电极的互电容技术[9]，并将这种技术分别在1998年和2003年授权给两家英国公司Zytronic和Visual Planet使用，直到今天这两家公司依然在销售这项技术。在苹果公司将投射电容式触控技术应用在第一台iPhone手机之前，这项技术一直默默无闻[3]。苹果手机极致的用户体验赢得了消费者的欢心，从而促使其他智能手机生产厂家开始接受这项技术。在接下来的五年中，消费者为投射电容式触控技术使用的满意度设定了一项极高的标准：

- 可以实现多点同时触控（“多点触控”最初仅仅应用在图像放大上）。
- 对极其轻的触碰也能做出反应（不需要使用者出力）。
- 屏幕表面平滑。
- 优越的视觉体验（特别是相对于模拟电阻式触控技术而言）。
- 屏幕滚动快速而流畅。
- 屏幕坚固并且耐用。
- 触控功能与手机充分整合，使用起来不费力且充满乐趣。

2.4.1.1 投射电容式技术的原理

投射电容式触控技术主要有两种：自电容式和互电容式。图2.7展示了这两种形式的触控技术的原理。自电容式技术（见图2.7a）把被感应的物体（如手指）作为另一个感应电极。当手指触碰屏幕时可在手指和传感电极之间产生一个小量电荷。相反地，互电容式技术（见图2.7b）测量的是一对电极，它扫描到的是通过相邻电极的耦合产生的电容。当被感觉的手指靠近从一个电极到另一个电极的电场线时，互电容的变化被感觉到，从而报告触碰位置[10]。

两种触控技术的最主要区别在于电极的感应方式，而并不是电极的排列方式。在不考虑电极排列方式和电极数量的情况下，在自电容式触摸屏中的电极是一次感应一个。例如，就算电极按照X轴－Y轴矩阵进行排列，检测电极的时候也是先逐个检测完X轴上的电极再逐个检测Y轴上的电极。当手指触摸到屏幕的时候，最近的X电极和Y电极都会被探测产生一个电容峰波。但是，如图2.7c所示，当两个以上的手指以对角线方向碰到触摸屏的时候，屏幕上的两个点都会检测到峰波，于是“鬼点”（即相对于真实触点位置的“假性触碰”）和“真性触碰”都会同时被检测到。

要知道的是，这一缺点并不能排除在自电容式触摸屏上进行多点触控的可能性。模糊的触点位置不好判断，但是检测触点的移动方向是可以实现的。这样一来，即使屏幕上的两个触点产生了四个峰波，但只要这对触点的移动方向是呈对角线的，那么用户想要放大图像的

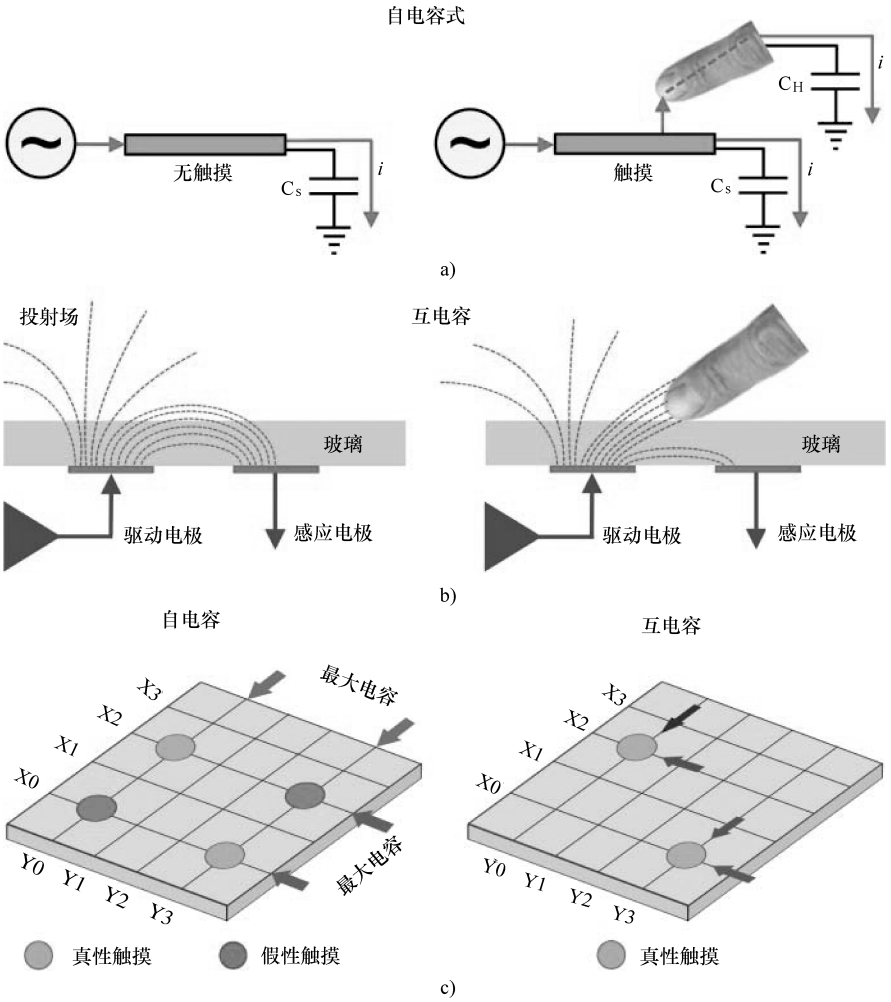


图 2.7　这些图片展示了自电容和互电容的区别。a）自电容本身包含一个单独的对地电容（C_S）；当手指触碰到屏幕时增加一个人体对地电容（C_H）。b）互电容包含两个电极之间的电容；当手指触碰到屏幕时两个电极之间的电容会减少。改编自 3M Touch Systems。c）自电容技术检测两轴上的每一个电极，因此当两轴上出现多个峰波时无法区分“真性触碰”和“假性触碰”（“鬼点”）（图示总共在 6 ×6 矩阵上测量 12 次）。互电容技术检测每个电极的交点，因此可以探测出多个触点的准确位置（图示总共在 6 × 6 矩阵上测量 36 次）。数据来源：改编自 Atmel

指令就能够被识别并且完成。因为这一点，再加上自电容式触控技术成本比互电容式触控技术更低，前者经常被应用在低端的手机生产上。

与此相对应，互电容式触摸屏上每个电极的交点都是单独被检测的。通常这可以通过架构两层导电层——驱动线和感测线来实现，运作上会轮流驱动一条 X 轴驱动线，并测量与这条驱动线交错的 Y 感测线是否有某点发生电容耦合现象。这一测量方法可以获知确切的触点位置。这使得互电容式触控技术成为厂商们制造高端移动设备的首选。

2.4.1.2 投射电容式控制器

投射电容式技术对电极的检测都是通过控制器来进行的。图2.8展示了一个互电容式触摸屏控制器的基本结构。感应器驱动会逐个激活X轴上的电极；模拟前端（AFE）则负责测量Y轴和X轴交汇处的电极，得出的数据会传送给模拟数字转换器（ADC）。然后由数字信号处理器（DSP）经过复杂而精密的运算对这一系列数据进行处理，再伴随着一系列诸如“手握压力抑制（消除人手握住无边手机时对屏幕产生压力的影响）”和“防止误触（消除无意识的触碰）”功能的处理，最后就将信号准确地反馈到触碰点或者触碰区域上。投射电容式控制器是专用集成电路（ASIC）的典型范例[11]。

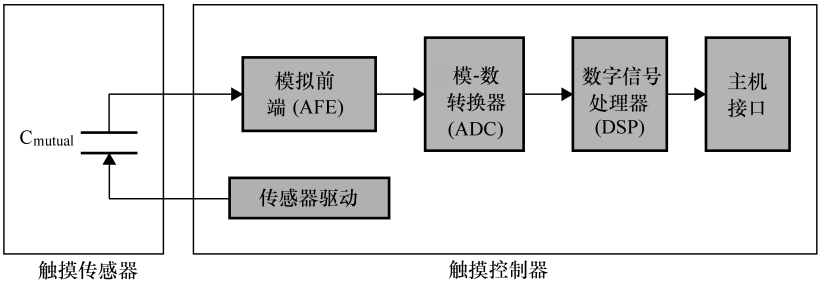


图2.8 图的右边部分显示了互电容式触摸屏的控制器基本结构。C_{mutual}代表一个交叉电极对的互电容

更多的创新发生在触控控制器设计领域而不是传感器的设计领域，那是因为触控控制器决定着触摸屏的灵敏度，而传感器则仅仅是接收电容数据的一个元件。但是，传感器的几何学结构一直对触控技术的提升贡献巨大。三个最出名的投射电容式控制器供应商（Atmel，Cypress，Synaptics，在2012年市面上基本所有的手机生产商都是采用这三大供应商的供货，除了苹果公司——其投射电容式控制器是自主设计并由美国博通公司生产的）[12]都是美国公司。这是投射电容式控制器产业愈发年轻的一个信号，因为大部分最终商品化的系统级专用集成电路的供应商基地都在亚洲。投射电容式控制器领域最近的一次创新是在2012～2013年，这期间触摸系统的信噪比大幅度提升。这一创新的价值在于使得投射电容式触摸屏可以支持笔尖仅为2mm的触摸笔进行输入，而不仅仅是手指。

如果一部智能手机能够支持细微笔尖的输入，那么它的价值就大大提升了。因为用户可以利用这项功能进行数据的“创造（画图、记笔记等）”，而不仅仅是被动地从传媒获取信息。在亚洲，人们经常需要在智能手机上输入汉字字符，而仅仅用手指无法实现这一点，因为在手写时指尖会挡住正在写的字。细微笔尖的触摸笔对于并不是为触摸而设计的操作系统而言也是一种很好的输入设备（例如，Windows 8的应用软件）。

2.4.1.3 投射电容式传感器

投射电容式传感器由一套透明的可传导电极的导电玻璃组成，这样的构成可以让控制器确定触摸点的位置。在自电容触摸屏里，导电玻璃通常被制造成一层或者两层，每一层上都存在着电极。当只有单层电极时，每一个电极都代表着一对不同的坐标并且和控制器相连接。当具有两层导电层时，电极以行和列的形式排列。每一行和每一列的交点代表着一对独

一无二的接触点坐标。但是就如前一节所提到的，在自电容触屏中，检测的是每一个单独的电极而不是电极的交叉点，因此该结构的多点触控的功能是受限的。

在互电容式触摸屏里，有两种最常见的电极的分布形式：

1）在空间上被绝缘层或薄膜或玻璃基板分隔开的纵横交错的垂直网格。

2）连锁菱形结构，相邻正菱形的两角由导线相连。

当该菱形结构用于两个隔开的表面时，每个表面的操作是很直接的。但为了使触摸屏尽可能的薄，该结构最常用于单一的共面层。这时的搭桥就需要额外的处理步骤以实现在跨越点处的绝缘。

图 2.9 展示了典型的互电容式触摸屏的叠层。为了使其和本章内其他类似的图示尽可能的简单易懂，我们做了如下的一些简化：

1）电极分布（第三行和第五行）呈分离的矩形网格状而不是更为常见的连锁菱形状；第三行显示了 Y 电极的端视图，第五行则是一个 X 电极的侧视图。

2）常用的光学透明黏合剂（OCA）省略；在第二行和第三行中间通常夹有 OCA。

3）图示的触摸屏使用玻璃基板；许多移动设备（特别是较大型的）的基板通常有两层聚对苯二甲酸乙二醇酯（PET）薄膜，每个对应每组电极。

4）LCD 内的薄膜晶体管（TFT）下面各层次（如底部偏光器、增亮膜、背光等）均省略。

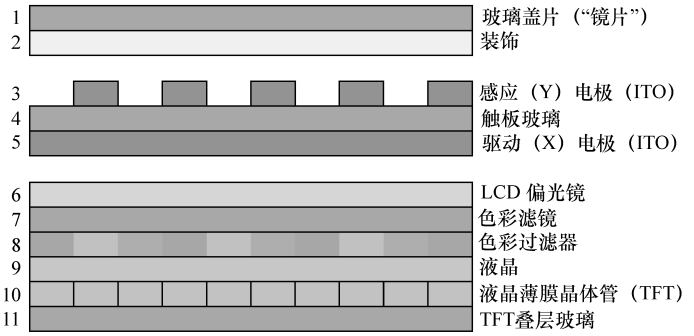


图 2.9　一个典型的互电容式触摸屏叠层简化图，位于 LCD 简化图之上。接触传感器基板（第四行）是一个两边有 ITO 的独立玻璃层

图 2.9 的一个重要方面是展示了触摸屏在叠层中增加了第四层玻璃。所有 LCD 都使用两层玻璃，而基本上每个移动装置都增加第三层玻璃（或塑料）作为保护和装饰层覆盖在 LCD 上。增加第四层玻璃总的来说没有必要，因为其增加了重量、厚度和设备成本。有以下两种移除第四层玻璃的基本方法：

1）触摸屏产业使用的方法，统称为“单玻璃方案”（OGS），但是不同公司的具体叫法不同，比如有叫“传感器玻璃盖片”的 。

2）LCD 产业使用的方法，称为“嵌入式触控”。这些方法之间存在直接竞争。

图 2.10 展示了 OGS，其中触摸屏电极被移至装饰玻璃盖片（“镜片”）的底面[13]。该

方案中，触摸屏制造商要么从合适的供应商处购买装饰玻璃盖片，要么就垂直整合和获取生成玻璃盖片必要的设备或技术。然后触摸屏制造商生产触控模块（传感器和控制器），把装饰玻璃盖片作为一个基板使用并将整个装配销售给移动设备 OEM/ODM（触摸屏制造商可能也会购买设备 OEM/ODM 规定的 LCD 并整合两者，以使 OEM/ODM 增值）。OGS 的好处是制造商可以持续从生产触控模块中获取利润而不是把利润送给 LCD 产业。

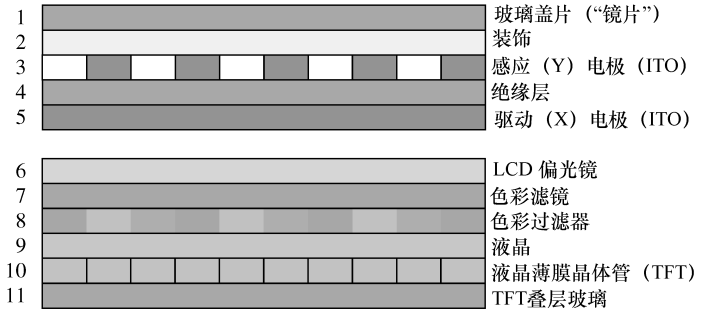


图 2. 10　OGS 互电容式触摸屏的一般叠层简化图，位于与图 2. 9 所示的同样的 LCD 之上。接触传感器装在屏幕玻璃盖片（第一行）的底面。该结构减少了图 2. 9 中触控传感器的独立玻璃层

图 2. 11 展示了最简化形式的嵌入式触摸屏（称为“外嵌”），其中第四层玻璃盖片因为触摸屏电极被装在彩色过滤玻璃盖片上、LCD 顶层偏光镜下而得以移除。注意外嵌结构具有与图 2. 9 和图 2. 10 所示的投射电容式结构完全相同的功能，只是电极的位置不同。外嵌方案的优势与 OGS 完全相同：移动设备由于移除第四层玻璃而更轻薄。外嵌方案对 LCD 制造商的优势是由于触控功能的附加值增加，他们的利润也将增加（但是触摸屏制造商的利润将减少）。

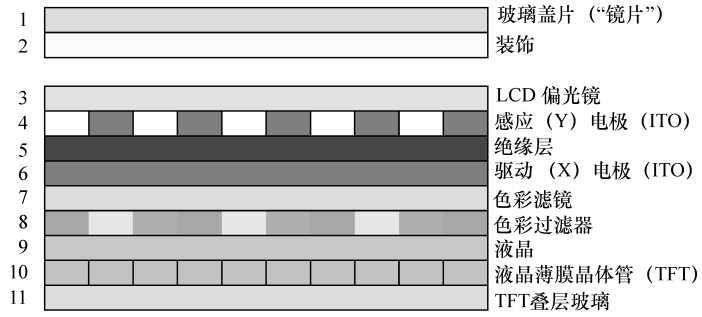


图 2. 11　外嵌式触摸屏的叠层简化图。该结构中，触控传感器安装在彩色过滤玻璃盖片的上方（第七行）、显示器顶部偏光镜的下方（第三行）。该触摸屏的功能与图 2. 9 和图 2. 10 所示相同，只是各传感器层的位置有变化

嵌入式触控的另一个有利因素是，触控传感器与 LCD 的整合使我们开始考虑把触控控制器与屏幕驱动器整合到一个单独的 ASIC 中，或至少建立起两个芯片的直接联系以促进协

作。产量是外嵌式触控的重大问题，因为在彩色过滤玻璃盖片上存储电极大大地增加了玻璃盖片的价值；如彩色过滤存储或解除电极存储有缺陷，两者都要丢弃。生产线管理也会成为LCD制造商面对的更为复杂的问题，因为他们可能需要推出10种不同的模型，每种要有500万的数量并装饰上独特的玻璃盖片，而不是给设备生产商输送5000万相同的LCD。

目前普遍认为触控功能与LCD的整合会自然驱使触控技术优化，成本降低。以上讨论明显指出了嵌入式触控并不一定比OGS更好。两者各有需要考虑的因素，且有些因素不仅是技术层面的，更是涉及市场和运行层面的问题。触控模块制造商与LCD制造商之间的竞争将成为各种嵌入式触控技术发展的主要因素。在2013年第二季度的触控预测中，Display-Search估计各类嵌入式触控到2017年前将仅占到所有投射电容式触摸屏单位出货量的26%[12]。

2.4.1.4　取代ITO的投射电容式线排传感器

在上述所有的关于投射电容式触摸传感器的讨论中，ITO被认为是制成导电玻璃的材料。但随着触摸屏变得越来越大，ITO的使用难度也随之增加，因为相对较高的基板电阻（50～200Ω/m^2）减慢了触敏处理过程，并且降低了产量。实质上增加了触摸屏的成本。除了极少的情况下，用ITO制成的触摸屏几乎没有32in以上的。

至少近十年内，大屏（大于32in）投射电容式触摸屏的导体材料的选择一直是10μm的铜线。铜线并非透明，但直径是10μm，接近人类视觉的较低区分度指数，因此几乎无法看到。40～100in的均有自电容（1～2个触点）和互电容（10多个触点）两种铜线触摸屏。大多数情况下，用在大型触摸屏内的基板是一层塑料薄膜（通常是PET）。10μm的铜线电极通常由一个自动机械装置铺成锯齿形的两层，两侧间放置某种绝缘体。尽管触摸屏感应器可能是以一卷薄膜的形式运输到合成商或设备制造商手中的，但是薄膜总会被压盖在基板的背部成为最终成品。其中一个最根本的原因是所有LCD的顶部都太软（仅有2H或3H的铅笔硬度），无法避免触碰造成的意外损坏。

2.4.1.5　投射电容式触控模块

“触控模块”一词仅应用于分离式触摸屏，因为嵌入式触摸屏只是显示器整体的一部分。前面主要关注了投射电容式触摸控制器和传感器；这些是投射电容式触控模块的主要元件。其次重要的触控模块元件是连接传感器和控制器的挠性印制电路（FPC）。触摸控制器一般安装在（和一些无源元件一起）FPC上，并接近于传感器以弱化噪声拾取。FPC的另一端通常被插入一个位于设备主板的连接器。

一个投射电容式触控模块通常以两种方式连接到显示器上：“沿框贴合”或“全贴合”。第一种方式下，将两边带黏性的密封垫片沿着显示屏的周边粘合，再把触控模块对齐显示屏，然后将两部分压紧。这会在显示屏触控传感器中间留出空气间隔；该间隔的范围在0.25～1mm以上不等，取决于显示屏的大小。这种沿框贴合法的优点是工艺成本低且产量高；缺点是它会产生额外的反射表面，在环境光强时将严重降低图像质量，整个装配也会稍厚。

在全贴合方法中，显示屏的整个上表面都要上一层高透明的黏合剂（干燥或液状）。对

齐之后，把触控模块按压在显示器上。普通使用的黏合剂有很多种；固化方法取决于类型。全贴合法的优势是光学性能总是较沿框贴合法更高，视差会更小，而且表层的耐用性也会增强（比如，它的规格能承受一个球从更高处落下的作用力）。劣势是该工艺的成本高、产量低。

今天大多数投射电容式触摸屏应用在消费者使用的设备中。根据 DisplaySearch 的报告，2013 年有超过 92% 的设备是智能手机和平板电脑[12]。剩下的消费产品包括笔记本电脑、多合一桌面电脑、便携式媒体播放器、便携式游戏机、电子书、便携式导航装置和相机。DisplaySearch 还称，2013 年，不到 1% 的所有投射电容式触摸屏是企业（商用）设备[12]。造成这种悬殊的原因是基本上整个投射电容式触控模块产业都在聚焦着这 92%（智能手机和平板电脑）。这意味着该产业对小批量、更高性能和环境规格的商业应用并不感兴趣，即使企业愿意为每台设备付出更多的成本。

相比之下，线排大型触摸屏（1% 的一部分）的应用常常与公众交互。其中一个最著名的应用当属“橱窗穿越”零售，即商家在非营业时间内接近潜在顾客，让顾客通过产品选择程序来跨越店铺橱窗并与商家交流。其他应用包括店内数码广告牌，公共信息服务站，如商场目录和自动贩卖机。

投射电容式触控技术的优缺点总结详见表 2.4。

表 2.4 投射电容式触控技术的优缺点

优点	缺点
无限、稳定的多点触控（如果正常运行）	成本高（主要是传感器；ITO 替代材料会帮助减少成本）
超轻的触碰（零压力）	接触物体必须有一定的接地电容（或是一支主动式触控笔）
平滑的触碰表面（无边）	难以集成（对每个新产品需要进行彻底的参数调整）
非常好的光学性能（特别是和模拟电阻比较）	因为隐形（ITO）电极而难以升级到 32in 以上
完全光滑和快速的滑动（如果正常运行）	没有绝对的压感；只是相对的手指接触面积
耐用的触控界面，不受刮痕和其他很多表面污染物的影响（受保护的传感器）	
可容许水在屏幕表面流过（但在 2013 年的消费产品中很少出现）	
可制成在特别厚的玻璃基板（约为 20mm）下运行	
可以按照 NEMA－4 或 IP65 的标准密封	

2.4.2 表面电容式触控技术（编号 2）

表面电容式触控技术由 MicroTouch Systems 公司发明并在市场推广。该公司成立于 1982 年，于 2001 年由 3M 公司收购并成为 3M Touch Systems 旗下的一个公司。由于表面电容技术缺乏在模拟电阻式触摸屏（当时主导的触控技术）中使用的易损塑料表层，它在 20 世纪 90 年代被认为是能够解决高难触控应用问题的方法。

如图 2.12 所示，表面电容式触摸屏传感器由一个透明导体匀质薄板组成，存放在玻璃

基板之上。用于表面导体触摸屏的最常见透明导体是掺锑二氧化锡（ATO），它能生成一个电阻率高达 1200Ω ~ 2000Ω/m^2 的高度均匀薄板。该技术的低成本方案是使用电阻较低的 ITO 或热解氧化锡（TO）替代。导电涂层和线性化的电极连接并被其包围，这些电极是由丝网印刷的银熔块制成，被连接到触摸屏的活动连接点（电极线性化的目的是纠正其电场内的本身的非线性（弯曲）属性，这与在矩形导电层内角到角流动的电流属性有关）。

导电涂层和线性化电极被一层焙干透明的绝缘硬膜覆盖，该硬膜通常由二氧化硅制成，还有防炫光（AG）功能。硬膜还总是抗粘连的，以减少手指和屏幕表面的静电摩擦；这使得拖拽物体（比如在视频扑克游戏中的卡片）更加简单。

图 2. 12 还显示了一般由 ITO 制成的备选保护层；其目的是保护导电层免受显示屏发射的电磁干扰（EMI）。由于底部保护层增加了触摸屏的成本、减弱了传递性（即降低了图像亮度），该保护层并不受欢迎。减少 EMI 效应现在往往是通过触摸屏控制器中的硬件实现的。

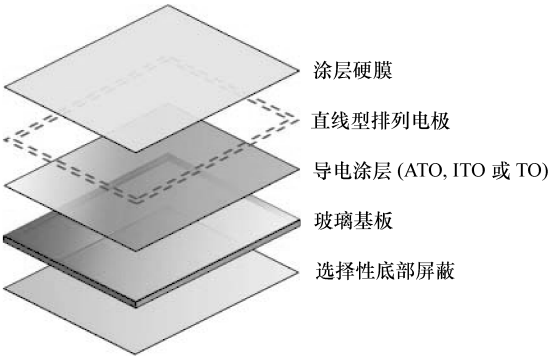


图 2. 12　一个表面电容式触控传感器的典型结构。触控感应器由一层均质透明导电硬膜组成并位于玻璃薄层的上方。导电涂层被线性化布置的电极包围，并由一层焙干的绝缘硬膜保护

表面电容使用一个贯穿导电涂层的匀质电场，这是通过将 AC 信号应用到涂层的四角而实现的。AC 信号（通常是 30 ~ 100kHz 频率范围内有 1 ~ 2V）是必需的，因为绝缘硬膜阻止了 DC 驱动信号与用户的手指连接。所有四个角都由完全相同的电压、相位和频率驱动。当用户的手指接触顶部硬膜，一小部分电能与用户电容耦合，导致一小部分的电流流过每个角落连接。控制器通过比较已知的在无触碰状态下的“基准”电流和用户触摸屏后的电流变化来识别触碰。触点的位置通过测量供应到各角的电流来定位，而且电流的大小与触碰位置到四角的距离远近成比例（表面电容式触摸屏的等效电路见图 2. 13）。电子控制器测量这些电流，将其转化成直流，对过滤噪声过滤，放大电流，再通过模拟数字转换器（ADC）将其转换成数字量，计算触点位置，增加合适的信息特征并向主机输出触摸位置坐标[14]。

表面电容的优缺点详见表 2. 5。

表 2. 5　表面电容式触控技术的优缺点

优　　点	缺　　点
在超滑表面的优越的拖拽性能	无多点触控
比模拟电阻的耐用性好得多	仅手指（或触控笔）
抗污染物	不如许多其他玻璃基板触控耐用
高度敏感（超轻触碰）	校准度渐变，易受 EMI 影响
	中等视觉质量（85% ~ 90% 光传输）
	不能在移动设备中使用

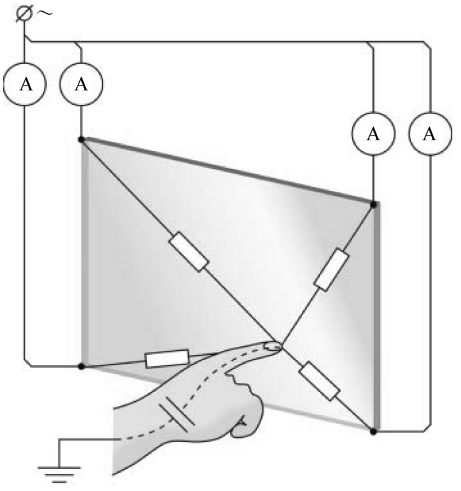


图 2.13 表面电容式触摸屏的等效电路。画圈的“A”代表了电流经过每个角落连接的测量。来源：Mercury13［CC－BY－SA－3.0（http：//creativecommons.org/licenses/by－sa/3.0）］，Wikimedia Commons。

表面电容是一项单触点技术。与模拟电阻相似的表面电容“模拟手势”功能是由 3M Touch Systems 的竞争对手在 2009 年开发的，但该功能效果有限，因为表面电容几乎绝大多数使用在商业性应用中，对多点触控的需求量相较于消费性应用要少得多。但是在不远的将来，某些商业应用的多点触控需求可能会改变。许多商业程序的用户（如公共咨询服务台用户和医疗器械用户）可能都会有投射电容式触摸屏的智能手机和/或平板电脑，因此他们会自然地对多点触控有所期待。终端机软件和医疗器械的开发商们可能会通过优化产品的多触点功能来满足用户的期待。反过来看，这也将把表面电容技术逐出市场，并以投射电容式技术取代之。

表面电容技术相当成熟；3M Touch Systems 已经对其不断改良，目前进一步优化的空间较小。3M Touch Systems 自 2001 年收购 MicroTouch Systems 以来一直保持着主要市场份额。但根据 DisplaySearch 的报告称，2013 年表面电容的全部市场价值仅为约 4500 万美元，相对于 2013 年整个触摸屏市场价值的 310 亿美元而言[12]，它并不是一个重要因子。

正确嗅到了未来触控技术的发展方向的 3M Touch Systems 已经将其注意力从表面电容转向了投射电容，这从 3M Touch Systems 在 2013 年展销会中表面电容的几乎全部缺席可以看出端倪。随着表面电容市场的萎缩，少数剩下的亚洲竞争对手也开始退出市场，这将加快该技术的消逝速度。结论是表面电容式触控技术正在走向其使用寿命的终结点；5～7 年内，该技术将永久地成为一项历史。

2.4.2.1 反向斜铺场电容

标准表面电容技术无法在移动中使用，因为它要求一个非常稳定的参考地来建立基准电流，从而获得“无触碰”环境的信息。CapPLUS™是一项运用了“反向斜铺场电容”（RRFC™）的表面电容技术，从而十分巧妙地解除了移动使用的限制[16]。RRFC 技术由 Touch Konnection Oasis（TouchKO）公司发明，该公司于 1996 年在得克萨斯州创立，2007 年被

Wacom 收购。

在标准表面电容中，导电基板是单一的静电场平铺。RRFC 使用的则是四个斜型的电场，如图 2.14 所示。通过在两个相邻角落的导体基板上安置一个 AC 电压，并在对面的两角安置一个 DC 电压，这样就能生成一个经过传感器和相应静电场的电压斜坡。触碰控制器按顺序对所有的四角组合重复这个命令，测量出四组由一个手指触碰产生的电流变化（两次 X 方向和两次 Y 方向）。这些数据的概念以四个垂直圆柱体的概念呈现在图 2.14 中。在测量中捕捉的信号数据继续经由额外数字信号处理，以过滤掉诸如接地物体变化、金属包边、EMI、皮肤干燥值或手指大小、薄手套等外部因素的影响。这使得触碰信号独立于所有外部环境的电容效果而仅仅来自手指触碰。

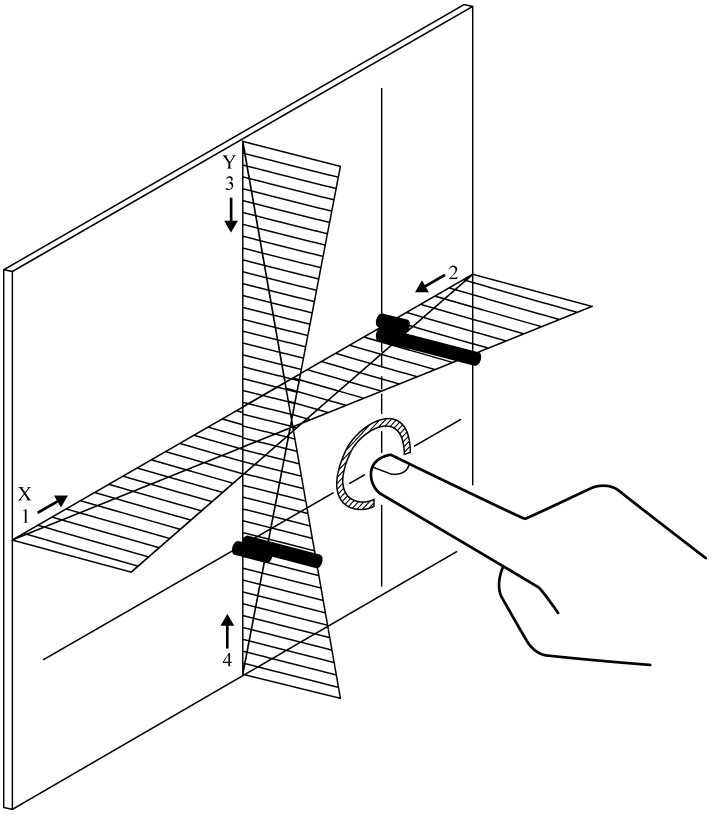


图 2.14　反向斜铺场电容（RRFC）触控技术使用了四个斜铺场（两个电压坡和两个静电场坡，图示为阴影三角形），有别于之前的表面电容使用的单一平面静电场。来源：转载获得 Wacom 的许可

- 这一技术改良的结果是显著解决了绝大多数表面电容遗留的问题。不幸的是，它仍有两个很大的缺点：RRFC 仍是一项单点触控技术；除了传统信息终端机以外的许多表面电容应用都明显地趋向于多触点技术。
- Wacom 是 RRFC 的唯一供应商；除非有一个压倒性的市场驱动力（比如，Wacom 的

数码笔使用在诸如三星 Galaxy Notes 这样的平板电脑中），OEM/ODM 倾向于回避独家技术供应商。

2.5 电阻式触控技术

2.5.1 模拟电阻式触控技术（编号 3）

模拟电阻式触摸屏通常认为是由 Elographics 公司于 1975 年发明的[17]。（Elographics 公司成立于 1971 年，于 1986 年更名为 Elo Touch Systems，于 2012 年更名为 Elo Touch Solutions。）然而，Elographics 公司原创的电阻式技术仅用于不透明的笔控操作仪，而不是透明的触摸屏。直到 1977 年 Elographics 公司才着手研发透明的版本（有弯曲度以适用于 CRT 显示器的表面）。该应用直到 1982 年的诺克斯维尔世界博览会上才作为商品面市[18,19]。

透明的模拟电阻式触摸屏由西屋公司率先发明。该项发明拥有美国专利，专利号为 3522664，专利申请时间为 1967 年，专利授予时间为 1970 年[20]。这块触摸屏由一块玻璃和一块聚酯薄膜（透明塑料）组成，两者均在表面覆盖了一层导电玻璃，并被间隔开。这是一个三线触摸屏（现已过时）的结构，所谓“三线”是指：

1）玻璃基底的相邻两面由二极管连接。

2）玻璃基底的另外相邻两面，也由二极管连接。

3）表层为聚酯层（更多细节详见专利记录）。

该发明并未投入市场。最早商业化的模拟电阻式触摸屏或为 Sierracin/Intrex 公司推出的四线模拟电阻式触摸屏，该触摸屏于 1979 年面市，品牌名称为“TransTech”[21]。

模拟电阻式触摸屏仅仅是用于定位触摸指令的机械开关。典型模拟电阻式触摸屏的结构如图 2.15 所示。一层玻璃基底和一层可弯曲的薄膜（通常为 PET 材质）均有一面被导电玻璃 ITO 覆盖。这两个涂层面一经接触，两个可导电表面便会被微小（50～250μm）、透明的绝缘点隔开。电压可以通过两层材料或其中的一层（取决于电阻式触摸屏的种类）。当手指点击可弯曲的薄膜层，两层材料的导电表面便可形成电流。ITO 材质产生的电阻在接触点形成了一个分压器，通过电压的比值便可得出触碰的位置。

2.5.1.1 模拟电阻式触控技术的变体

电阻式触控技术有以下三种主要的变化形式：

1）根据“导线”的数量。

2）根据层级结构。

3）根据选项。

导线的数量是指传感器之间的连接数。有三种常见类型，分别是四线、五线和八线。

在四线触摸屏中（见图 2.16），其中一个导电层左右边缘的母线相连接（即连接 X），另一导电层的上下边缘的导电层相连接（即连接 Y）。控制器产生通过 X 连接的电压，并计算其中一个 Y 连接上的电压，从而得出触点的 X 坐标。反之，控制器产生通过 Y 连接的电

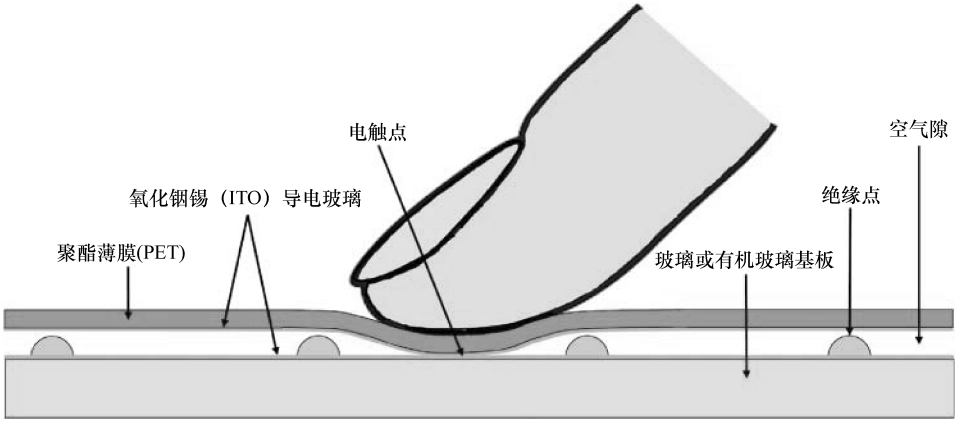


图 2.15　模拟电阻式触摸屏是一个用于定位触摸指令的机械开关。两个导电层被微小的绝缘点隔开；当两个涂层被触压在一起时，就形成了电接触。通过导电层上的电压比就能计算出触点的位置。改编自 Elo Touch Solutions

压，并计算其中一个 X 连接上的电压，从而得出触点的 Y 坐标[22]。

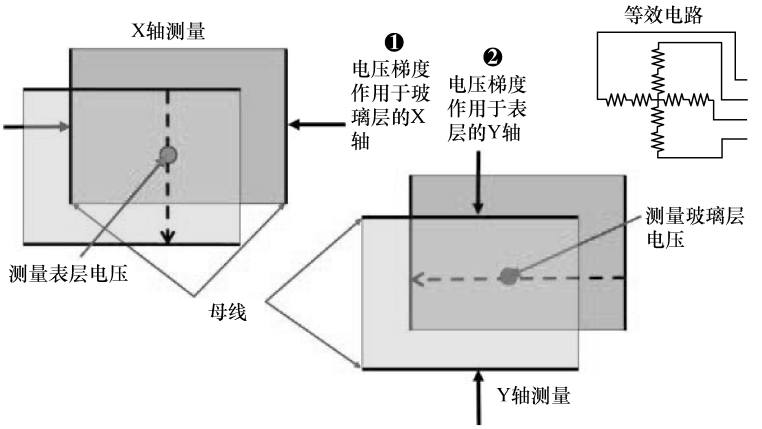


图 2.16　在四线触摸屏中，电压梯度作用于玻璃层的 X 轴的两个母线，产生的电压是在上层测算得出。反之，电压梯度作用于上层 Y 轴的两个母线上，结果电压则在下面的玻璃层上测算得出

在五线触摸屏中（见图 2.17），X 电压和 Y 电压作用于下面的导电层的四个角，上面的导电层的作用仅仅是接触点（接触刷）。控制器形成电压作用于 X 轴右边的两个角，并使 X 轴左边的两个角接地。上面的一层（第五根线）的作用相当于用来计算 X 位置的电压探针。同理，控制器反向进行该过程，把形成的电流作用于 Y 轴上面两个接触点并使 Y 轴下面两个连接接地，上层便可作为电压探测器来测量 Y 坐标。五线触摸屏时刻为触碰做好准备，

在触碰产生前，四个角被相同的电压作用，与此同时上层被高电阻接地。没有触碰时，上层的电压为零。当屏幕被触碰时，如前所述，控制器检测到电流增加并通过上层，就开始了计算位置的过程[23]。

四线触摸屏和五线触摸屏最主要的不同在于使用寿命。四线将1个手指的点击换算为100万次触击（或者将手写笔的一次点击换算为10万个字符），而五线触摸屏则将其换算为3000万次触碰。出现这样的区别原因在于上面导电层的不同作用：当它仅仅作为接触点而不是电阻分压器的时候，就能在导电涂层进一步退化之前停止运转。

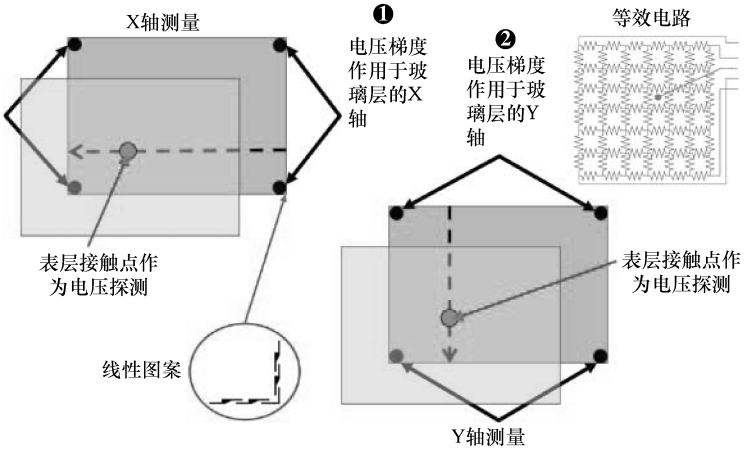


图2.17 在五线触摸屏中，电压梯度作用于玻璃层的X轴，而上层（即第五根线）的作用相当于电压探测器。当电压作用于玻璃层的Y轴上，上层的作用仍然一样

八线触摸屏就是在四线触摸屏的基础上多加一根导线，这根导线将每一个母线相连，这样就能直接测算传感器上的电压。这项技术的核心优势通常称为“四端测压”，也就是通过分离电压和电流，从控制器向传感器传导电流的四根导线的阻抗作用得以消除，这样就减少了屏幕校准误差。

过去，也出现过所谓的“六线”和“七线”电阻触摸屏。一般来说，它们都是触摸屏制造商试图回避Elo Touch Solutions公司对五线触摸屏的专利权的产物。六线就是在玻璃底层的背面多加一个接地层，然后这并没有什么实质作用。七线就是多加两根导线，用于减少因环境变化产生的误差，但是效果不佳。这些不同产品本质上与五线触摸屏无区别。

电阻式触摸屏有七种不同的层级材料组合，包括：

1）聚酯薄膜/聚酯薄膜。

2）聚酯薄膜/玻璃。

3）聚酯薄膜/塑料。

4）聚酯薄膜/聚酯薄膜/塑料。

5）聚酯薄膜/聚酯薄膜/玻璃。

6）玻璃/聚酯薄膜/玻璃。

7）玻璃/玻璃。

以上组合中的第一个材料用于顶层（即有五种材料组合是用聚酯薄膜做顶层），最后一个材料运用于底层。前面两种材料组合占据了80%的元件市场份额，大多数的材料供应商都在中国[12]。第一种材料组合在通信设备中应用最广（特别是手机），而第二种是通信和商业领域都适用。第三种主要用于不能出现玻璃破损情况的产品（例如儿童玩具）。第四种，触摸屏是聚酯薄膜构成，下面的基底是坚硬塑料材质，以增强耐用度。第五种与第四种基本相同，除了为增加硬度，以刚性平板玻璃作为基底（通常用于数字电阻）；第六种被誉为“装甲”，因为它解决了上层聚酯薄膜材质的耐用性不足问题。第七种因其稳定性主要应用于汽车领域。

相比其他的触控技术，电阻式触摸屏提供的选择非常多。常见的选项如下（详见本书2.13节）：

- 坚硬涂层——可提高耐用性。
- 抗反射涂层——可减少反射扩散。
- 反眩光涂层——变镜面反射为扩散反射。
- 防指印涂层——可防止指印带来的油脂附着在表面。
- 防污染（或“防腐蚀”）涂层——可防止类似永久标记墨水一类的墨水附着。
- 抗菌涂层——可减少附着在医疗设备上的细菌。
- 加固基底——可提高耐用性。
- 装甲表面——将聚酯薄膜/玻璃材料组合中的顶层锻压为微型玻璃，以提高耐用性。
- 高透射率/低反射率——提高户外使用的可视性。

2.5.1.2　模拟电阻式触摸屏的特性

模拟电阻是单点触控技术，也就是说，它不支持真正的多点触控。正如本章介绍投射电容部分提到的一样，随着数以十亿计的智能手机和平板电脑在市场上的推出，消费者对支持多触点触摸屏的需求也在不断提高。2008年，电阻控制器（也有时被称为“模拟手势”）的进一步改进，成为解决多触点技术空缺问题的营销变通方案。今天，许多标准的电阻控制器都能实现模拟手势的功能[25]。

实现模拟手势有几种方法，其中一种是测算在操作期间被传感器消耗的电流。当受到单点触控时，电流通常是恒定的，因此不受监控。但当有两个接触点出现时，两个导电层成为并联电阻，这就增加了电流消耗。这使得模拟电阻能够支持一些简单的两指操作，例如放大缩小和旋转，但它无法通过标准的多点触控测试，比如Microsoft Windows Touch Logo。

模拟手势功能在触摸屏的营销方面尤为重要，因为它让低端的模拟触摸屏至少可以在某一方面与投射电容式触摸屏媲美。事实上，电阻式的模拟手势带来的用户体验非常不同，这不仅因为其手势功能有限，也因为大部分电阻式触摸屏比投射电容式触摸屏需要更用力的触击，这样在同时移动两指进行操作时，持续用力按压非常吃力。

模拟电阻式触控技术的优势和劣势见表2.6。

表2.6中的前面四项劣势与投射电容式实际产生的新标准直接冲突。这些劣势导致模拟电阻式触摸屏在消费电子应用领域很快被投射电容式触摸屏抢夺了市场份额。根据DisplaySearch报告，模拟电阻式触摸屏2012年仅占据消费类单位出货量的16%，这当中的73%是用于手机。

而在商业应用领域的情况却大不相同。根据DisplaySearch报告，模拟电阻式触摸屏占据了88%的单位出货量。其主要的商业应用领域包括汽车、工业设备、零售/销售终端(POS)、信息点终端（POI)、自助服务设备，以及复印机、打印机之类的办公设备。电阻式技术在商业应用领域不断强势的原因如下：

- 电阻式技术作为标准的触控技术已经超过30年，它的短板已经为许多应用领域所接受。
- 虽然某些方面对多点触控的需求正在增长，但商业应用领域对于多点触控的要求不高。

表2.6 模拟电阻式触控技术的优缺点

优点	缺点
可用手指、手写笔及一切非尖锐物操控（触控物无限制）	无法实现多点触控（仅有模拟手势）
最低价的触控技术：每对角线英寸只要1美元甚至更低	光学质量差（20%的显示层发出的光线会因为层级反射而丢失）
广泛的供应渠道，有100家供应商（一个商品）	耐用性差（聚酯薄膜表层容易损坏）
可以按照IP65或者NEMA－4的标准密封	相对需要以更大力度触控
防屏幕污染	
耗电量低	

- 商业应用领域中，绝大多数的触控都是点击的形式，不会用到滑动手势，因此对于触控力度没有严格要求。
- 为了满足商业应用领域对齐平包边触摸屏的快速增长需求，大部分电阻式触摸屏供应商将五线触摸屏改进为齐平包边外观[26]。
- 现在对于手写笔的需求也相当大，而电阻式触控技术致力于各种无电源手写笔的使用。
- 苹果手机出现以后消费电子应用领域发生了天翻地覆的变化，而商业领域尚未出现这种飞跃性的变革。

电阻式触控技术只能应用于消费电子产品领域和商业电子产品领域。它在消费电子产品领域的主要优势在于低价和手写笔操控功能。然而，投射电容式技术将在五年内吸收这些优势，以至于将电阻式技术在消费电子产品领域的市场份额压缩至个位数。

在商业电子产品领域，电阻式技术将被投射电容式技术抢占大部分市场份额，至于早晚则取决于以下几个因素：

1）投射电容式触摸屏降价的速度。

2）更多投射电容式触摸屏供应商加入适应商业应用领域更多专业需求行列的速度。

3）每个应用领域对投射电容式触摸屏的关键性能的需求增长速度。例如，在消费品领域，对齐平包边触摸屏的需求增长更快，比如在保健用品和信息终端方面的需求就比销售终端、工业设备应用方面大。同理，休闲对于多点触控技术的要求相比销售终端增长更快，难以想象一家快餐店的点菜终端需要用到多点触控。

DisplaySearch 预计，在商业应用领域，电阻式触摸屏的单位出货量份额只会略微下跌，从 2012 年的 88% 下降到 2017 年的 72%[12]。

2.5.2　数字多点电阻式触控技术（编号 4）

有一种类型的电阻式触摸屏被称为“矩阵电阻”。这种矩阵电阻式触摸屏中，导电玻璃层被划分为网格状的行和列。它其实是第一种电阻式触摸屏技术开发的。Sierracin/Intrex 公司在 1973 年率先销售 ITO 涂层的聚酯薄膜。根据当时在该公司工作的雇员说，Sierracin/Intrex 公司发明了一种矩阵电阻井字游戏用以演示他们聚酯薄膜的产品[27]。这使得他们的客户迅速开发了一系列兼容矩阵电阻式触摸屏的产品。当然，行列交叉的矩阵电阻式技术在当时并不是独创的，它更早地被应用于膜片开关面板的不透明（金属）导体上。

在数字电阻式触摸屏中，两层（基板和盖板）上的 ITO 涂层都被划分成横平竖直的条块，相互之间形成特定的角度，如图 2.18 所示。当触摸屏被按压时，ITO 涂层上的一个或多个十字交叉点形成电触点，而每一个十字交叉点都形成一个独立的开关。条块的间距取决于所需的切换矩阵布局。这当中没有对称性的要求，因此矩阵的行数和列数都是任意的（例如 4 行 12 列）。大多数数字电阻式触摸屏是根据客户需求定制的，而且不需要控制器[28]。

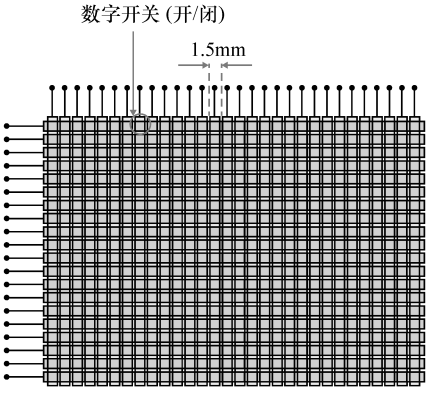


图 2.18　数字电阻式触摸屏由加了 ITO 涂层的两层（基板和盖板）构成，涂层被划分成横平竖直的条块，相互之间形成特定的角度。当触摸屏被按压时，一个或多个十字交叉点形成电触点，进而形成完整电路。具体的例子是一个 Stantum iVSM 的代表性触控传感器（见正文）

20 世纪 70 年代，数字电阻式触摸屏被广泛地用于商业性产品中，如工厂自动化、复印机、传真机、计算器和自动取款机。当四线和五线模拟电阻式触屏开始在 80 年代普及时，数字电阻式触摸屏日渐式微，这是因为它较低的分辨率以及无法处理写字和画图。

在 JazzMutant 创立的 2002 年以前，数字电阻还是一项单点触控技术。这个法国音乐播放器生产企业在 2005 年推出其 Lemur 产品之后，这一情况发生了改变。虽然多点触控自 1982 年以来就已经开始研究[5]，但 Lemur 音乐控制器实际上是第一个开始应用多点触控界面的产品[29]。2007 年，当 JazzMutant 决定单独营销他们的多点触控技术时（同年第一台

iPhone 手机发布），他们将公司重新命名为 Stantum。

Stantum 的技术是数字多点电阻（DMR），冠名“插入电压传感矩阵”（iVSM）。除了增加了一个复杂的多点触控控制器，该技术使用的透明转换矩阵概念与 30 年前引入市场的技术基本相同。核心区别如下：

- ITO 的各平行和各垂直线之间仅距离 1.5mm，比之前使用的要窄得多。这虽然使控制器获得更高的分辨率，但是也极大增加了控制器的连接数（如一个 10in 的屏幕需要 400 个连接线）。
- 触摸激活作用力相对较轻，仅为 8～15g。
- 支持多达 10 个多点同时触控的控制器最佳使用于手指触控和触控笔触控。这意味着它具有“防手掌误碰”的功能（忽略除了笔尖之外的任何触碰），这对有效的使用触控笔十分关键。

由于他们是一家很小的法国初创企业，Stantum 决定启用一个许可经营的商业模式，而不是成为触摸屏硬件的供应商。Stantum 开始将它们的控制器授权给两家 ASIC 制造商（ST Microelectronics 和 Sitronix），并与美国的一家专营商业应用的触摸屏制造商 Gunze 合作。2009～2011 年间，Stantum 在商业和军事应用方面的表现一般，因为手指和触控笔的结合在当时更受青睐。2012 年 Stantum 与 Nissha Printing 合作开发了一款 iVSM 产品，称为“精准触控 Z”。它在两层基板间加入了一层 Peratech 公司的透明压感材料，极大地增强了触屏的压感能力[30]。虽然 Stantum 较前些年略显低调，但它一直与合作伙伴致力于设计商用产品，如 K－12 教育平板电脑。尽管它不是唯一一家数字多触点电阻式技术供应商，但其名气在该领域业界毫无疑问是最大的。

数字多点电阻式触控技术的优劣势见表 2.7

表 2.7 数字多点电阻式触控技术的优缺点

优点	缺点
真正的多点触控	光学质量差（20% 的显示层发出的光线会因为层级反射而丢失）
可用手指、手写笔和其他非尖锐物体操控	耐用性差（聚酯薄膜表层容易损坏）
比投射电容式触摸屏价格更低	所需触控力度小 但仍大于投射电容式触摸屏
简单而成熟的电阻式技术	需要大量传感器连接
可以按照 IP65 或者 NEMA－4 的标准密封	供应商数量有限
防屏幕污染	传感器通常需要定制
耗电量低	

2.5.3 模拟多点电阻式触控技术（编号 5）

当 2007 年苹果手机引发了全球消费者对于多点触控技术的无尽需求时，模拟电阻式触摸屏行业发明了交融数字电阻式和模拟电阻式的触控技术，称为“模拟多点电阻”，作为投射电容式的低成本替代品。2008 年，中国台湾的 JTouch 公司是首家将这种技术商业化的触摸屏供应商。但一些触摸屏供应商将自己推出的版本以品牌命名。例如 Touch International 称其为“多点触控模拟电阻式传感器”（MARS）。有些供应商只是宣传“矩阵电阻式触摸屏”，这种触摸屏需要通过检查数据表来判断是模拟电阻式技术还是数字电阻式技术。判断

方法之一是看传感器边缘的连接。如果在四边都有许多连接，就是模拟电阻式技术。如果只有两边上有连接，就是数字电阻式技术。

如图 2.19 所示，在这种技术中，每个导体表面都呈横纵交叉的条块状，这样条与条之间的重叠交叉处就形成了一个个方形，每个方形相当于一个迷你的四线触摸屏。也就是说，在任意方形中，判定触击位置的方法与单点触控电阻式触摸屏一样，都是通过模拟电阻分压器法。然而，当两个触击在同一个方形中的时候，这两个触击动作就被均分，被当作一个单独触击处理，正如在一个四线触摸屏上操作一样。

除了 ITO 层的布局，模拟多点电阻式触摸屏在物理结构上与四线模拟电阻式触摸屏极为相似。模拟多点电阻式触摸屏通常只采用聚酯薄膜 – 玻璃结构，虽然有时玻璃 – 聚酯薄膜 – 玻璃结构的耐用性更高。模拟多点电阻式触摸屏的控制器有着如 Texas Instruments 公司这样的标准货源[31]。一些触摸屏生产商也自己生产模拟多点电阻式控制器，例如 AMT（Apex Material Technology）公司[32]。

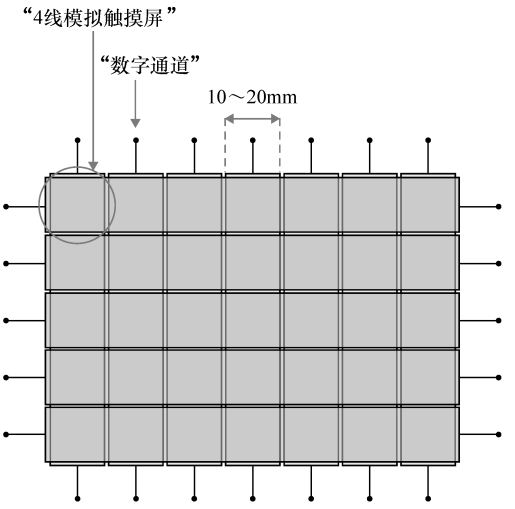


图 2.19　在模拟多点电阻式触摸屏中，通常均匀的导电层被划分为条状，这样条与条之间的交叉部分就形成一个个方形，这些方形一般边长为 10 ~ 20mm。每个方形如同一个独立的四线模拟触摸屏

模拟多点电阻式设计的初衷是为运行 Windows 7 系统的消费类一体化台式机以低成本解决多点触控需求。为了减少 22in 一体式触摸屏上的传感器连接数量（也就是降低成本），每个方形的宽度一般在 10 ~ 20mm。问题是，这意味着当用户将两个手指紧挨在一起并在屏幕上进行拖拽的时候，随着触摸位置的不同，触摸屏输出会在一个或两个触击命令间随机切换。通常在这种情况下，消费者会认定触摸屏有缺陷迹象。除了会产生有严重缺陷的用户体验之外，模拟多点电阻式触摸屏在消费电子产品市场上还有如下问题：

- 与投射电容式触摸屏相比它没有明显的价格优势。
- 它很难做到尺寸适宜，尤其是在大尺寸的情况下。
- 它也有电阻式触摸屏的基本限制通病（相对高的触控力量、低光学性能和低耐用度）。

在推向市场几个月之后，一家主要的原厂委托制作企业就召回了他们基于模拟多点电阻式技术的一体机。实际上，其他基于模拟多点电阻式技术的一体机在一两年后也未在真正意义上被消费电子产品市场所接受。这让电阻式触摸屏产业得到了教训。到了 2013 年还未被市场淘汰的模拟多点电阻式触摸屏是在以下几个方面做了改进：

1）尺寸缩小。

2）方形足够小，小到两个手指不能触控同一个方形。

3）专门针对商业和军事应用领域。

总的来说，模拟多点电阻式研发的初衷是，在消费电子产品领域，在多点触控性能方面以低价抗衡投射电容式技术。但它已经败下阵来，成为了一项无足轻重的小众技术。

模拟多点电阻式技术的优劣势见表2.8。

表2.8 模拟多点电阻式触控技术的优缺点

优点	缺点
多点触控，但相同方形内无法实现两点	视觉质量差（20%的显示光由于反射层丢失）
手指、触针或任何其他非尖锐物品操控（任何物体碰触）	耐用性差（易损耗聚酯薄膜表面）
简单成熟的电阻式技术	触力小，但仍高于投射电容式触摸屏
可按照IP65或NEMA－4环境标准密封	大量传感器相连（连接距离小到能足够感应两个手指并在一起）
防屏幕污染	供应商数量有限
耗电量低	

2.6 声波触控技术

2.6.1 表面声波触控技术（编号6）

目前周知的表面声波（SAW）是由著名发明家Robert Adler博士于1985年在Zenith发明的[33]。（Adler博士以共创1956年首度问世的超声波电视机遥控器而闻名[34]。）Zenith在1987年向当时美国Raychem所有的Elo Touch Solutions（当时名为Elographics）出售表面声波触摸屏技术。Robert Adler在售后继续为Elo提供咨询服务，为表面声波技术在20世纪90年代的商业化进程做出了积极的贡献。

如图2.20所示，表面声波传感器相对简易，由一个普通钠钙基板、4个压电换能器和4个波导部分反射器组成，该反射器由低温玻璃熔块制成并丝网印刷在表面上以火烧制。压电换能器成对安装，一个给X轴，一个给Y轴。X和Y轴的发射换能器发送穿过基板表层的超声瑞利波（Raleigh waves），瞄准X轴和Y轴的发射反射器。

在4～10MHz范围内的频率是可行的，但是出于历史原因，目前大多数的表面声波触摸屏在5.53MHz运行工作。发射换能器由一组呈45°角的隆起物组成；随着瑞利波击中这些隆起物后脊线，它们部分会被反射到屏幕上。相邻隆起物脊线的距离空间是基板上的扩散波波长的整数倍。这在波列穿行时能够防止隆起物对其产生的巨大干扰，波列在经过每个隆起物时会部分折射。在屏幕相对边缘处的一组匹配的接收反射器将波导向X轴和Y轴接收换能器。

任何特定的瑞利波从发射换能器到接收换能器的传输时间取决于路径的长短；反射器首反射的平行波耗时比起反射器尾反射的时间要短。这种运用“飞行时间”的方法在中介介质呈非色散时是可行的，也就是当波速在测试的频率范围内没有重大波动的时候。这样，碰触屏幕的物理位置就能折射到时间区域。当手指或其他柔软（吸音）物体碰触基板时，基板能吸收一部分特定的X轴和Y轴的瑞利波。

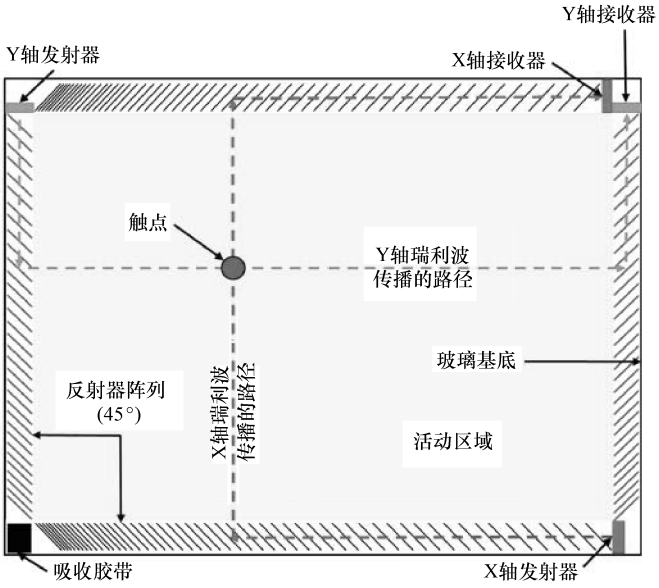


图 2.20　表面声波触控传感器由一个玻璃基板、两个发射换能器和四个 45°角的反射器组。瑞利波从一个发射换能器下行至一个反射器，穿过屏幕，上行经过相对的反射器，然后抵达接收换能器。改编自 Elo Touch Solution

如图 2.21 所示，触控位置是由测量波幅在 X、Y 轴的时间区域内的减少决定的。测量波幅的减少能够得出 Z 轴的触控力，尽管在实践中很少这样做。

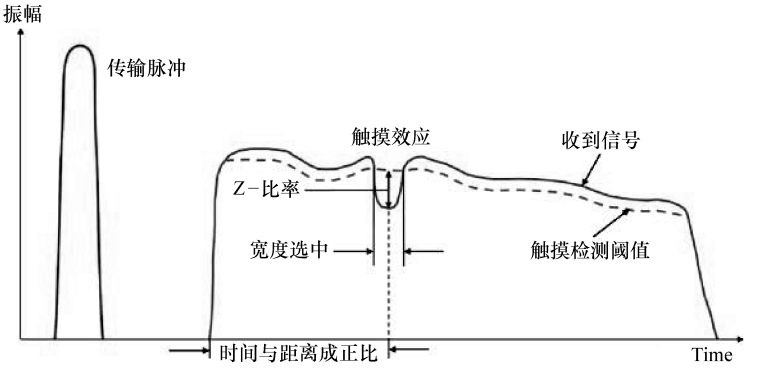


图 2.21　在表面声波触摸屏中，瑞利波在 X 或 Y 轴的传输脉冲信号在时间区域生成了一个振幅特性曲线；在轴线上的触控位置是由振幅下降的时间区域位置决定的。来源：改编自 Elo Touch Solutions

表面声波控制器本身具有自适性。为了忽略触控表面的大多数污染物质，它们持续监控在无触点条件下接收的波形，在环境发生改变时自行调整，并在必要时更改噪声阈值。

表面声波触摸屏有 6～52in。但由于瑞利波在钠钙基板中的相对高衰减性（吸入性），大于 24in 的屏幕要特别使用低衰减性的硼硅基板或钡基板（此基板能增加 30dB 到 42in 表面声波触摸屏的信噪比中[35]）。但是，随着触摸屏尺寸接近 42in，光触控技术变得性价比

更高。因此，虽然从技术上说制造一个52in的表面声波触摸屏是可行的，但在实际应用中却非常少有。

表面声波起初只是一项单触点技术，但是其两个最大的供应商——Elo Touch Solutions和General Touch（两者占据大多数市场份额）在2009年和2010年分别开发了支持双触点的触摸屏。Elo的方法是在除了45°角之外（如15°或75°）增加另一组的反射器，以为触控位置提供另一种数据来源[36]。双触点表面声波的主要不足是需要很大的压力来记录触点——在20~80g之间，具体取决于产品情况。即便是单一触点，这也比投射电容式要求的力度（基本为零）大得多。两个手指保持足够的压力，同时做出诸如放大或旋转的手势，或是需要用力下压来划过，这些都不是良好的用户体验。一些一体化的Windows 7联想和三星电脑型号与双触点表面声波曾经一起营销，但此外就再无大型的消费市场存在了。Windows 8已经不再考虑在消费市场投放表面声波，因为Windows 8触控规范要求至少有五点触控[37]。

标准表面声波的另一个问题在于其要求使用边框来遮罩玻璃边缘的反射器。Elo Touch Solutions和General Touch都已经发明了无边框（又称“零边框”“齐平包边”或“边对边”）的两点触控版本。Elo Touch Solutions的方法是将换能器和反射器移到玻璃以下以及环绕玻璃的边缘，这样瑞利波就能平滑地从前表面流向玻璃表面的后部。由于LCD框架使得玻璃以下几乎没有空间，Elo Touch Solutions使用单组并使其多路传输[38]。成型的边缘和换能器、反射器的位置使得该结构比投射电容式更难以整合在无边框的装置中。

由前面可知，电阻式触控技术目前在商业触控应用市场中占有很大的利润比例，而表面声波和表面电容技术则在竞争剩余的市场空间。表面声波的主要应用包括公共咨询台（移动信息站）、电子销售机（POS）、自动取款机和游戏机等。表面声波较表面电容技术有更多的应用，这是因为它成本低、可视功效好、耐用程度高、装配简单且供应商更多。这些优势再加上其两点触控功能，意味着表面声波可能继续发挥它的商业化用途。表2.9列举了表面声波触摸屏的优势和劣势。

表2.9 表面声波触控技术的优缺点

优点	缺点
由平面玻璃基板导致的高可视质量	没有多点触控（>2点）
手指、戴手套的手和软触笔可以激活	对表面污染非常敏感，特别是水
非常耐用；可以用钢化玻璃或化学加强玻璃防爆屏	需要相对较高的碰触压力（一般20~80g）
相对容易安装；防水和/或防尘版本可用	要求一个软（吸音）触控物体

2.6.2 声学脉冲识别触控技术（编号7）

声学脉冲识别（APR）技术和色散信号技术（DST，下一节内容）均使用了弯曲波。弯曲波是一种由某物体作用刚性基板表面而产生的机械能量。它不同于其他表面波之处在于它穿行整个基板的厚度，而不仅仅是在材料的表面；由此产生的一个优势是它的耐刮性。

当诸如手指或触针碰触基板时，触碰位置会产生向手指外扩散的弯曲波。因为弯曲波向外传递，它在扩散现象的影响下逐步分散扩展。弯曲波通过固体材料传播的速度取决于波频。由触碰引起的推力在基板内生成了许多不同频率的弯曲波。由于扩散，它们以不同的速

度传播到玻璃边缘，而并非以统一的波阵面。结果，基板边缘或角落的传感器就接收到与原始脉冲完全不一样的波形；波的形成过程被来自基板内层的反射进一步修改。最终生成的是大量的混乱波集，在整个基板内相互影响。声学脉冲识别和色散信号技术的核心区别是这股混乱的波集是如何处理的。

在声学脉冲识别触摸屏中，玻璃基板是事先通过机器在其上千个方位进行敲打“定性”的。每个弯曲波的“独特标记”方位被抽样并记录在一个查阅表内，该表存储在可长久保存的与某个基板有联系的内存里。操作时，碰触产生的弯曲波由四个不对称分布在基板周边的压电换能器感知（见图 2.22）。不对称性可以确保独特标记尽可能的复杂；高度的复杂性则有助于区分标记。控制器处理四个换能器的输出来获得当前触碰的标记，并将其与查阅表中存储的样本进行比对；采样点间插值被用来计算正确的触碰位置[39]。

声学脉冲识别的概念在 21 世纪初期由 Tony Hardie – Bick 提出，他是一个有着自己公司的个人发明家。SoundTouch Ltd. Elo Touch Solutions 在 2004 年前后收购了 SoundTouch 公司。在开展了一些商业化活动之后，2006 年公布了该项技术。这是为了取代模拟电阻式而开发的一项更耐用、成本更低的技术，出现在 2007 年苹果公司使多点触控成为产品不可或缺的特性之前。

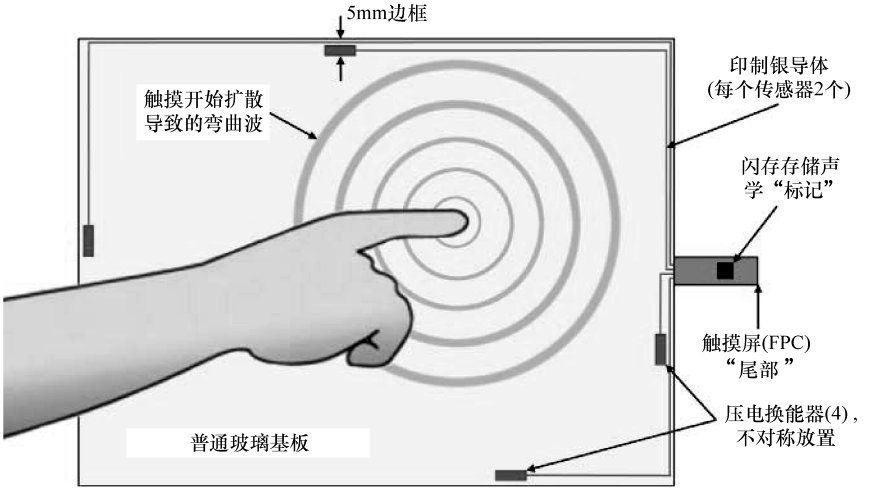


图 2.22　一个声学脉冲识别触控传感器由一个玻璃板和四个在玻璃层后部的压电换能器组成。当手指或其他物体接触玻璃时，弯曲波就会在玻璃基板内产生并由换能器采样；控制器决定了接触位置。改编自 Elo Touch Solutions

一项相似的基于感知弯曲波基本原理的触控技术也同时在法国 Sensitive Object 公司（当时称作“ReverSys”）独立开发完成[40]。该公司与 Elo Touch Solutions 的知识产权并未被相互损害，然而它们十分近似的交叉存在。因此两家公司在 2007 年发布产品后的不久就执行了交叉许可证协议。签署后，两家公司继续独立发布产品，因为协议只是为了避免对现有知识产权的诉讼而并非往后共享知识产权。Sensitive Object 的核心发明是只需几步就能快速定

性基板的方法，相对于 Elo Touch Solutions 用机器手敲击基板上千次的方法。2010 年 1 月 Elo Touch Solutions 以 6200 万美元收购了 Sensitive Object 公司[41]。两家公司知识产权的合并可谓是强强联手。

然而，即使组合性能增强，声学脉冲识别由于其对弯曲波的依赖性仍存在许多局限。最严重的局限是声学脉冲识别并没有“保持”功能（相当于长按拖拽鼠标）。当接触物体停止移动时，弯曲波也就不再生产了。这意味着在 Windows 桌面上普遍使用的拖拽 – 停 – 拖拽次序无法实现，因为声学脉冲识别驱动程序必须在开始“停”的时候发出一个自动“放开鼠标”的命令。这实质上限制了该技术仅在商业范围内使用（即非 Windows 用户界面）而并不面向个人消费者。

声学脉冲识别的另一个重要不足是它需要“敲击”来产生足够的可以被探测到的弯曲波。如果一个胆怯或犹豫的用户悄悄地接近声学脉冲识别触摸屏并按下而不是有意识的敲击（即便是用户用力地按下），这个触控将无法被识别。除了缺乏长按拖拽功能和需要明显敲击之外，还有一个局限是其本质的单点触控技术。在多点触控越来越普及的今天，单点触控技术显得越发无足轻重了。

声学脉冲识别对弯曲波的依赖产生了三个额外缺点。首先，该触控技术并不具有确定性。多次触碰完全相同的位置会在靶点坐标周围产生一个“点集”，这意味着每次如果用触控笔比划并不会产生完全相同的结果。这和模拟电阻式有很大的不同。后者在触碰完全相同的位置总能产生相同的靶点。

其次，声学脉冲识别的弯曲波侦测算法由于以下两个原因无法最优化：

1）由一系列快速敲击产生的间歇性弯曲波，比如出现在自动贩卖机的应用中。

2）由拖拽产生的持续弯曲波，比如在相同的自动贩卖机签下自己的名字。

对“一般通用程序”的优化使得快速敲击和拖拽的性能无法实现最佳。

第三个也是最后一个局限是 APR 固定（夹紧）工艺触摸屏对优化性能的重要性。这只需要想想敲击一个自然悬挂的玻璃面和一个四方夹紧固定的玻璃面的区别，就很容易得知了。也就是说，全球的产品制造商和系统装配商都必须接受关于如何适当装配声学脉冲识别触摸屏的培训。因此，声学脉冲识别作一个零部件并未在市场推广，只是被 Elo Touch Solutions 装配到了触控系统（最终产品）中。

鉴于投射电容式技术的主导地位和许多上述局限性，声学脉冲识别如今已不可能成为主流的触控技术。但是 Elo Touch Solutions 正在把声学脉冲识别有限的市场潜力与 POS 应用结合，使得上述局限不再重要。除了触摸屏显示系统之外，Elo Touch Solutions 也许可以吸纳 ReverSys 在制造触敏平面上的优势，从而开发更多非传统的专营市场。比如，可以让智能手机的后壳实现触敏。

由于声学脉冲识别和色散信号技术非常相似，两种技术的优缺点可以合并总结在表 2.10 中。

表 2.10　Elo Touch Solutions 的声学脉冲识别技术和 3M Touch System 的 DST 基于感知弯曲波的优缺点

优点	缺点
用手指、触控笔或其他触碰物体	无“长按拖拽”
由于平面基板而具有高可视质量	无多点触控
十分简易的传感器（玻璃基板 + 四个压电传感器）	要求足够的触控速度（敲击）以生成弯曲波
抗表面污染物；可在表面受刮或接触外部物体的情况下工作	包边内的固定夹紧工艺十分重要
很容易制成无边框屏幕（齐平包边）	非确定性操作（“多点集”）
	由于快速敲击（间歇性弯曲波）和拖拽（持续性弯曲波）难以实现优化

2.6.3　色散信号技术触控技术（编号 8）

色散信号技术（DST）是 3M Touch System 的商标名，该技术是一项基于感知弯曲波的触控技术。

Elo Touch Solutions 的声学脉冲识别（前一节所述）和 3M Touch System 的色散信号技术的核心区别在于，色散信号技术能够实时分析弯曲波以计算触点位置，而不是把碰触生成的弯曲波与存储的特性样本进行比对。图 2.23 展现了弯曲波在玻璃基板上的效果。第三幅图表现了声学脉冲识别采样并比对的波形；第四幅图体现了色散信号技术实时算法处理的样式结果。

在介绍声学脉冲识别时提到，经过基板的弯曲波传输速度随频率的改变而发生改变，这能导致信号的扩散或推广。接收到信号后，色散信号技术将会重组扩散的信号，该过程包括运行允许延迟和频率差别的程序，再运行四个传感器之间的相关性估算，最终三角测距出原始的触碰坐标。实际上，这属于本身耐受信号反射和干扰的扩展频谱技术，本质上能容忍信号的反射和干扰[42]。

3M Touch System 在 2003 年获得了英国 NXT PLC（New Transducers 公司）色散信号技术核心技术的独家许可。NXT（后于 2010 年更名为 HiWave Technolgoies PLC，于 2013 年再次更名为 Redux Labs）以首创平面扬声器闻名。在该设备中，压电换能器安装在硬基板的边缘上并受到音频信号的驱使，使得基板具有扬声器振动膜的功能。NXT 已经意识到（并获得专利权）相反的假设，即基板的振动（弯曲波）可以被换能器感知并用来定位波源（触点位置）。

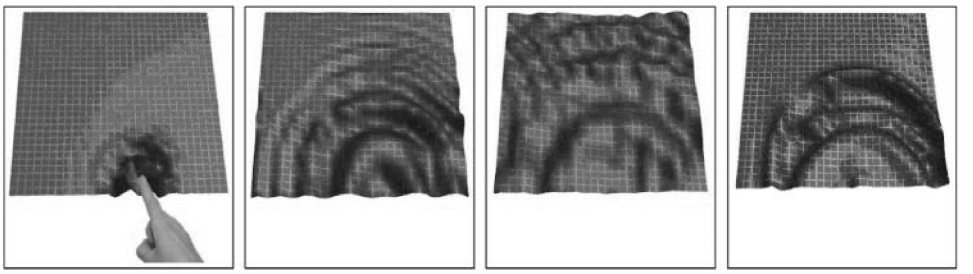


图 2.23　这组图展现了弯曲波在玻璃基板上呈现的效果。3 是声学脉冲识别脱机采样和比对的典型模式；4 是通过色散信号技术实时计算处理的模式结果。来源：改编自 3M Touch System

3M Touch System 和 NXT 合作开发了大量的商业化色散信号技术。2004 年 3M 提前发布了其首款色散信号技术产品的声明并于 2006 年实际发行。起初上市并不顺利；3M Touch System 一年多后宣布首款产品下线，并终于在 2007 年重新发布。由于当时 3M Touch System 的主打产品是大小在 5.7 ~ 32in 的电容式触摸屏，为了避免冲击这一产品市场，它把色散信号技术定位在尺寸 32 ~ 55in 的大画幅显示器开发上。相反，Elo Touch Solutions 则聚焦于 32in 以下的产品——并不是出于竞争的考虑，而是因为其无法使声学脉冲识别技术在数码广告牌应用中发挥最佳性能（Elo 的声学脉冲识别数码广告牌产品于 2012 年从市场下线）。

色散信号技术应用程序类似那些使用摄像光学和传统红外技术原理的应用；交互信息和数码广告牌是其主要关注点。色散信号技术和声学脉冲识别具有很多相似的基本局限，见表 2.10 总结。

大约在 2011 年底，3M Touch System 停止了对所有色散信号技术的进一步开发。没有持续的研发意味着该技术迟早会失去竞争力。虽然 3M Touch System 仍在持续在现有交互信息和数码广告牌应用中使用该技术，但它有可能在五年内从市场消失。

2.7 光学触控技术

2.7.1 传统红外线触控技术（编号 9）

首个广为人知的红外触摸屏范例于 1972 年诞生于伊利诺伊大学的 PLATO IV 教学系统中[43]。该系统内，一个 16 × 16 的网格红外触摸屏被覆盖在了一个橘黄色的等离子位图显示器之上，目的是为了提供手动选择的功能。

其中一个最早的红外触摸屏商业产品是于 1983 年问世的 HP - 150——惠普的第一个触控微电脑（它有一个 9in 的 CRT 并带有 CP/M 操作系统）[44]。在 20 世纪 80 年代和 90 年代，Carroll Touch 曾被认为是红外触摸屏的领军供应商。AMP 于 1984 年收购了 Carroll Touch。1999 年，Tyco International 收购了 AMP 并随后在同年内收购了 Raychem，后者在 1986 年就已经收购了 Elo Touch Solutions（Elographics）。这样，Carroll Touch 在 1999 年就成为了 Elo Touch Solutions 的一部分。

如图 2.24 所示，一个传统的红外触摸屏在屏幕框架相邻两边装有红外 LED，另两边则是红外光检器。每个 LED 按序列脉冲，产生的光由对面的光检器接收（该序列脉冲使该项技术又称作“扫描式红外光”）。因此，在 X 和 Y 方向的红外光束网格得以在屏幕表面的上方形成。一旦手指或任何非红外透光体阻隔光束，控制器就会计算触点位置。

20 世纪 90 年代初期，Elo Touch Solutions 对红外线技术做出的一个细微却意义重大的改进是提出了“单发多接”的概念。即改变原先的发射和接收一对一的对应模式，而使每个 LED 发射器能够为最多五个接收器所见。这提高了 Touch System 的稳健程度，因为故障的接收器不会再在触摸屏上留下盲点。检查静止物体和排查物体变化的概念得到改进。这防止了污染物（如一团花生酱）制造屏幕盲点；使用多个接收器可以“四处寻找”污染物，因此

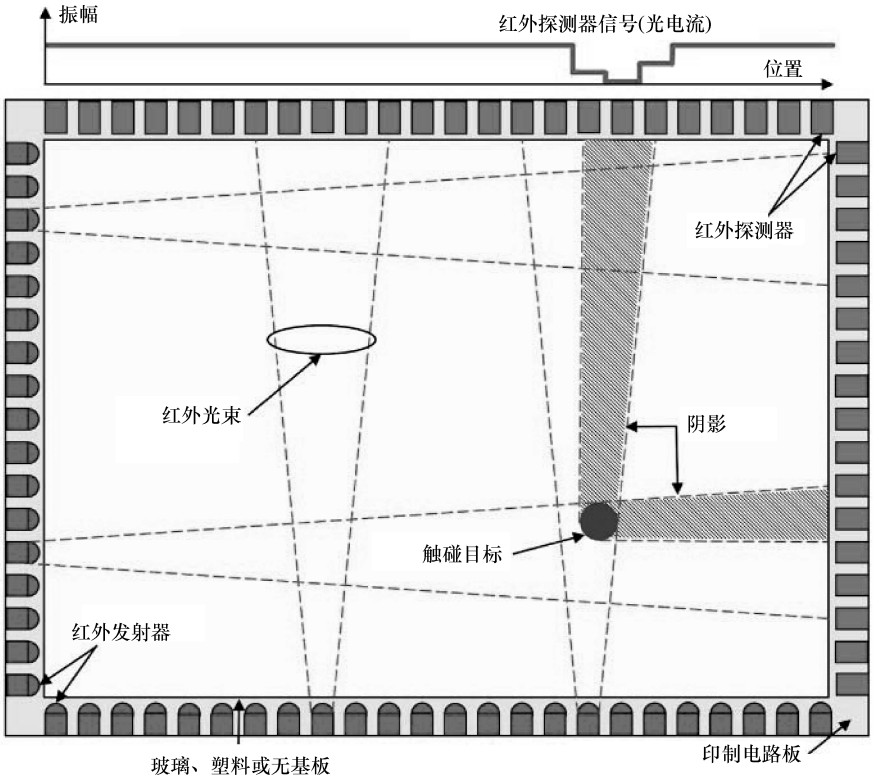


图 2.24　传统的红外线触控传感器是一个相邻两边是红外 LED、另两边是红外光检器的框架。LED 产生了红外光网格；当一个无法透过红外线的物体阻挡光束时，触点就可以被识别。改编自 Elo Touch Solutions

减小了它的影响。“传统”一词多指提及的红外线类型与 20 世纪 90 年代的在根本上是相似的。近期还开发了一些更新型的红外线，这些将稍后在本节讨论。

红外线在已经讨论过的主流触控技术中所占市场份额最小；据 DisplaySearch 报告，2012 年所有市场规模约有价值 4000 万美元[12]。几乎所有份额都属商业应用领域，包括自动取款机、自动销售点终端、各种资讯服务台以及诸如交互数码标识和交通导视系统等大画幅显示器。红外线是最为稳定的触控技术之一，能够容忍恶劣的环境，比如，它可以承受日光直射，也可以密封防污染。因此，它常被用在室外触摸屏应用中。红外线的独特之处在于它实际不需要任何基板——红外线光束能够被直接放置于显示器上而无需介入玻璃。也因如此，它通常被称为“红外触框”而不是“红外触摸屏”。

除了大屏之外的其他大多数应用程序里使用红外线是因为：①设备 OEM 长期使用红外线并认为其自有市场内最好的技术（比如，IBM 在自动销售点的应用）；②它的环保性。这些原因说明了红外线相对不太可能被投射电容式取代。但是在室内大型屏幕应用中，它正面临着来自摄像光学技术的巨大挑战，特别是后者在超大显示器中成本较红外线而言更低。

红外线起初是一项单点触控技术；当多点触控显得越发重要时，主要供应商纷纷开始支持某种程度上的两点触控。由于只有两种可用的信息轴（X 和 Y），两点触控并不能在缺乏额外信息时发挥作用（同样的问题也出现在自电容的“假性触碰”中）。这类有限的多点触控有时又被称为“一个半触控”。在 2010 年左右，Elo Touch Solutions 开发出了一个巧妙的方法：使用对角光束增加另一个信息轴（被称为额外维度 U）[45]。除了在双触点恰好与对角光一致（相互封挡）的特殊情况下，这在大多数情况下使得双触点清晰可辨。可惜的是，由于其成本很高，Elo Touch Solutions 从未将这项技术大量投入生产。

传统红外光可能作为一种独特的技术会继续存在，特别是对需要抵抗外界环境影响的应用来说尤为重要。未来五年内中小规模的市场份额将保持相对稳定，但是其在大画幅应用的市场占有率则可能由于摄像光学技术的发展而下降。

传统红外线触控技术的优缺点见表 2.11。

表 2.11 传统红外线触控技术的优缺点

优点	缺点
可延展至很大面积（超过 100in）	大多数为单点触控；对两点触控的支持有限（“一个半触控”）
无需触力，可经由任何非红外穿透物体激活	事先触碰（触点在实际接触屏幕表面之前激活）
由于平面玻璃基板而具有高光学性能	横截面高度（红外线发射器和接收器投射高于触碰表面）；印刷电路板（PCB）必须完全环绕屏幕
容易装配；有无基板均可；甚至可以作为自己安装的框架部件	包边的设计必须包括一个红外线透明窗口
非常耐用；可以用钢化或化学强化玻璃防止爆屏	非常难以防止误碰（衣袖、昆虫等）
可以环保密封用于户外	较低分辨率和准确度
可以抗外界密集红外线干扰（比如 75klx）	最小碰触物体尺寸通常大于 5mm（不可使用尖头触控笔）
	比起摄像光学耗费成本相对更高（成本随着参数比例改变）

2.7.1.1 波导红外技术

始于 2000 年左右，一家澳大利亚的名为 RPO 的初创企业开始开发针对“最后一公里”远程通信市场的聚合物光波导。2002 年，由于纤维和光纤产品的过度扩张，“最后一公里”市场也随之萧条。当时 RPO 重组并开始寻找新的应用产品。2004 年，他们决定在一系列传统的红外触摸屏中开始使用其开发的光波导技术[46]，并将其命名为“数码波导触控（DWT）”。如图 2.25 所示，这一概念是指使用单一光源和两套光波导传输 X 和 Y 方向的光，同时另一对波导收集光束并将其导向一个多像素光检测器。生成波导的制造工艺与 LCD 使用的照相平印技术相似。这实现了高分辨率，光导渠道也小到每个 10μm。

在现实中实施这个概念的困难和限制常常发生，这导致出现了一个稍微更为复杂的设计，如图 2.26 所示。

在有限的边界空间的限制下，RPO 仅使用一对光波导来搜集光并将其导向接收光检测器（一个含有 X 和 Y 100 多个像素的线路扫描 CMOS 感应器）。为了将光束散布到整个基板，RPO 使用双透镜的 IR－LED，一个针对 X，另一个针对 Y。基板本身用作一个光波导

[使用全内反射（TIR）]，对面两侧的抛物面反射器将光的方向进行180°改变，并帮助在基板上扩散光源。图2.26的白线指出了一个从LED左侧触摸屏到光检测器的光束路径，分别是顶视图和侧视图。产生的结果是一个低成本、高性能的2～15in（非固定）范围内最优的红外触摸屏[47]。

图2.25　RPO波导红外触控技术的概念图。来源：RPO

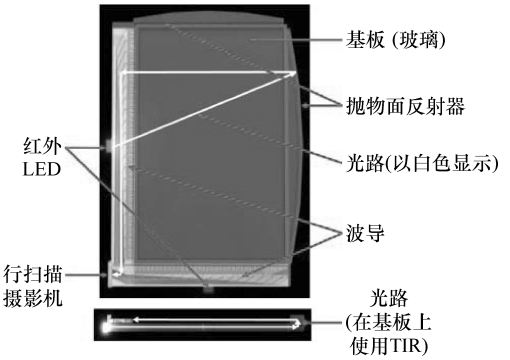


图2.26　RPO的3.5in波导红外触摸屏的实际结构。来源：RPO图片；作者附注释和箭头

该技术对带有反射屏幕的装置最为适用（比如，使用电子墨电泳显示器的电子书）。红外线能够在屏幕上操作而无需加层与反射屏幕形成了绝佳配比，因为后者需要有效地使用每个存在的光子（RPO的玻璃基板作为一种波导，可以在电子书显示器的下面安装）。但是正如所有的触控技术，波导红外线在其应用中也有一些如下内在的局限：

- 由于仅有两个位置的信息源，多点触碰仅限于两点；假性触碰减少到最小，但并未完全消除。
- 需要包边来保护波导和反射器。总体高度只有约1.5mm，但仍然不可为零，正如今天的智能手机和平板电脑屏幕上的包边处理。
- 该技术对屏幕上的残留物相对敏感，因为波导通道距表面仅200μm。

RPO于2007年发布该项技术，2008年改进了其性能，2009年增加其尺寸，并在2010年在一台13.3in的笔记本电脑中应用了该技术——这些都在国际信息显示学会（SID）的展示周会议上呈现。RPO当时与一家非常大的LCD电子消费品制造商合作（一个大客户）。当合作关系在2010年末突然终结时，RPO未对其他资金渠道做好充分准备，导致其无法支持电子消费品市场对生产量的要求。在超过10年的总价值5500万美元的投资之后，RPO最终于2011年清算。其资产（专利）的销售在2012年进行。目前尚不清楚该技术是否再次投入使用。

该项技术的一个更深历史层面值得一提。早在RPO开始设想他们的发明之前，有一项十分类似波导红外的触控技术已经发明并获得了专利。这是由位于加利福尼亚州硅谷的初创企业Poa Sana（斯瓦希里语“真酷”的意思）公司研制。Poa Sana的第一项专利在1997年申请并于1999年发布[48]。在1997～2002年间，Poa Sana在商业化推广技术方面并无太大建树，并把筹集的3500万美元主要用于研发上。2003年，当时在寻找进入触摸屏市场机会的美国国家半导体公司收购了Poa Sana公司的专利权。在花费了几年时间钻研该技术和评

估其市场机会后，国家半导体公司最终认为该技术的前景并不光明，于是他们又将 Poa Sana 的专利权还给了创始人。两家公司间没有发生过任何法律纠纷，因为没有哪家可以支付足够的金额诉诸法律。

大概有三个基本原因导致波导红外技术在当时的失败。总结如下：

- 应用该技术的最佳产品（电子书和其他带有反射显示器的产品）仅在 2010 年占有一个非常小的专营市场（要知道，“好钢要用在刀刃上”，任何技术必须要在至少一个应用中脱颖而出才能获得成功）[5]。
- 本质上该技术无法支持真正的多点触控和齐平包边设计，两者在 2007 年苹果手机上市后成为消费者电子产品的基本元素。
- 波导技术限制了触摸屏尺寸，不超过 14in 的大小使其应用无法在许多其他潜在市场推广。

任何技术都有其最佳的时间和地点，因此可以说波导红外的机遇还尚未到来。至少一家公司（Nitto Denko）已经提交申请并在其技术领域获得了专利批准。这样看来该技术并未完全消逝。

2.7.2 多点触控红外技术（编号 10）

多点触控红外技术是一个依靠红外 LED 发射器和光检测器的新成像方法，两组配件较其在传统红外触摸屏的使用并无不同。新方法支持达到 32 个以上的手指同时触碰；主要不同之处在于控制器对发射器和接收器的管理方式和过程。在大多数的应用中，该控制器使用尽可能多的接收器来捕捉所有单个发射器生成的触碰屏幕物体的阴影，而不仅仅是寻找成对的受干扰光束。

使用在这种成像方法的红外发射器和探测器主要有三种设计。前两种由其创始企业发现：①研发了这个技术并在 2009 年 1 月发布首个产品的 PQ Labs；②Image Display System；③目前尚无法找到具体的发明单位。

如图 2.27 所示，一个红外 LED 发射器一闪光，许多或所有的在对面两边或三边的红外光检测器就会记录下它们的光强度，生成一个单像素“图像”，它能显示所有 LED 和光检测器之间的物体阴影。每个图像像素（即来自某个光检测器的数据）通常都以灰度图展示，这对勾画移动物体的轮廓十分有用，因为现实的阴影在物体位移时并没有显著的轮廓。每次红外 LED 闪烁都会“重复”这个照相过程。这个过程速度极快，归并一起并按照数学数组排序的图像[49]能够使相对大量的有影物体被同时追踪。

用于所有三种结构的硬件也都相对相似。用户对不同的多点触控红外产品的体验大多是由使用算法的质量决定的。这些算法用来分析“阴影图”的数据，剔除重影点，处理遮挡并追踪移动和非移动物体。

目前该技术的供应商为数不多，最有名的是 PQ Labs（该技术的创始人）、Citron（DreamTouch™品牌）、Image Display System（PulseIR™品牌）、TimeLink 和 ZaagTech。

该技术使用的资源的两个核心特征是高速排序和大量持续的图像处理。因为该技术可实

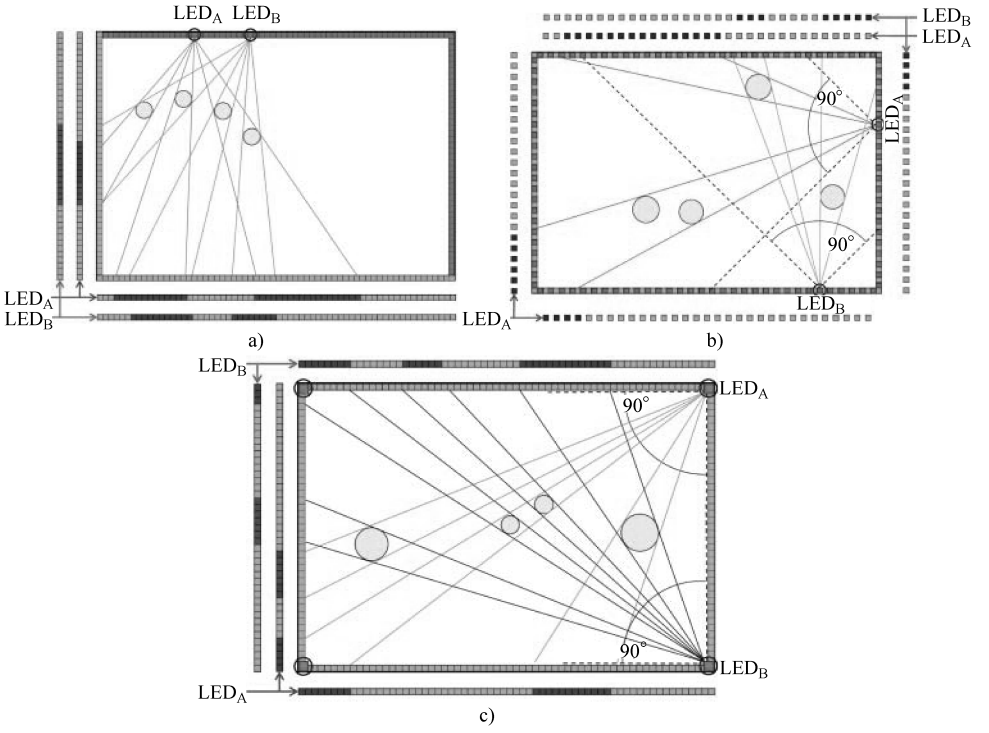


图 2.27　多触点红外触摸屏通常使用“成像”法。目前有三种结构：PQ Labs 的图 a，Image Display System 的图 b 以及不明来源的图 c。这些结构都比较相似，每个红外 LED 发射器一闪光，许多或所有的在对面两边或三边的红外光检测器就会记录下它们的光强度，生成一个单像素“图像”，它能显示所有 LED 和光检测器之间的物体阴影

现的最大分辨率与红外光检测器之间的间隔（通常直径为 5mm）密切相关，这种类型的触摸屏线路图可以显示其呈“阶梯状”或“锯齿状”。同样，随着接触物体离红外 LED 越来越近或变得越来越大，它们的阴影也会放大，这样就减少了每个图像可承载的数据量。

令人不解的是，该技术的主要问题在于缺乏清晰的应用。它似乎更多的是为了迎合主流消费对多触点的热衷而并非出于满足实际应用的需要。当下的商业红外触摸屏应用极少需要超过两点的触控，而且也从未有人定义过任何一个真正的 20 ~ 40 点的触控应用。另一个障碍是识别哪个触点归属哪个用户的问题，这尚未开发出良好的（实际的）解决方案。在大型水平显示器（游戏桌面）上进行的多人游戏也许是多点触控红外技术的最佳机会，但是不清楚该技术是否能够满足程序对速度和清晰度的要求。

多点触控红外技术因为受限于分辨率、速度和最小接触物体尺寸而不大适合交互白板应用。交互白板应用总体上需要一支触控笔以及快速识别抬笔小于 1mm 的距离。由于红外发射器放置在屏幕表面，任何形式的红外线要达到抬笔识别的要求非常困难。虽然多点触控红外技术也应用于白板产品（特别是来自 PQ Labs 的亚洲竞争商），获得的用户体验大多并不积极。

多点触控红外技术的优缺点见表2.12。

对当前多点触控红外技术发展状况的评价可以通过访问PQ Labs的网站（该技术的绝对先锋）了解。在本书编写的2013年，该网站的展示页面[51]囊括了七部各2.5min的视频；在全部18min的视频中，几乎没有任何应用显示使用超过两点的触控。最引人注目的应用是一个双人空中曲棍球比赛，每人使用两把球棍。

另一方面，继续发掘PQ Labs网站可以发现关于消除该技术诸多缺点的市场呼声很高。比如，他们声称通过加入10个“轻型处理器”（某种尚未定义的CPU）分散了控制器的处理工作量，因为这些处理器能够执行大多数在触摸屏框架内的多点处理，因此优化了触控速度和准确度。PQ Labs声称最小触体尺寸仅为1.5mm（相比起大于5mm的通常规范）——只是该描述并未出现在产品规范中，而是在其市场产品的宣传中有所提及。PQ Labs声称已经“优化的手写算法和独特的白板模式”能够使其产品支持清晰复杂的手制绘图，比如数学公式。他们还称产品不受“翘曲”和“恶劣的照明环境”影响——但并未提供一个“最小平坦度”或“最大环境红外光照度指数”的具体规范。

表2.12 多点触控红外技术的优缺点

优点	缺点
触点可从2点增至32点以上（仅控制器改变）	大屏市场中对多点触控的需求不显著
同多数传统红外优势相同（可调整大小、激活面广、零触碰作用力、光学性能高、耐用、可密封）	同多数传统红外缺点相同（预触碰感知、横截面高度、电路板环绕、包边设计复杂性、误触碰、低分辨率和准确度、最小接触物体大小）
物体大小识别（成像方法的副产品，可以从多个有利位置捕捉单个物体视图）	性能常常不如传统红外技术好（反应更慢、阶梯形轮廓和振动更频繁等）
	比传统红外技术价格高很多（也许由于暂时的“市场定价”）

上述部分通过例子指出了整个触控产业存在的一个根本问题，即市场宣传过多且缺乏足够的产品规范描述。

2.7.3 摄像光学触控技术（编号11）

尽管摄像光学触控仅于2009年随着Windows 7的发布才问世，但该项技术已经存在30余年之久。1979年，Sperry Rand集团首度定义了使用两个红外线性图像传感器［它们那时还是电荷耦合组件（CCD）］来定位显示器表面的触碰位置的概念，并获得了该项技术的专利。加拿大的SMART Technologies和新西兰的NextWindow在2000年前后独立开发了首个商业用途的基于光学Touch System的互补金属氧化物半导体（CMOS）。十年间，SMART将该技术应用于一部分自有产品，但直到2010年，该技术一直没有广泛使用。

惠普是第一个在桌面产品中使用光学触控技术的商家，它于2007年发布了TouchSmart™一体化电脑附带NextWindow触控技术。2009年4月，SMART状告NextWindow侵犯其专利，并于2009年6月许可Pixart使用该项技术。Pixart很快开始向Quanta供应光学传感器以应对2009年10月发布的Windows 7。Quanta成为了NextWindow的主要竞争对手。SMART在次年4

月收购了 NextWindow，于是中止了诉讼并缓解了 Quanta 作为竞争对手的经济冲击。融合两家公司的光学触控的知识产权比起一个公司的专利被另一个公司贬值（可能的诉讼结果）要明智得多。

摄像光学是一种遮光红外触控（此处“摄像”是指包括图像传感器、透镜、红外滤光片、外壳以及数据线的常规装配集合）。在最普通的摄像光学触屏形式中（见图 2.28a），屏幕角落有一个通过红外 LED 提供的外围背光源，屏幕周边有一个反射器（反射器是一种能把光从其射入的方向折射回去的材料，无论入射角角度）。由于反射器的作用，从屏幕边缘射出的光穿过屏幕表面 CMOS 线路扫描或区域成像器（相机）被安装在屏幕的两个或更多的角落；当手指触碰屏幕时，边缘光受遮挡，其阴影被相机捕捉。

要注意的是，尽管相机使用的是区域成像器而不是单像素线路扫描成像器，其仍然无法看见触碰手指的灰度图像；它只能辨别有光或者无光。控制器处理来自相机的数据，并用三角测距来确定触碰手指的位置[52]。

图 2.28b 展示了通过一个 512 像素光学传感器看到的光强度图。图中像素为 358 的急速下降是手指触碰屏幕的结果（即所有边缘背光都被遮挡的时点）。在 250 像素左右的缓慢下降是屏幕两个边缘的交点（即下边和右边边缘，从左上边的相机角度看去）；这是距离相机最远的点。270 像素左右的高峰出现在反射器把光原路射回相机的时候。

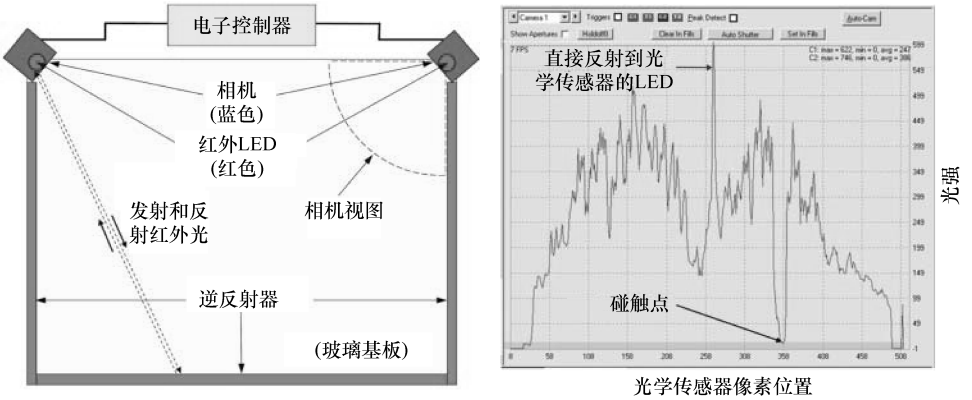


图 2.28　摄像光学触控使用由红外 LED 在屏幕的角落所生成的背光和一个边缘反射器。CMOS 行扫描传感器（相机）被安置在屏幕的两个或多个角落；当一个阻挡红外光的物体碰触屏幕，边缘光就会被遮盖，相机就能看到阴影。来源：改编自 NextWindow

2009 ~ 2012 年应用在桌面产品中的多数摄像光学触摸屏仅有两个 CMOS 传感器，这主要是考虑到成本问题。使用三角测距需要两个相机，这才能计算单个触点的 X 和 Y 位置。如果两个同时出现的触点能够被两个相机捕捉（即每个相机都能看到两个明显的阴影），那么就可能会出现四个触点——两个真正的触点和两个“重影”触点（指在位置上关联真正的触点）。这是在自电容投射电容、传统红外和单点触控表面声波中也同样存在的问题——

所有只能通过双轴获取信息的触控系统均是如此。在光学触控中区分真正的触点和重影需要有能够操纵多套触点的复杂算法。

高级算法还在另一种情况下彰显重要性——当同时出现的双触点无法被相机分清时(即一个触点被另一个遮挡)。图 2.29 展示了重影点和遮挡的情况。

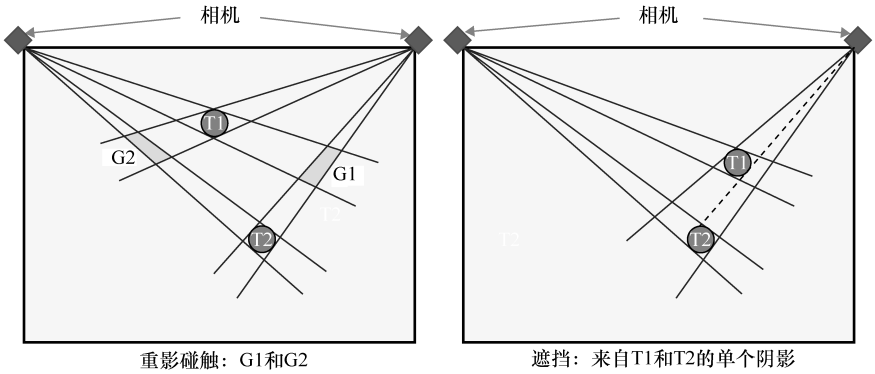


图 2.29 在有两个相机的光学触摸屏中尝试定位双触点时，重影点碰触（G1 和 G2，如左图绿色所示）和遮挡（一个触碰物体 T1 和 T2 的单个阴影，如右图所示）的问题无法解决，因为没有足够的位置来源数据

在一个双相机光学触控系统中的触摸屏控制器有大部分的处理时间是用在运行算法以消除重影点和弥补遮光。实际上，多点触控体验的质量在双相机光学触控系统中是取决于算法的复杂程度，而不是硬件的质量。因此，一些大型的（大于 30in）光学触摸屏使用四个相机来提供更多的数据来源。四相机能够带来两个清晰的触点，除了一种特殊情况，即当两个触点都位于相机间的一条对角线上时，两个相机看到的是被遮挡的物体。

前面描述了由红外 LED 在屏幕角落生成的背光被反射的系统。这被称作“被动”背光系统，因为相机感应到的阴影是由反射光生成的。然而背光也可“主动”，也就是说，它可以直接发射光。构建主动背光主要有两种方法。一种常用的方法是把大量的红外发射 LED 装置在触摸屏的四周边缘，这些可以直接发射被触碰物体阻挡的光。这种方法的主要优势是更高强度的光生成更高的触控系统信噪比，从而增强触控功能的稳定性。该方法的主要缺点是元件的增加成本和周围的印制电路板。

在第二种方法（由于知识产权的问题仅由 Lumio 使用过）中，呈管状分布在屏幕周边的光导重新指引了红外光的方向，该红外光由位于每个波导段末端的 LED 生成，因此得以分散至屏幕表面。该方法的主要优势是较低的成本和非常低的横截面高度（3～4mm 对比以往的 6～10mm）；主要的劣势则是其较低的光强度和单一来源的本质。

2013 年的摄像光学触控应用主要在以下两个主要领域发展：

1）桌面多合一触控电脑和触控显示器。

2）大屏交互信息系统、数码广告牌、会议和培训室，以及在教育应用中使用大型交互 LCD 替代白板书写。

桌面应用系统的发展主要由于微软的 Windows 7 Touch Logo 的规范是基于摄像光学编写的，它在当时是成本最低的可支持双点触控的技术。而 Windows 8 Touch Logo 的规范则是围绕投射电容编写的，具有最低五个同时触点的要求。NextWindow 已经能够通过使用六个相机来满足 Windows 8 Touch Logo 的规范——每个角落各一个，另外两个在三等分屏幕的顶边。目前，没有其他摄像光学供应商（除了 SMART Technologies，即 NextWindow 的母公司，仅在出售完整的系统而不是触摸屏）声称拥有可以满足 Windows 8 Touch Logo 规范的产品。

由于投射电容式在桌面尺寸（15～30in）的成本较高，摄像光学被看作是 Windows 8 消费产品的理想替代品。但是 PC OEM/ODM 总体上更青睐多元资源，这样直到有除了NextWindow之外的供应商出现，否则让摄像光学进入 Windows 8 桌面产品的可能性将十分有限。

摄像光学技术在大屏显示器应用方面的主要对手是传统红外线；其他的竞争者还包括有线投射电容、表面声波和 3M 出品的色散信号技术（后两项技术局限在尺寸约为 52in 内）。摄像光学比起传统红外的主要优势是其可延展性，这有助于它用在更大触摸屏时仍然保持较低成本。任何尺寸的传统红外触摸屏必须要在屏幕周边布满印制电路板，但是摄像光学触摸屏只要使用可连接到塑料或金属托架上的印制逆反射器即可。后者的成本要低很多。摄像光学技术的另一个优势是其高于传统红外的分辨率和速度。

在大画幅应用中，光学触控和传统红外触控均有各自优势，因此两者皆可能在今后数年继续出现在大画幅市场中。随着时间的推移，摄像光学技术将取代传统红外，因为其硬件要更简单，而更多的性能可通过软件增加。摄像光学触控技术的优缺点在表 2.13 中做了总结。

表 2.13　摄像光学触控技术的优缺点

优点	缺点
可延展至很大尺寸（达 120in）	剖面高度（相机模组投射在接触表面）；随屏幕扩大而增加
触碰力为零的任一阻挡红外光物体可以激活（无需触控笔）	预触碰（在实际接触到屏幕表面时触控已经激活），但没有传统红外严重
多点触控（通常 2～5 个触点，但使用 20 个相机已经可达 40 个触点）	需要相对高的基板平坦度（如 ±2mm），特别当反射器作为背光源时
相对高的分辨率和准确度	很难防止意外碰触（衣袖、昆虫和屏幕上的残留物等）
物体大小识别（从多个有利位置捕捉每个接触物体的图像的结果）	多点触控性能由于较少的数据来源客观低于投射电容；带双相机的两点触控性能非常低
由于平面玻璃基板而具有高光学性能	屏幕各处的准确性不统一，特别在仅有双相机的时候
比传统或多点触控红外技术成本低	最小接触物体尺寸取决于屏幕尺寸；针对大屏可以增至 7mm
有无玻璃基板均可	对环境红外比对传统红外更敏感（缺乏解决方案）
	屏幕相邻放置会相互影响，由于红外光可以被互相探测到

2.7.3.1　光电二极管光学技术

光电二极管光学技术是摄像光学技术的一个分支，由于知识产权的考虑，目前仅在 Baanto 公司可以找到。该光学触控技术使用光电二极管（p－i－n 半导体结构的二极管）作为光传感器，而不是前面描述的 CMOS 相机。光电二极管具有以下可以简化光学触摸屏结构

的特征[53]：

- 光电二极管直接读取光强度并在高达 10000 帧/s 的速度下运行，支持了高性能的触碰系统。
- 光电二极管无需透镜，而且可装配接近 180°的视场（FOV）（见图 2.30），因此可以无需依靠装置在触摸屏角落的传感器并消除需要电脑计算纠正的光学像差。

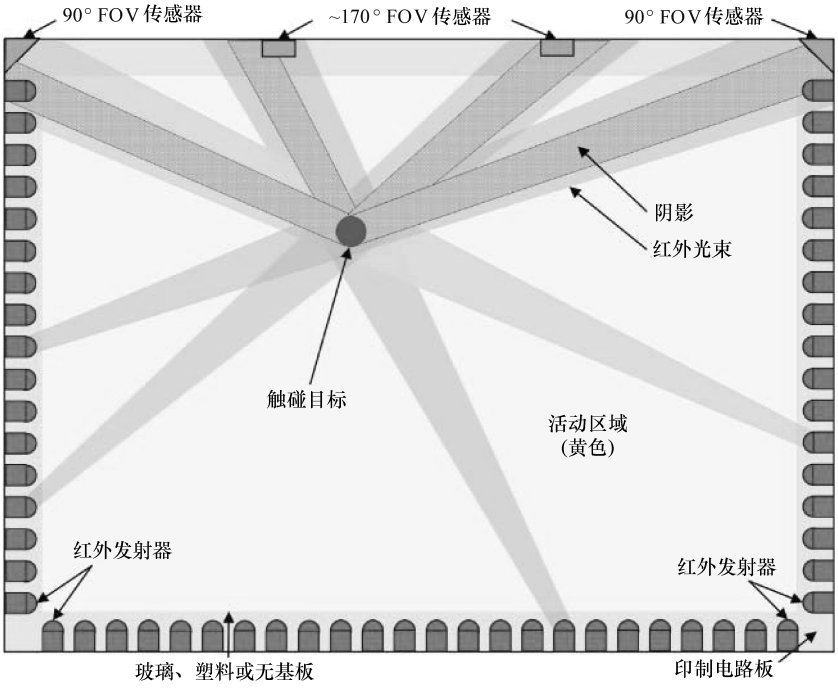


图 2.30 光电二极管光学触摸屏的一个例子，该技术在角落使用了两个 90°的视场传感器，并在顶边使用两个 170°的视场传感器。触摸屏另外的三边含有红外 LED。光束被吸引集中，以显示一个传感器能观察到两个接触物体（取决于该物体的大小）的阴影和边缘光。来源：改编自 Baanto 公司

- 光电二极管无需曝光控制，因为传感器性能不会因为接触物体的距离或速度而受到光照变化影响。
- 光电二极管的场深无限，意味着位置探测算法不会随着接触物体的位移而发生改变。

光电二极管完全在模拟领域运行，比起更为数码导向的图像处理技术，它更能让触摸屏控制器以不同的方式使用接触物体的阴影信息（Baanto 的技术被称为“阴影感知”）。它包含的一些在其他种类的摄像光学触摸屏中不多见的性能包括以下方面[54]：

- 可选择的接触面积（提供防误触，忽略雨滴到屏幕上，或设置要求的最小手指接触压力）。
- 可选择的“驻留时间”（在报告有效触碰之前，一个触碰物体必须停留在屏幕的最少帧值，这能实现排除短暂的意外触碰）。

- 可选择的阴影密度（使接触物体的红外穿透性成为判断有效碰触的标准）。
- 更好地排除高度的环境红外干扰（达 100k lx）。
- 更容易上调尺寸（Baanto 至今的最大触摸屏是在一面 266in 的视频墙上）。

Baanto 的光电二极管光学触摸屏的背光是主动的，它使用了 940nm 的红外 LED，相互间隔 5mm 并环绕在触摸屏的边缘。触摸屏使用完全照明、部分遮挡、完全遮挡的触摸事件之间的比例使屏幕能够耐受周边变化的 LED 功率值，从而无需使用分档的（匹配的）LED。另外，因为控制器算法使用读数之间的比例，传感器接收到的总功率变化不会影响对触碰位置的计算。

2.7.4　玻璃光学触控技术（平面散射检测）（编号 12）

平面散射检测（PSD）是一种玻璃光学触控的特别形态。该技术由一家 2007 年初期创立的瑞典公司 FlatFrog 发明，于 2012 年 5 月首度发布了该产品。FlatFrog 的触控技术被称为“光学波导分析”。所谓“波导”是一块接触基板，可以是任何一种尺寸稳定的透明材料，且无硬度和平坦度的要求，这种特性在光学触控系统中十分少见。FlatFrog 系统的基本工作原理如图 2.31 所示。在一个 PSD 触碰传感器中，光由多个红外 LED 射入光学基板边缘并由全内反射（TIR）限制在基板以内。由于受抑全内反射（FTIR），触碰分散了一部分光；多个与基板边缘的 LED 交错的红外光检测器探测到剩余的（减弱的强度）TIR 光。复杂的算法经过分析光射线强度和进行 1D 到 2D 的重构，即可决定表面所有物体的位置[55]。

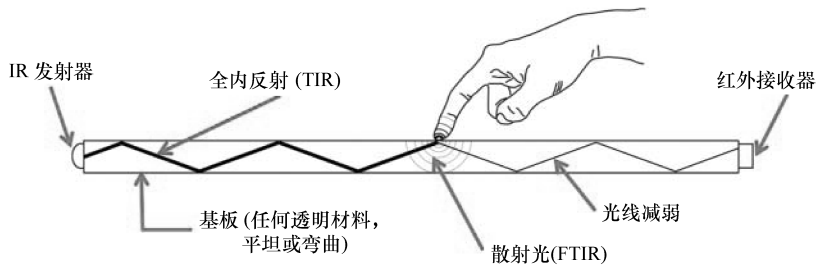


图 2.31　该图是一个解析平面散射检测（PSD）的原理图。红外光被射入基板（波导）；随后由于 TIR 保留限定在基板内。由于 FITR 使触碰分散了一部分光；强度变弱的光持续存在，直到基板对面边缘的光检测器检测到。经过分析全部光射线的强度，使用复杂的算法即可计算出碰触位置。来源：改编自 FlatFrog

PSD 不同于标准 FTIR 的一个方面是 FTIR 光在基板内得到分析而无需像视觉触控需要其离开基板。另一个不同是一次触碰仅消耗给定光射线的一小部分，这样多触点可以在直光下被检测到，剩余足够的光仍然可以在边缘被感知。和传统红外光一样，PSD 要求基板四边都要装置电路板，但是红外发射器和接收器的数量比传统红外触摸屏的要求要少些（7～8mm 的元件间隔距离相较于以往的 5mm）。不同于传统红外光，PSD 触摸屏有完全齐平的包边，因为显示器屏幕上没有任何投射。

唯一一个 2013 年发布的 PSD 产品是一个由 FlatFrog 组装的 32in LCD 触控显示器。换句

话说，目前该触控技术并未作为组件提供。但是 Intel Capital 已经注资 FlatFrog，并与 FlatFrog 直接合作将该技术市场化。有可能到 2014 年前，PSD 触控技术将成为多合一电脑（如 23in）显示器的组件之一。FlatFrog 预计在大规模电子消费产品中使用许可证商业模式，并在小规模大型商业屏幕中使用产品销售商业模式。PSD 可能成为传统红外和摄像光学大画幅技术的强劲对手，甚至对笔记本电脑大小以上的投射电容技术构成威胁。PSD 触控技术的优缺点见表 2.14。

表 2.14 玻璃光学（PSD）触控技术的优缺点

优点	缺点
十分稳健的多点触控（32in 屏幕上 40 多个触点，所有触控技术中与投射电容的用户体验最为接近）	红外发射器和接收器需要周边布置印制电路板（9mm 宽），每 12 对组件还需要一个 ASIC 驱动器
边到边（无边框）或含边框（类似投射电容）	由于缺乏悬浮不能满足 Windows 数字笔界面的规范，不适合触控笔应用
实际大小范围在 14～84in 之间（优于投射电容）	软物体意外触碰可能导致 FTIR
高分辨率（400dpi）和准确度，满足 Windows 8 的需求（等同于投射电容）	对外部红外光敏感（未来可改进）；由于 FTIR 改变，触控表面的尘土或烟雾会影响性能
平面玻璃或塑料基板产生高光学性能（优于投射电容）	新兴触控技术；价格竞争优势和产品规模有待证实
很轻的触碰（类似投射电容）；随着压力增加，手指的光学性质发生改变，导致 10bit 压敏的产生	截至 2013 年，直接代替投射电容需要在玻璃盖片和 LCD 之间留出更多空间（3mm）且不能直接键合；直接键合需要 PCB 部件安装在 LCD 边框之外，这增加了屏幕边框宽度
可用手指、手套、被动软尖触控笔（任何能导致 FTIR 的软物体）	
对电磁干扰/射频干扰不敏感（优于投射电容）	
成本低于桌面尺寸投射电容；只有 ASIC 驱动器和固件/软件是独特组件	

2.7.5 视觉光学触控技术（编号 13）

此处视觉光学触控技术是指使用“电脑视觉”检测和处理接触平面所发生的触控。虽然同样的术语也用于（可能更经常）描述经过 2D 或 3D 摄像头检测并处理的手势命令，但是因为后者并不包括接触显示器，本章对此技术不作讨论。“电脑视觉”也暗指大量使用图像分析软件以判断碰触位置和其他接触屏幕表面的信息。

2.7.5.1 投影

产生视觉触控的方法目前有三种：

1）投影。

2）LCD 后方的多个广角相机。

3）内嵌光感。

视觉触控使用的投影方法通常是背面投影，即在投影仪旁边安装相机（见图 2.32）。受抑全内反射（FTIR）（见图 2.32）是最常用的产生由触摸投影表面导致的红外光“光团”（明亮发光物体）的方法[56]。2007 年发布的 Microsoft Surface 1.0（以及随后四年出现的许

多相似的“触控桌”）是背面投影视觉触控的最佳范例。该方法的主要优势是系统可以以非常低的成本组装[57]；主要劣势是背面投影系统的实际大小，以及由于背面投影导致的相对较低的图像质量。

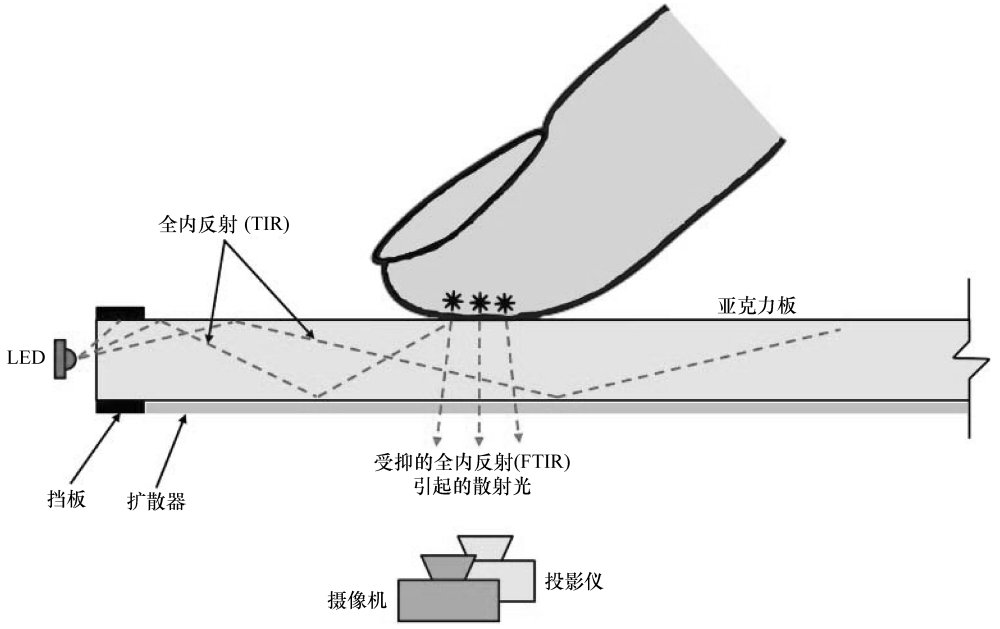


图 2.32　最常用的产生视觉触控的方法是使用背面投影，这需要在投影仪旁边安装相机。FTIR 最常用来产生由触摸投影表面导致的光。来源：改编自 Perceptive Pixel

除了投影视觉触控法能生成红外光团以外，还有其他三种方法：

1）扩散照明（DI）。

2）激光平面（LLP）。

3）扩散表面照明（DSI）[58]。

扩散照明指在触碰表面的背面均匀散布红外光。这通常依靠与屏幕间隔一定距离安装的一个或多个红外发射器实现。Microsoft Surface 1.0 就使用了这个方法。激光平面指使用激光在触摸屏上方生成一层很薄的（1mm）红外光平面；当手指打破这个平面时，红外光团即产生。通常，在屏幕各角上会装置两个或四个激光发射器；每个发射器上装有一个呈 120° 角的线路滤波器以扩散光束。扩散表面照明指使用一种特殊的丙烯酸纤维在表面均匀分散红外光。丙烯酸纤维含有小反射颗粒；当红外 LED 光射入纤维边缘时，这些颗粒把光反射分散到纤维表面。该效果与扩散照明相似，但有着更强的均匀性。

2.7.5.2　集成相机

当前唯一一款集成多广角相机的 LCD 的视觉触控产品是来自芬兰 MultiTouch 公司的 MultiTaction™（见图 2.33）。在该款触控显示产品中，相机被集成到 LCD 的背光中。这种方法的主要优势是相比投影的屏幕更薄（8in）以及性能更高。主要的劣势是成本、复杂性

和厚度。MultiTaction 产品的一些更有价值的性能如下[59]：

- 不受外界光环境的影响（通过识别环境光和嵌入在背光中的红外反射器发出的光）。
- 不受限的触点数和用户数（触摸屏软件同样识别手，而不仅仅是触点）。
- 运用2D 标记和/或总体形状识别来辨认物体。
- 用红外发射触控笔工作（清晰区分手指和触控笔）。
- 模块化的触摸屏显示可以形成多用户交互墙。

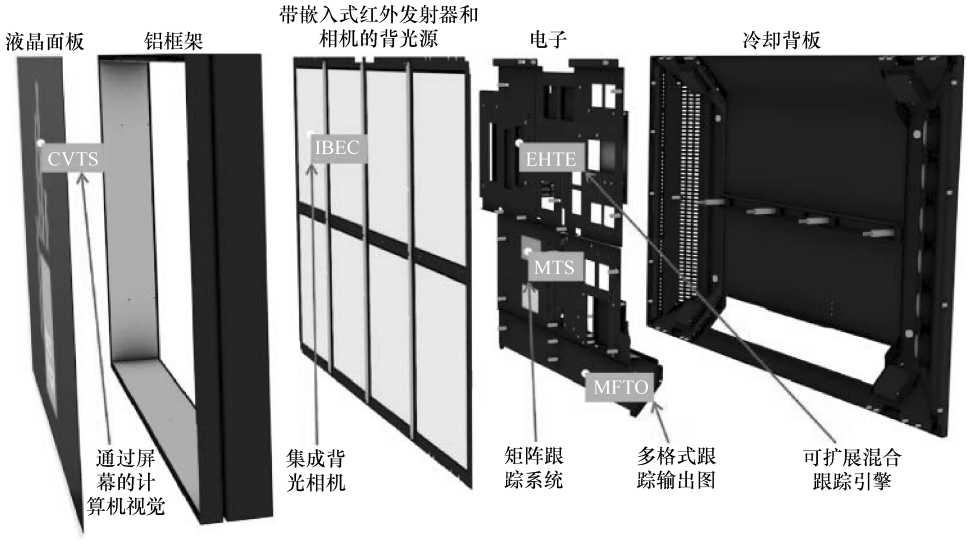


图 2.33 来自芬兰 MultiTouch 公司的 MultiTaction 视觉触控系统把红外发射器和红外相机融入了标准 LCD 背光中，并由嵌入的处理器运行复杂的物体追踪软件来支持该系统。来源：改编自 MultiTouch

2.7.5.3 嵌入式内嵌光感技术

嵌入式内嵌光感技术将在随后的“嵌入式触控技术”部分进行更为细致的探讨。运用于 Microsoft Surface 2.0 中的三星 SUR40 触摸屏是目前唯一一个使用嵌入式内嵌光感技术来实现视觉触控的市场化产品（Surface 2.0 在 2011 年发布，并在 2012 年由 Microsoft PixelSense 重新命名，为了使其名“Surface”可以在平板电脑市场使用）。嵌入式内嵌光感技术的主要优点是触控完全与显示器实现一体化而无需增加任何厚度。该技术（在 SUR40 中实现）的最大缺点在于对环境红外的极端敏感性（以至于 SUR40 无法使用在大多数光亮的环境中[60]），以及由于需要更长时间处理嵌入式光传感器的数据而导致触摸延迟性。

2.7.5.4 视觉触控技术小结

视觉触控的应用主要可以分成两类：

1）由于多触点的高性能和自制的低成本，许多高校建立了研究触控技术的公用平台[57]。

2）一个传统商业应用的新平台，如零售店内的多样化触控桌产品和公共场所的交互视频墙。

虽然视觉触控技术的发展仍处于萌芽期，但是至少可以说它已经是一项拥有专营市场的技术，因为它并不直接与其他触控技术竞争。技术的“视觉”特征使其可以完成其他触控技术无法实现的任务。一个简单的例子是通过使用赋予物体的图形记号来识别物体[61]。举例来说，它可以使带有记号的数码相机或智能手机放置在触摸屏上，应用软件自动通过蓝牙把设备中的图片下载下来并显示在屏幕上供编辑或排列——无须用户发出任何指令。视觉触控技术的优缺点详见表 2.15。

表 2.15　视觉触控技术的优缺点

优点	缺点
用于图像处理软件分析的理想数据来源	无法作为元件获取，仅为大型显示器的一个系统存在
通过使用图形记号进行物体识别（不仅是大小或形状识别）	由于缺乏悬浮无法达到 Windows 数码笔界面规范要求，不太适用于触控笔应用
多触点十分稳健	投影方法需要不少空间，且光学性能低
非常轻微的触碰	集成相机法的成本、复杂性和相对厚度
可使用手指、手套或被动软头触控笔（投影法）；可使用发光触控笔（内置相机和嵌入式方法）	嵌入式方法对环境红外光敏感
自制成本低（投影法）	作为新兴触控技术，已有的应用数量有限
通过多个集成相机和强大的嵌入式图像处理能力而实现的复杂功能（MultiTaction）	
嵌入式触控（SUR40）的标准优势	

2.8　嵌入式触控技术

如“投射电容式触控传感器”部分所述，术语“嵌入式”指的是由显示器制造商集成在显示器中的触控功能，而“分离式”则意味着触控功能是独立于显示器之外制造的。谁提供了触控功能实际上是决定嵌入式触摸屏的根本因素，而并非是技术本身的细节。在决定一个 OEM/ODM 设备是嵌入式还是分离式的触摸屏产品时，商业问题往往比技术问题更为重要。比如对于智能手机，一个嵌入式触摸屏和一个分离式 OGS（单片触控面板）触摸屏的技术差别实际并不大，详见如下所述[62]：

- 嵌入式触控智能手机显示屏通常比分离式的同款显示屏薄 100～150μm。由于使用嵌入式触控的智能手机模型的厚度变化约 1.0mm，100～150μm 的差别对大多数用户来说并不显著。
- 嵌入式和分离式触控性能大致相同。一些显示器制造商仍在迎头赶上，但长久来看这些性能是趋同的。
- 嵌入式和分离式触控的重量相同，因为两者都使用三块玻璃板（屏幕两块，还有一块玻璃盖板）。
- 嵌入式和分离式触控的功耗大致相同。随着时间的推移，带有嵌入式触控的更高效

的集成应该能使其功耗更低。

- 嵌入式和分离式触控的成本惊人的相似。因为嵌入式触控可以实现更高效的集成，目前在智能手机控制器和排线上可能节省 2～4 美元。随着嵌入式触控拓展到平板电脑大小的显示器中，分离式触控由于替换了 ITO 成本实际上更低，取代的材料可以是金属网。
- 屏外触点图标（比如安卓智能手机里的“菜单”图标）可以用分离式触控轻易创建，因为保护玻璃层总是比实际显示区域要大。但是嵌入式触控必须使用额外的部件（如虚拟按键）来获得屏外触点图标。

嵌入式触控技术发展了至少 10 年之久。触控新方法仍在不断探索，并不断地并入 LCD 的组成部分，影响着 LCD 的设计与应用[63]。主要的触控方法如下：

- “指压电容”（又称“电荷传感”，首次由三星批量生产）的原理是指压产生的屏幕压力导致液晶的介质常数发生改变。变化的介质常数改变了增加到部分或全部像素中的电极对之间的电容[64,65]。
- “光感”，首次由夏普批量生产，指红外光检测器被并入部分或全部像素中。光检测器既可以在强环境光下读取触碰物体的阴影，也可以在弱环境光或黑暗中读取触碰物体的折射背光[66]。
- “电压传感”（又称“数字转换”，由三星首创），指 X 和 Y 位置的微动开关被并入部分或全部的像素中。屏幕上受到的压力关闭了微动开关，从而定位了压力源[67]。

这些方法中没有哪个是完美的，虽然三星确实推出了数百万个使用指压电容的傻瓜数码相机。它们未能获取完全成功的主要原因如下：

1）信噪比不足，无法实现稳定操作。

2）要求屏幕表面实际上发生弯曲（这样消除了使用玻璃保护层的可能性，因此增加了屏幕的易损坏性）。

3）把屏幕尽可能的压近框架是不可靠的，因为彩色滤光片几乎无法移动。

随着投射电容的互电容广泛用于智能手机分离式触控技术中，显示器行业愈发意识到把投射电容集成到显示器中应用是一个正确的决定，而不是无谓的进行重复劳动。表 2.16 总结了三个集成投射电容的嵌入式触控方法。

表 2.16 外嵌、内嵌和混合嵌入式触控技术以及向市场大批量投放每类技术产品的公司

方法	定义	首次发布
外嵌（on－cell）	触控传感器是在 LCD 的彩色滤光片或 OLED 的封装玻璃上表面的 ITO 电极阵列；功能同标准投射电容一样	三星，2010 年 OLED 智能手机
混合内外嵌（in－cell/on cell）	触控传感器由 ITO 电极阵列组成，其中感应电极层在彩色滤光片上面（外挂），驱动电极在 LCD 面板内。驱动电极可在 TFT 玻璃上（在 IPS（共面转换）的 LCD 内）或在彩色滤光片的底面（在非 IPS 的 LCD 内）	索尼，2012 年索尼和 HTC 智能手机
内嵌（in－cell）	触控传感器位于 LCD 显示面板内部（夹层在 TFT 和彩色滤光片之间）。传感器可以是 ITO 电极阵列（互电容式）或光感元素	苹果，2012 年 iPhone 5（电容）；夏普，2009 年上网本（光感）

从出货量看，在2010年2月，三星公司是第一个大批量在消费市场发布面板上外嵌式触控的生产商，并将其应用于S8500 Super AMOLDED™智能手机（“Super AMOLED”是三星自有的主动矩阵OLED的品牌冠名，该产品应用了外嵌式触控技术）。2012年5月，索尼成为首个发布混合内外嵌式触控技术的公司，XperiaP™和HTC EVO Design 4G™智能手机均使用了该项技术；Synaptics为这些产品开发了触摸控制器[68]。2012年9月，苹果公司首次在iPhone 5中使用了内嵌式触控。从发明（专利）的角度说，是应该把所有形式的电容嵌入式触控的发明归功于一家公司（如苹果），还是应该把每种类型的电容嵌入式触控分别归功于某个公司——这还有待观察。

2013年，嵌入式触控主要应用于智能手机（小）屏幕，原因是该技术在较大屏幕中的应用仍然在开发中。到2015年以前，该技术有望升级到笔记本电脑大小的（15in）显示器中。升级面临的主要问题如下：

- 更大的屏幕具有更多的电极，因为感应电极和驱动电极的数量是实际屏幕尺寸的一个函数，而不是屏幕的像素分辨率。电极也会更长。这两个因素增加了完成完整触摸屏扫描所需的时间。
- 更大的屏幕往往有更高的像素分辨率，这缩短了显示屏电子环境的安静时间。嵌入式触控感应通常需要在这些安静的时间内完成（实际上，为了优化时间，触摸控制器和显示控制器经常协同运行）。

这两个问题结合起来——每次扫描需要更多时间，而用于扫描的时间变得更少——一直妨碍着嵌入式触控升级到7in以上的批量产品中，实验室内使用的限制是12in。

2.8.1　外嵌互电容式（编号14）

外嵌互电容是概念上最简形式的嵌入式触控。投射电容式触摸屏并非安装在独立的玻璃基板上或在保护玻璃下，而是安装在LCD彩色滤光片或OLED密封玻璃上方（见“投射电容式触控传感器”部分的图2.11）。各层叠加的功能本质上与分离式投射电容是相同的。外嵌式最普遍电极阵列分布是连锁菱形，因为它能与金属连接线安装在一个阵列基板上。

如前所述，首个批量生产的内嵌式触控产品是一台OLED智能手机。外嵌式触控用于OLED其实比用于LCD更简单（产量也更高），因为OLED封装玻璃下并没有任何装置。而LCD的彩色滤光片下至少有彩色过滤材料；如果是非共面转换（IPS）的LCD，其上面也会有普通电压（Vcom）电极（由ITO形成）。对一个显示屏生产商来说，问题就变成了要先制造哪一边。如果先制造触控面板，高温煅烧就可以使用，从而可以改进ITO的质量，提高触摸屏的性能。但是玻璃层在密封LCD面板之后就无法保持平时的薄度了，可能使显示器增厚约0.3mm。如果先制造彩色滤光片，那么玻璃层可以变薄，但是触摸屏无法高温煅烧（若这样做就会损坏彩色过滤材料），导致触摸屏性能更低[69]。制造商通常选择后者，因为不会干扰LCD制造环节的生产，而且薄度总是被视为一个极其重要的因素。

2.8.2　混合互电容式（编号15）

顾名思义，混合嵌入式触控是指触控面板一半嵌入LCD，一半在LCD外。感应电极存

储在 LCD 彩色滤光片上（外挂）。有些 IPS 显示屏在彩色滤光片上有用均匀 ITO 涂层制成的抗静电隔离层；如存在，则该层以长条形电极平行排列而成。该层仍起屏蔽层作用，因为除了其中一个条形电极会随时感应之外，其他均接地。这种防静电屏蔽的方法正逐渐被位于显示屏顶部偏光器内的导电层所取代。该情况下，感应电极就作为新的一层被置于彩色滤光片上。

驱动电极是通过分组和改变 LCD 的 Vcom 电极用途而产生的，是触摸屏面板的一部分，也具有普通的升级显示屏功能[70]。在 IPS 显示屏中，这些电极位于 TFT 层上。在非 IPS 显示器里，这些电极位于彩色滤光片的底下。由于两个表层仅距离几微米，两种结构在性能上没有差别。组合形成单一嵌入式投射电容驱动电极的 Vcom 电极数量取决于像素分辨率、屏幕大小以及合适的电极间隔。

比如，一个 7in、1280×720 的 LCD 有 155mm×88mm 的活动区域。如果条件是嵌入式投射电容式驱动电极要在显示器的水平长边形成，且最好有约 4.8mm 的电极间隔，那么把 1280 像素分成 32 组就能产生由 40 个接地 Vcom 电极构成的投射电容式驱动电极，间隔略多于 4.8mm。感应电极（在彩色滤光片上，如前面解释）会沿显示屏短边垂直运行，相同的间隔下产生 18 个感应电极。图 2.34 展示了一个混合嵌入式结构，使用了 Japan Display（索尼公司的前身）的“像素眼”[71]作为例子。该图展示了电极的物理分布和叠加层（不按比例）。

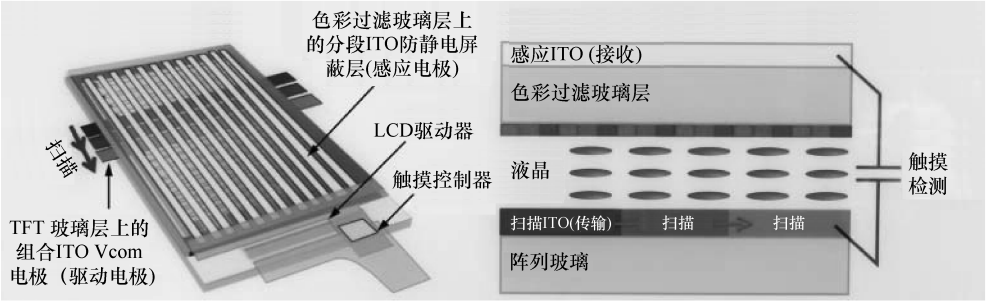


图 2.34 Japan Display 的产品“像素眼”混合嵌入式触摸屏结构。右图展示了嵌入式 TFT 层的驱动电极，同时感应电极在彩色滤光片上。左图展示了更多物理结构信息，说明了感应电极是通过切分 ITO 防静电屏蔽层形成的，而驱动电极则是由分组 Vcom 电极形成的。该图也展示了包括显示屏和触摸控制器的 FPC 以及与其连接的两组电极。来源：Japan Display，图解由作者标注

以上描述实际只是数种混合嵌入式方案中的一种。另一种方式，与苹果、三星的专利描述一样，是在黑色矩阵上（在彩色滤光片的底面）覆盖金属作为驱动电极，而不是在 TFT 层上进行设置。

有了外嵌式触控，使用无需连接 LCD 的标准触摸控制器 ASIC 的可能性是很大的。毕竟外嵌式触控电极只是稍微比 OGS 一体化触控模组更接近 LCD 一些，因此增大的 LCD 噪声并不是一个大问题。然而，一旦部分触控系统内嵌 LCD，触控系统和 LCD 的同步化对解决噪声问题来说就十分必要了。

2.8.3　内嵌互电容式（编号 16）

顾名思义，内嵌式（in－cell）触控是指触控系统完全嵌入 LCD 内部。前面指出，2012 年 9 月，苹果公司是首个在其发布的 iPhone 5 智能手机中批量使用嵌入式触控的公司。iPhone 5 的配置将驱动电极和感应电极放在了 IPS LCD 的 TFT 层上。这是使用与前述相同的分组和改变 Vcom 电极的基本技术来实现的，除了其 Vcom 电极是被分成了两组——一组驱动和一组感应。实际操作比听起来要复杂得多。

图 2.35 展示了分组的过程是如何进行的。该图由 BOE Technology Group 的中心研究院发布，旨在以浅显易懂的方式解释苹果公司的专利[72]（但是要注意，该图的视角不大准确；图示中各正方形实际应该是 54 像素高 ×126 像素宽，因此它们应该是长方形）。如图所示，标有 TX 的行是数组横向（X）切分的 Vcom 电极，由 ITO 组成并由金属栅（黑色）连接。每行都通过触控面板金属（记号为“TP”）与触摸控制器相连。因为该图显示屏有 1080 像素高且有 20 个驱动电极（行），所以每组包含 54 个 Vcom 电极。较宽的列是垂直的（Y）通过触敏探测金属（记号为“S/D”）相连的 Vcom 电极。每 10 列也包含 54 个 Vcom 电极。从触摸屏的角度看，驱动电极和感应电极是对称的。较宽列两边的窄列是电气连接，但彼此独立的仿制 ITO（这种仿制 ITO 常常用在触控电极中以实现更为整齐的外观）。

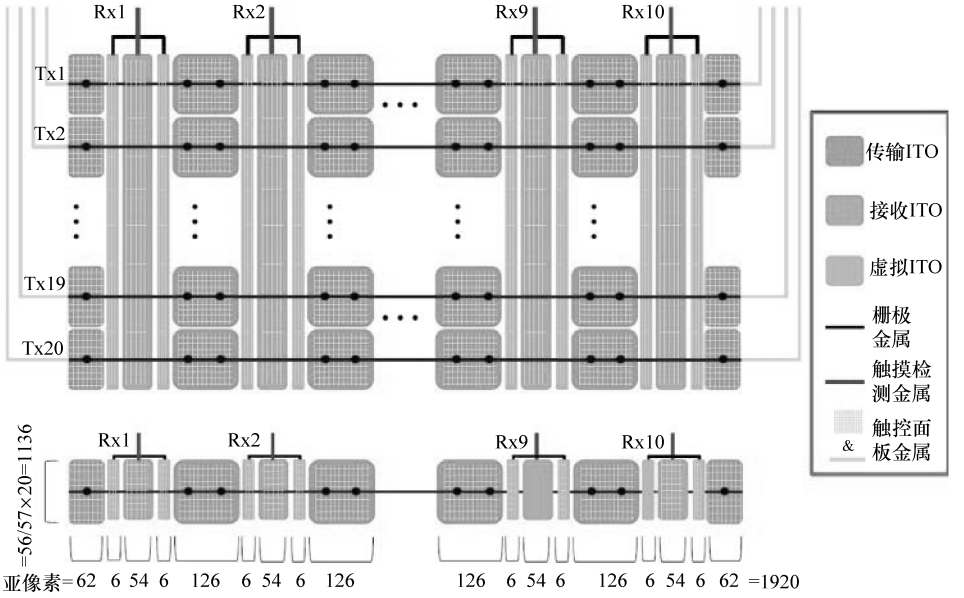


图 2.35　该示意图展示了 iPhone 5 内 TFT 阵列中的 Vcom 电极是如何分组形成驱动电极和感应电极的。来源：改编自 BOE Technology Group 中心研究院

很重要的一点是，苹果公司 iPhone 5 使用的技术仅是数种不同内嵌式触控方案的其中一种。比如，苹果和三星的专利描述了一种通过切分 Vcom 来获取驱动电极的方法，但是感应

电极是由覆盖在黑色矩阵上的（在彩色滤光片的底部）金属制成的。这两家公司在其专利中描述了另一种方法，其中黑色矩阵上的金属作为感应电极（如前面所述），而驱动电极则是存放在介电层上的 ITO 线条，介电层下面是彩色过滤材料。夏普公司在 2012 年的一次显示系统会议上描述了一种将两种触控电极都存放在彩色滤光片底部的方法，这在黑色矩阵和彩色过滤材料应用之前。LG Displays 在它的其中一份专利文件中描述了一个仅运用切分的 Vcom 电极的自电容法。

有两种方法可以使触摸控制器和屏幕控制器（TCON）同步：稍微调整两个控制器并通过数根导线将其连接；或把两个控制器并入一个芯片中 。Synaptics 是第一个实验了两种方法的触摸控制器公司。第二种方法的主要优点是减少了一两美元的触控系统的材料清单（BOM）成本，但是因为芯片的开发增加了一次性工程费用（NRE）成本；主要的缺点是合并的控制器只能使用于特定的显示分辨率和像素组成。显然，第二种方法仅对大规模产品（至少几百万）有实际意义。

电容嵌入式触控技术的优缺点总结见表 2.17。

表 2.17 电容嵌入式触控技术的优缺点

优点	缺点
大多数投射电容的优点（如果在控制器算法中正确应用，可实现稳健的多点触控；非常轻的触控；允许齐平包边保护玻璃；出色的视觉性能；可密封；等等。）	内嵌式和混合式仅对大规模的显示器（数百万）有实际意义；外嵌式可以有较少的产量，但是它可能会减少 LCD 的制造量
对产品的每个新实现来说，“参数调整”的需要比分离式电容少	目前无法升级到 12in；也许永远无法像分离式一样可升级
比分离式电容（OGS）的成本低，但是差别并不显著，而且绝非“免费”；在 OGS 开始使用 ITO 替换材料后，差别可能减小	更难以实现与分离式相同的触控性能；显示器制造商也许会比改进触控性能更重视提高产量
比分离式更薄（通常 100～150μm）	控制器制造商可能会较慢地把投射电容的改进方案（如抗水性、活跃触控笔等）引到嵌入式方案中
功耗可能比分离式稍微低些，特别是在集成了触摸和显示控制器的情况下	显示器制造商也许不愿像分离式触控面板制造商那样生产多种保护玻璃，或愿意做直接结合
有机会减少触碰延时，特别是在集成了触摸和显示控制器的情况下	没有额外元件无法支持屏外图标（不像分离式）
	可能无法使用极厚的玻璃基板
	并非绝对的压力传感；只有相对的手指接触区域（与投射电容相同）

2.8.4 内嵌式光感（编号 17）

内嵌式光感触控技术是通过在部分或全部 LCD 像素中增加光检测器而实现的（见图 2.36）。正如表 2.16 所示，夏普是第一个在 2009 年 5 月大规模地在消费产品中使用了内嵌式光感技术的公司；该产品是一台带有显示式触摸板的上网本。其显示器采用了 4in 大小的 LCD，运用了夏普 854×480 像素（245ppi）的连续晶粒（CG）硅技术。一开始夏普想尝试

每个像素一个传感器，以使显示屏同时作为扫描仪，但发现这样做会过多降低显示的孔径比。于是他们使用每九个像素一个传感器，这样产生了只有 27ppi 的扫描分辨率——对实际操作来说不够高。即使是每九个像素一个传感器的比率，夏普仍发现处理光传感器的输出需要更多的 CPU 带宽，超过了触摸控制器可以提供的范围，因此该显示式触摸板的性能变得出乎意料的低（仅为一个普通的电容触摸板速度的 25%）。

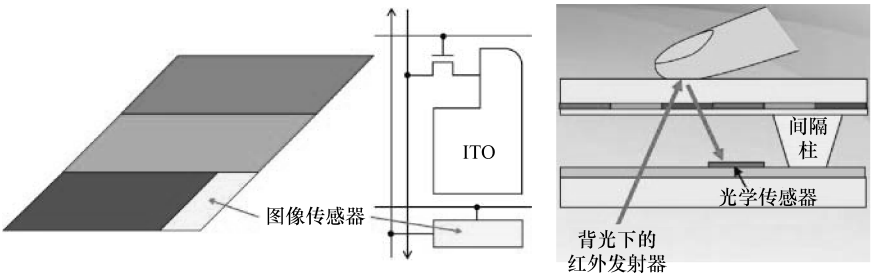


图 2.36　内嵌式光感触控的概念图。光传感器是一个 aSi 光检测器；它放置在蓝色（最低）亚像素中，因为传感器的最大灵敏度是在蓝 - 绿色频谱范围内。右图展示了红外光从红外 LED 射出，背光由手指触碰屏幕反射并由光传感器感知。来源：改编自参考文献［4］（左）和三星（右）

该领域的研究在 2000 ~ 2005 年主要集中在可见光传感器，预计该传感器能够在光亮环境中察觉触碰物体的影子，或在昏暗外光下察觉触碰物体的背光折射[74,75]。在 2006 年研究人员意识到可见光的局限性之后（比如，它无法穿透 LCD 上的黑色图像），他们转向了红外光传感器。这一改变也意味着红外发射器需要加入到背光中。因为 LCD 只是对红外光适度透明，而且光线在两次穿过 LCD 的过程中有所削弱，因此红外发射器必须有相对高强度的发射。在夏普显示器中，红外 LED 的额外功耗严重降低了上网本的电池续航能力。

从 2009 年夏普的上网本到今天（2013 年），仅有两款商业产品使用了光感嵌入式触控。第一个是由中国台湾 IDTI（集成数字技术公司）开发的 21in 显示器[76]。开始的设计中只有一支光笔，后来的版本则改进为支持手指触控，通过可见光阴影或折射法实现（如前一段提及）。

第二个产品是由三星联手微软为 Microsoft Surface 2.0 而开发的 SUR40 40in 多合一桌面电脑[77]。该款 LCD 包括背光内红外发射器，并实现每八个像素使用一个红外光传感器。为了改进触控系统的敏感性，三星运用了 aSiGe（非晶硅锗）光传感器，它们比普通的硅光传感器敏感 15 倍。虽然这绝对改善了触摸灵敏度，但它也带来了一个新问题：对环境红外有极度的敏感性。该现象太过于严重，致使三星发布了一份手册，记录了在每种室内光下触控系统可以容纳的最大光照度；白炽灯的数值仅为 50lx[60]。该产品包括一个测量环境红外光和显示器红 - 黄 - 绿区域的应用程序，以指示是否光照度达到足够低的程度。

三星目前正在研究为 OLED 显示器所用的光感嵌入式触控[78]。该概念与 LCD 描述相对应，指在 OLED 面板中嵌入红外发射像素，同时在主动矩阵底板中装上红外探测传感器。

虽然对内嵌式光感触控的研发已经进行了 10 多年之久，这仍是一项尚未攻克的新兴技

术，至今尚未有大量的消费产品成功发布。表2.18提供了对内嵌式光感触控技术更为详细的优缺点对比。

表2.18 内嵌式光感触控技术的优缺点

优点	缺点
部分嵌入式电容的优点（超轻触碰识别；允许齐平包边保护玻璃；可密封；等等）	仅对大规模的显示器有实际意义，因为本质上它是一款独特的显示设计
也许是最低成本的嵌入式触控技术（只有一组传感器；因所需材料更少而更容易和LCD集成），虽然图像处理的要求可能抵减这一优势	难以达到分离式或嵌入式电容相同的触控性能；处理传感器输出的数据需要使用CPU/GPU密集型图像处理软件（同视觉触控）
没有严格的尺寸限制（截至2013年最大尺寸为40in）	降低的光学性能（因光传感器而存在较低的LCD孔径比；因红外发射器而存在更低的OLED光输出）
稍微比分离式电容更薄（通常100~150μm），但是与嵌入式电容厚度相同	对环境红外光敏感；使用越多光检测器时敏感度越高；在十分光亮的红外环境下难以避免光检测器饱和
比分离式和嵌入式电容触敏需要更少的“参数调整”	当通过触碰物体折射的红外光和外部红外光相同时，在交叠点处的信噪比低（低触敏）
对外部RFI/EMI的敏感性更低	光检测器的低信号水平增加了触敏对内部干扰的敏感度（比如，临近的光检测器的杂散电流）
	触敏随着接触平面远离LCD而降低（即，空气隙，更大屏幕要求适配更厚的保护玻璃，等等）
	光学传感器密度（ppi）不足以使显示器兼具扫描器的功能
	由红外发射器导致功耗增加
	如果没有额外元件就无法支持屏外图标（同嵌入式电容）

2.9 其他触控技术

2.9.1 压力感测（编号18）

压力感测总是被视为触控技术的“圣杯”，因为最简单的检测触碰的方法应该只需在基板多个位置上测量触碰的压力并通过三角测距找到源触点——如果真是那么简单该多好！

基于压力感测的最早闻名的商业产品当属IBM的“TouchSelect”触控系统，该触控技术应用于1991年生产的2~19in的CRT显示器上。它使用了应变仪来安装触摸屏。然而这并未取得成功，仅在市场出现了不超过三年就销声匿迹了。接下来的压力感测产品诞生于2007年的美国，由QSI——一家制造人机互动产品和移动数据终端的犹他州公司出品。该技术冠名InfiniTouch™，巧妙地用搭建支架的方法装配应变仪，从而消除了触碰水平方向的作用力[79]。为了避免影响其现有销售，QSI在2008年使压力感测技术自立门户并命名为Vissumo[80]，成为了一个单独的子公司。该子公司融资不足，无法完成将一项崭新技术打入竞争激烈的市场的艰巨任务，因此他们在2009年就耗尽了资本并关门停产了（QSI在2010年

被 Beijier Electronics 收购，并从此使用该名）。

在压力感测技术的另一次商业化尝试中，芬兰的 MyOigo 于 2000 年开发了一个用于智能手机高级用户界面的压力感测触摸屏。MyOigo 在 2004 年由其管理层买断并更名为 F – Origin 重新开始。F – Orign 2005 年在芬兰倒闭，其资产由一个美国投资商收购，并在美国于 2006 年重组 F – Orign。随后的 2007 ~ 2008 年，F – Origin 继续开发压力感测技术（冠名 zTouch^{TM}）），但由于受到投射电容的迅速发展和普及的影响，其仍未能在电子消费市场中获得任何的影响力。2009 年，该公司被 TPK 收购注资（世界上最大的投射电容供应商），实现重组，并于 2010 年冠名 zTouch 在市场上开始出现。F – Origin 目前专注于商业性应用产品，在这方面，压力感测触控技术的耐用性和抗环境干扰性具有特别价值[81]。

压力感测触控通过使用压力传感器支撑显示屏（或保护玻璃层）运行，这些传感器通常是应变仪或压电式传感器。为了获取触碰表面的准确压力读数，显示器和/或保护玻璃的移动必须限制，这样才能确保它们在 Z 方向上移动。有几种方法可以完成这个任务。图 2.37 展示了目前 F – Origin 使用的悬挂弹簧臂法。

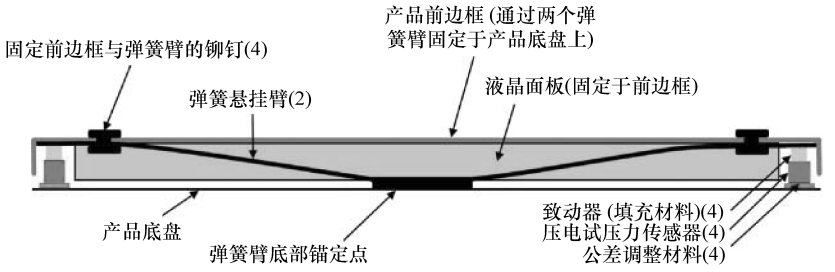


图 2.37　在 F – Origin 压力感测触控技术中，LCD/保护玻璃/前边框组件是由两个悬挂弹簧臂支撑的，弹簧臂的中部固定在设备外壳（产品底盘），其尾部与该组件连接。四个压电式压力传感器安装在组件的各角，位于组件和设备外壳之间。由于悬挂弹簧臂的限制，组件只能在 Z 方向上移动，任何该组件（如显示屏保护玻璃）之上的压力就能被检测并由四个压电式压力传感器定位。来源：改编自 F – Origin，图解由作者标注

压力感测触控技术的优缺点见表 2.19。应注意到该分析不包括新的尚未公布的压力感测方法，比如在显示屏下使用压力感测有机晶体管阵列的 NextInput 产品[82]。

本章一开始，我们就预测了多点触控将最终在商业应用中发挥重要作用。基于该预测，压力感测技术可能会在下一个五年内消失，或者最多是一个非常小众的市场。

表 2.19　压力感测触控技术的优缺点

优点	缺点
由于平面玻璃基板而产生的高光学效果	难以达到多点触控（两个触点要求八个传感器；这之后传感器数量急剧增加）
接触面独立（面板可以是任何硬材料）；可以是带有嵌入式移动物体的 3D 基板（这是独特的）	最小触摸作用力接近零，但不为零，类似投射电容

（续）

优点	缺点
可以被任何接近零触点作用力的物体激活（触控笔独立）；比电阻式要更好	很难但并非不可能实现齐平包边的屏幕设计
压力敏感；压力可取代悬浮（轻触屏幕获得选项显示；重触屏幕给出选择命令）；压力也可用来减少假性碰触	由于缺乏五点触控，无法满足 Windows 8 触控规范；这限制其向商业或向非 Windows 的消费者产品的发展
耐用特性；可以很容易设计来处理多变的外部环境	触感机制的机械本质降低了可靠性
由于传感器的简化实现的相对低成本（基板加上四个压电式感应器）	大量传感器增加了系统负荷（厚度或占用空间）
没有预触控（用户必须实际触碰基板）	
对 EMI/RFI 和环境光不敏感	
连续校准可以过滤诸如振动等环境条件	
已经升级到 42in；理论上还可以继续升级	

2.9.2 组合触控技术

本章介绍的信息清晰地回应了“没有一项触控技术是完美的”的概括。任何一个单项技术无法满足所有应用的要求。组合这些技术倒是一个可以更好的制造触摸屏的方案。这样的例子在平板电脑、电子书和电子销售终端机中常常可见。例子如下。

最新的 Microsoft 平板电脑常把投射电容式指控触摸屏与电磁（EM）数码笔结合。主导的 EM 数码笔销售商（日本的 Wacom）提供可以同时驱动触摸屏和数码笔的控制器，实现了手笔输入模式的自动切换[83]。

在 2011 年 5 月，Hanvon 公布了一项可以实现相同手笔操作目标的新技术组合方法。Hanvon 将 EM 数码笔和一组压感压电式电容组装在相同的面板中，EM 传感器在显示屏下方。同样用在笔尖上的压电式电容能够通过电子书显示屏（EPD）本身（而不是表面）感受到手指的作用力。

一家自动贩卖机的主要供应商偏爱在其产品中使用传统红外光。但为了减少“预触碰”的问题（即手指干扰了红外光束，在没有实际触碰显示器表面的情况下发出了触控指令），供应商在触摸屏组装中增加了一个压电换能器，确保触点坐标只能在用户真正触碰屏幕的时候才生成。在该应用中，换能器检测的是触碰的“发生”，而传统红外触摸屏探测的是触点的“位置”。

触控技术的组合很可能在未来五年内持续存在，尽管主要技术的组合常常受到成本的限制。捆绑主要技术与边缘技术的可能性很大，如上述提到的组合传统红外和压力传感器的例子。结合现有的触控技术与新兴的人机界面（HMI）技术同样很可能出现；比如，投射电容与低成本、迷你 3D 相机的组合可以检测触摸屏之上及近场空间内的手势，从而超出一般的悬浮检测范围。最终应该可以实现的是，在投射电容式触摸屏上操纵一个物体，然后将其从 2D 屏幕“拽出”并拖入显示器和用户之间的 3D 空间内，从而在投射电容式触摸屏和 3D 相机之间无缝转换物体。

2.10　结语

本章较为细致地介绍了当前 18 项触控技术（外加几种已经停用的技术）。总结这些笨拙信息的最佳办法莫过于预测这 18 项触控技术将如何在未来的 5～10 年 1 内演变发展，见表 2.20。

表 2.20　该表记叙了本章描述的 18 项触控技术在未来 5～10 年内的可能演变过程

编号	触控技术	预测
1	投射电容	继续保持消费市场第一的位置；在商业应用中显著增长
2	表面电容	5～7 年内从市场消失
3	模拟电阻	显著减少，但不可能完全从市场消失
4	数字多触点电阻	专营商业和军事应用
5	模拟多触点电阻	专营商业和军事应用
6	表面声波	在商业应用中适度增长
7	声学脉冲识别	专营非显示器应用（触敏表面和设备）
8	色散信号	不到五年就会从市场消失
9	传统红外线	大屏应用适度减少；商业应用显著减少；带反射显示的移动设备可能增加
10	多点触控红外线	在多用户游戏和/或合作的应用普及且技术成本降低之前保持有限增长
11	摄像光学	仅大于 40in 的显著增长
12	玻璃光学	大屏专营市场的应用，可能在消费市场中的多合一桌面电脑中应用
13	视觉光学	大屏专营市场的应用
14	外嵌式电容	大规模消费设备中显著增长；由于对 LCD 屏的生产干扰最小，将成为最普及的嵌入式触控形式
15	嵌入式混合电容	仅在大规模消费设备中显著增长；普及程度排名第二
16	内嵌式电容	仅在大规模消费设备中显著增长；由于对 LCD 屏生产的调整较大，普及程度排名第三
17	内嵌式光感	不到五年内从市场消失，除非能解决当前存在的问题
18	压力感测	不到五年内从市场消失或存在于非常小的专营商业市场

2.11　附录

所有本章提到的触控技术供应商（不再经营的除外）均按照首字母排序，见表 2.21，其相应开发的技术和网页地址也一并列出。

表 2.21 该表列出了本章提到的所有触控技术供应商（不再经营的除外）及其相应开发的技术和网页地址。

表　2.21

公司	技术编号	网址
3M Touch Systems	1，2，8	www. 3mtouch. com
Apex Material Technology（AMT）	1，3，5	www. amtouch. com. tw
苹果	1，15	www. apple. com
Atmel	1	www. atmel. com

（续）

公司	技术编号	网址
Baanto	11	www. baanto. com
Citron	10	www. citron. de
Cypress Semiconductor	1	www. cypress. com
Elo Touch Solutions	1，2，3，6，7，9	www. elotouch. com
FlatFrog	12	www. flatfrog. com
F – Origin	18	www. f – origin. com
General Touch	1，3，6，9，10，11	www. generaltouch. com
Gunze USA	1，3，4	www. gunzeusa. com
IDS Pulse	10	www. idspulse. com
Integrated Digital Technologies（IDTI）	17	www. idti. com. tw
Japan Display（JDI）	16	www. j – display. com
JTouch	1，3，5	www. jtouch. com. tw
LG Displays	1，15	www. lgdisplay. com
Lumio	1，6，10，11	www. lumio. com
微软	13	www. microsoft. com
MultiTouch	13	www. multitaction. com
Nissha	1，3	www. nissha. com
Peratech	4	www. peratech. com
Planar	1，3，6，10，11	www. planar. com
PQ Labs	10	www. pqlabs. com
Quanta	11	www. quantatw. com
三星	14，15，17	www. samsung. com
夏普	15，17	www. sharp – world. com
Shenzhen TimeLink Technology	10	www. timelink. cn
SMART Technologies	11	www. smarttech. com
Stantum	4	www. stantum. com
Synaptics	1，16	www. synaptics. com
Texas Instruments	1，3，5	www. ti. com
TPK	1	www. tpk. com
Visual Planet	1	www. visualplanet. biz
Wacom	1，2	www. wacom – components. com
ZaagTech	10	www. zaagtech. com
Zytronic	1	www. zytronic. co. uk

参考文献

1. Johnson, E.A. (1965). Touch Display – A Novel Input/Output Device for Computers. *Electronics Letters* **1**(8), 219–220.
   Further Reading:
   Johnson, E.A. (1967). Touch Displays: A Programmed Man-Machine Interface. *Ergonomics* **10**(2), 271–277.
   Orr, N.W., Hopkins, V.D. (1968). The Role of Touch Display in Air Traffic Control. *The Controller* **7**, 7–9.
2. Shneiderman, B. (1991). Touch screens now offer compelling uses. *IEEE Software* **2**,93–94, 107.
3. Walker, G. (2007). Touch and the Apple iPhone, *Veritas et Visus Touch Panel* **12**, 50–54, (http://www.walkermobile.com/Touch_And_The_Apple_iPhone.pdf, retrieved 10/15/13).

4. DisplaySearch, 2008–2013. *Touch-Panel Market-Analysis Annual Reports*.
5. Buxton, B. (2007–2013). *Multi-Touch Systems that I Have Known and Loved*. Microsoft (www.billbuxton.com /multitouchOverview.html, retrieved 9/25/13).
6. Wigdor, D. (2011). The Breadth-Depth Dichotomy: Opportunities and Crises in Expanding Sensing Capabilities. *Information Display* **3**, 18–23, (http://informationdisplay.org/IDArchive/2011/March /EnablingTechnologyTheBreadthDepthDichotomy.aspx, retrieved 10/15/13).
   Further Reading:
   Wigdor, D., Wixon, D. (2011). *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann (Elsevier), Burlington, MA.
7. Stumpe, B. (1977). A New Principle for X-Y Touch System, CERN (http://cds.cern.ch/record/1266588/files /StumpeMar77.pdf, retrieved 10/15/13).
   Further Reading:
   Stumpe, B., Sutton, C. (2010). The First Capacitive Touch Screens at CERN. *CERN Courier* (http:// cerncourier.com/cws/article/cern/42092, retrieved 10/15/13).
   CERN Bulletin (2010). *Another of CERN's Many Inventions*. BUL-NA-2010-063, (http://cds.cern.ch/record /1248908, retrieved 10/15/13).
   Beck, F., Stumpe, B. (1973). *Two Devices for Operator Interaction in the Central Control of the New CERN*. CERN (http://cds.cern.ch/record/186242/files/CERN-73-06.pdf?version=1, retrieved 10/15/13).
   Stumpe, B. (1978). *Experiments to Find a Manufacturing Process for an X-Y Touch Screen*. CERN (http://cds.cern.ch/record/1266589/files/StumpeFeb78.pdf, retrieved 10/15/13).
8. Logan, J. (1991). *The History of MicroTouch: 1982–1992, a Decade of Touch Input*. Pamphlet published by the founder of MicroTouch Systems.
9. Binstead, R. (2009). *A Brief History of Projected Capacitance Development by Binstead Designs* (http://binsteaddesigns.com/history1.html, retrieved 10/15/13).
10. Barrett, G., Omote, R. (2010). Projected-Capacitive Touch Technology. *Information Display* **3**, 16–21, (http://informationdisplay.org/IDArchive/2010/March/FrontlineTechnologyProjectedCapacitiveTouchT.aspx, retrieved 10/15/13).
    Further Reading:
    3M Touch Systems (2011). *Touch Technology Brief: Projected Capacitive Technology* (http://solutions.3m.com /3MContentRetrievalAPI/BlobServlet?lmd=1332776667000&locale=en_US&assetType=MMM_Image& assetId=1319224170371&blobAttribute=ImageFile, retrieved 10/15/13).
11. Wang, T., Blankenship, T. (2011). Projected-Capacitive Touch Systems from the Controller Point of View. *Information Display* **3**, 8–12, (http://informationdisplay.org/IDArchive/2011/March/FrontlineTechnologyProjected CapacitiveTouchS.aspx, retrieved 10/15/13).
    Further Reading:
    Lawson, R. (2012). *Challenges and Opportunities in Touch-Controller Semiconductors, HIS*.
    SID-IHS Future of Touch and Interactivity Conference, Boston, MA.
12. DisplaySearch (2013). *Touch-Panel Market-Analysis Annual Report*.
13. Poor, A. (2012). *How It Works: The Technology of Touch Screens*. Computerworld (http://www.computerworld .com/s/article/9231961/How_it_works_The_technology_of_touch_screens?taxonomyId=12&pageNumber=1, retrieved 10/15/13).
14. Bauman, C. (2007). How to Select a Surface-Capacitive Touch-Screen Controller. *Information Display* **12**:, 32–36 (http://informationdisplay.org/IDArchive/2007/December/HowtoSelectaSurfaceCapacitiveTouchScreenCo.aspx, retrieved 10/15/13).
15. Harrah's Entertainment (2006). *Profile of the American Casino Gambler* (http://www.org.id.tue.nl/ifip-tc14 /documents/HARRAH′S-SurveyCasinoGambler-2006.pdf, retrieved 10/15/13).
16. Wacom (2013). *RRFC™ Reversing Ramped-Field Capacitive Touch-Technology* (http://www.wacom-components.com/english/technology/touch.html, retrieved 10/15/13).
17. Wikipedia (2013). *Touchscreen* (http://en.wikipedia.org/wiki/Touchscreen, retrieved 10/15/13).
18. Emerson, L.G. (2010). *Samuel Hurst – the 'Tom Edison' of ORNL*. OakRidger (http://www.oakridger.com /article/20101214/NEWS/312149981?tag=1, retrieved 10/15/13).
19. Elo Touch Solutions (2013). *History of Elo* (http://www.elotouch.com/AboutElo/History/default.asp, retrieved 10/15/13).
20. Westinghouse Electric (1970). *Interface Device and Display System*. US Patent 3,522,664 (http://www .freepatentsonline.com/3522664.html, retrieved 10/15/13).

21. Sierracin/Intrex (1979). TransTech product brochure.
22. Downs, R. (2005). Using Resistive Touch Screens for Human/Machine Interface. *Texas Instruments Analog Applications Journal* (SLYT209A).
23. Barrett, G. (2012). *Decoding Touch Technology: An Insider's Guide to Choosing the Right Touch for your Display*. Touch International, white paper (http://touchinternational.com/literature/choosing-touch-technology-whitepaper.html, retrieved 10/15/13).
    Further Reading:
    Barrett, G. (2006). *Frit and the Better Touch Screen*. Touch International, white paper (http://touchinternational.com/literature/whitepapers/FritandtheBetterTouchScreen.pdf, retrieved 10/15/13).
24. DMC Co (2011). *Technologies of Touch Screens – Analog 8-Wire Resistive* (http://www.dmccoltd.com/english/museum/touchscreens/technologies/8-wire.asp, retrieved 10/15/13).
25. Semtech (2011). *SX8677/SX8678 Haptics-Enabled Multitouch 4/5-Wire Resistive Touchscreen Controller with Proximity Sensing*. Datasheet (http://www.semtech.com/images/datasheet/sx8677_8.pdf, retrieved 10/15/13).
26. Elo Touch Solutions (2013). *AccuTouch ZeroBezel Technology Specifications* (http://www.elotouch.com/Technologies/AccuTouch/accutouch_zero-bezel_specifications.pdf, retrieved 10/15/13).
27. Fenn, J. (2013). *Turned Around a Potential Loss to Revolutionize a Business*. Professional Resume (http://www.corporatewarriors.com/john7820/fenn.doc, retrieved 10/15/13).
28. DMC Co (2011). *Technologies of Touch Screens – Digital Matrix Resistive* (http://www.dmccoltd.com/english/museum/touchscreens/technologies/Matrix.asp, retrieved 10/15/13).
29. Largillier, G. (2007). Developing the First Commercial Product that Uses Multi-Touch Technology. *Information Display* **12**, 14–18 (http://informationdisplay.org/IDArchive/2007/December/DevelopingtheFirstCommercialProductthatUses.aspx, retrieved 10/15/13).
30. Stantum (2012). *Stantum's Newest Digital Resistive Touch-Panel* (http://www.leavcom.com/stantum_061012.php, retrieved 10/15/13).
31. Texas Instruments (2011). *Analog Matrix Touchscreen Controller*. Datasheet (http://www.mouser.com/ds/2/405/sbas536b-94322.pdf, retrieved 10/15/13).
32. Apex Material Technology (AMT) (2013). *Multi-Finger (MF) Touch* (http://www.amtouch.com.tw/products/advanced-resistive-touch-screen/multi-finger-mf-touch/, retrieved 10/15/13).
33. Adler, R., Desmares, P.J. (1985). *An Economical Touch Panel Using SAW Absorption*. IEEE Ultrasonics Symposium 1985, 499–502. (10.1109/ULTSYM.1985.198560).
34. Kent, J. *et al.* (2007). *Robert Adler's Touchscreen Inventions*. IEEE 2007 Ultrasonics Symposium, 9–20 (10.1109/ULTSYM.2007.18).
35. Elo Touch Solutions (1997–2001). *Acoustic Touch Position Sensor Using a Low Acoustic Loss Transparent Substrate*. US Patent 6,236,391 (http://www.freepatentsonline.com/6236391.html, retrieved 10/15/13).
36. Elo Touch Solutions (2010–2013). *Acoustic Condition Sensor Employing a Plurality of Mutually Non-Orthogonal Waves*. US Patent 8,421,776 (http://www.freepatentsonline.com/8421776.html, retrieved 10/15/13).
37. Microsoft (2013). *Windows Certification Program: Hardware Certification Taxonomy & Requirements for Windows 8.1, Device Digitizer Requirements* (http://msdn.microsoft.com/en-us/library/windows/hardware/jj134351.aspx, retrieved 10/15/13).
38. Elo Touch Solutions (2011–2013). *Bezel-less Acoustic Touch Apparatus*. US Patent 8,576,202 (http://www.freepatentsonline.com/8576202.html, retrieved 10/15/13).
39. North, K., D'Souza, H. (2006). Acoustic Pulse Recognition Enter Touch-Screen Market. *Information Display* **12**, 22–25 (http://informationdisplay.org/IDArchive/2006/December/AcousticPulseRecognitionEntersTouchScreenMar.aspx).
40. Kent, J. (2010). New Touch Technology from Time Reversal Acoustics: A History. *IEEE International Ultrasonics Symposium Proceedings* 1173–1178.
41. Butcher, M. (2010). The $62 Million Sale of a Touch Tech Startup Adds to the Tablet Revolution. *TechCrunch* (http://techcrunch.com/2010/01/27/the-62-million-sale-of-a-touch-tech-startup-adds-to-the-tablet-revolution/, retrieved 10/15/13).
42. 3M Touch Systems (2008). *Dispersive Signal Touch Technology: Technology Profile*. White paper (http://multimedia.3m.com/mws/mediawebserver?mwsId=66666UF6EVsSyXTtmxTXoxfaEVtQEVs6EVs6EVs6E666666--&fn=DST%20Tech%20Profile.pdf, retrieved 10/15/13).
43. Wikipedia (2013). *PLATO Computer System* (http://en.wikipedia.org/wiki/PLATO_IV, retrieved 10/15/13).

44. Wikipedia (2013). *HP-150* (http://en.wikipedia.org/wiki/HP_150, retrieved 10/15/13).
45. Elo Touch Solutions (2009). *Elo TouchSystems IntelliTouch Plus Multi-Touch & Windows-7*. Communication Brief (http://www.elotouch.com/pdfs/faq_ip.pdf, retrieved 10/15/13).
46. Charters, R. (2009). High-Volume Manufacturing of Photonic Components on Flexible Substrates. *Information Display* **12**, 12–16 (http://informationdisplay.org/IDArchive/2009/December/FrontlineTechnologyHighVolume Manufacturingof.aspx, retrieved 10/15/13).
47. Thompson, M. (2009). *RPO Digital Waveguide Touch*. DisplaySearch 2009 Emerging Display Technologies Conference, San Jose, CA.
    Further Reading:
    Maxwell, I. (2007). An Overview of Optical-Touch Technologies. *Information Display* **10**, 26–30 (http://informationdisplay.org/IDArchive/2007/December/AnOverviewofOpticalTouchTechnologies.aspx, retrieved 10/15/13).
48. Poa Sana (1997–1999). *User Input Device for a Computer System*. US Patent 5,914,709 (http://www .freepatentsonline.com/5914709.html, retrieved 10/15/13).
49. PQ Labs (2010–2012). *System and Method for Providing Multi-Dimensional Touch Input Vector*. US Patent Application 2012-0098753 (http://www.freepatentsonline.com/y2012/0098753.html, retrieved 10/15/13).
50. Zytronic (2013). *Zytronic Reveals a New Dimension in Touchscreens for Gaming Machines*. Press release (http://www.zytronic.co.uk/assets/Uploads/ZY370-G2E-2013-Zytronic-reveals-a-new-dimension-in-touchscreens-for-gaming-machines.pdf, retrieved 10/15/13).
51. PQ Labs website, (http://www.pqlabs.com, retrieved 10/15/13).
52. Walker, G. (2011). Camera-Based Optical Touch Technology. *Information Display* **3**, 30–34 (http://informationdisplay.org/IDArchive/2011/March/FrontlineTechnologyCameraBasedOpticalTouchT.aspx), retrieved 10/15/13.
53. Baanto (2011). *ShadowSense™ Touch Detection*. White paper (http://baanto.com/uploads/Image/pdfs /whitepapers/shadowsense_touch_detection.pdf, retrieved 10/15/13).
54. Baanto (2011). *Rain and Fluid Discrimination for Touchscreens*. White paper (http://baanto.com /uploads/Image/pdfs/whitepapers/shadowsense_rain_rejection.pdf, retrieved 10/15/13).
55. Wassvik, O. (2013). *PSD: In-Glass Optical Touch for Larger Form-Factors*. DisplaySearch Emerging Display Technologies Conference, San Jose, CA.
56. Walker, G., Finn, M. (2010). Beneath the Surface. *Information Display* **3**, 31–34, (http://informationdisplay .org/IDArchive/2010/March/EnablingTechnologyBeneaththeSurface.aspx).
57. Castle, A. (2009). Build Your Own Multitouch Surface Computer. *Maximum PC* (http://www.maximumpc.com /article/features/maximum_pc_builds_a_multitouch_surface_computer?page=0,0, retrieved 10/15/13).
58. NUI Group Authors (2009). *Multitouch Technologies*. e-book, NUI Group (http://www.google.com/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&ved=0CCwQFjAA&url=http%3A%2F%2Fnuicode .com%2Fattachments%2Fdownload%2F115%2FMulti-Touch_Technologies_v1.01.pdf&ei= C7yBUqn6N4G3igKz9YDIDQ&usg=AFQjCNGWgAwHLy64d0YfObchamgKkeby8g&bvm= bv.56343320,d.cGE, retrieved 10/15/13).
59. Anttila, H. (2012). *Multi-User Interactive Technology Advancements for Any Size Display*. IHS Touch-Gesture-Motion Conference, Austin, TX.
    Further Reading:
    MultiTouch (2011). *MultiTaction: Technology Platform for MultiTouch LCDs of Any Size*. White paper http://multitouch.s3.amazonaws.com/resources/brochures_for_print/whitepaper_MultiTaction_v1-1_USletter _print.pdf, retrieved 10/15/13).
60. Samsung Electronics (2011). *Samsung SUR40 for Microsoft Surface Venue Readiness Guide* (http://www.samsung.com/us/pdf/sur40/SUR40_Venue_Readiness_Guide.pdf, retrieved 10/15/13).
61. Microsoft (2012). *Tagged Object Integration for Surface 2.0*. White paper. (http://download.microsoft.com /download/D/7/B/D7BE282A-FCB2-4A2C-AC48-6BC8441AB281/Tagged%20Objects%20for%20Surface% 202.0%20Whitepaper.docx, retrieved 10/15/13).
62. Walker, G. (2013). Embedded Touch: The Touch-Panel Makers vs. The Display-Makers, FPD International Conference, Yokohama, Japan, (http://www.walkermobile.com/FPD_International_2013_Touch_Futures.pdf, retrieved 10/15/13).
63. Walker, G., Finn, M. (2010). LCD In-Cell Touch. *Information Display* **3**, 8–14. (http://informationdisplay .org/IDArchive/2010/March/FrontlineTechnologyLCDInCellTouch.aspx, retrieved 10/15/13).
64. Lee, J. *et al*. (2007). Hybrid Touch Screen Panel Integrated in TFT-LCD. *SID 2007 Symposium Digest* 24.3.

65. Samsung Electronics (2004–2007). *Liquid Crystal Display Device Having Touch Screen Function and Method of Fabricating the Same*. US Patent 7,280,167 (http://www.freepatentsonline.com/7280167.html, retrieved 10/15/13).
66. Den Boer, W. *et al*. (2003). Active Matrix LCD with Integrated Optical Touch Screen. *SID 2003 Symposium Digest* 56.3.
67. Samsung Electronics (2007–2011). *Touch Screen Display Apparatus and Method of Driving the Same*. US Patent 8,072,430 (http://www.freepatentsonline.com/8072430.html, retrieved 10-15-13).
68. Ozbas, M. *et al*. (2012). An In-Cell Capable Capacitive Touchscreen Controller with High SNR and Integrated Display Driver IC for WVGA LTPS Displays. *SID Symposium 2012 Digest* 485–488.
69. Mackey, B. (2013). *Touch + Display, Any Way You Want It, Synaptics*. SID Display Week Conference (Session M8), Vancouver, Canada.
70. Synaptics. (2010). *Capacitive Sensing Using a Segmented Common Voltage Electrode of a Display*. US Patent Application 2010-0238134 (http://www.freepatentsonline.com/y2010/0238134.html, retrieved 10/15/13).
71. Noguchi, K. (2012). *Trend of In-Cell Touch Panel Technologies*. Sony, FPD International Conference, Yokohama, Japan.
72. Apple (2010). *Integrated Touch Screen*. US Patent 7,859,521 (http://www.freepatentsonline.com/7859521.html, retrieved 11/15/13).
    Further Reading:
    Apple (2011–2013). *Segmented Vcom*. US Patent 8,451,244 (http://www.freepatentsonline.com/8451244.html, retrieved 10/15/13).
    Apple (2009–2011). *Integrated Touch Screen*. US Patent 7,995,041 (http://www.freepatentsonline.com/7995041.html, retrieved 10/15/13).
    Apple (2010–2010). *Integrated Touch Screen*. US Patent 7,859,521 (http://www.freepatentsonline.com/7859521.html, retrieved 10-15-13).
73. Wu, C.W. (2013). *On/In Cell Touch Sensor Embedded in TFT-LCD*. BOE Technology Group, FPD China Conference, Shanghai, China.
74. Toshiba Matsushita Display (TMD) (2003). *Toshiba America Electronic Components Demonstrates First System on Glass (SOG) Input Display with Built-In Image Capture*. Press release (http://www.toshiba.com/taec/news/press_releases/2003/to-314.jsp, retrieved 10/15/13).
75. Toshiba Matsushita Display (TMD) (2005). *Toshiba Matsushita Display Announces World's First LTPS TFT LCD Prototype with Finger Shadow Sensing Input Capability*. Press release (http://www.toshiba-components.com/prpdf/5615e.pdf, retrieved 10/15/13).
76. Chen, Z.H. (2011). IDTI: In-Cell Optical Touch Panel. *Optoelectronic Notes Blog* (http://ntuzhchen.blogspot.com/2011/03/idti-in-cell-touch-panel.html, retrieved 10/15/13).
77. Samsung Electronics, 2014, Product Webpage for Samsung SUR40 with Microsoft PixelSense (formerly Microsoft Surface 2.0), (http://www.samsung.com/ae/business/business-products/large-format-display/specialized-display/LH40SFWTGC/XY, retrieved 03/18/14).
78. Samsung Mobile Display (2011–2012). *Organic Light Emitting Display Having Touch Screen Function*. US Patent Application 2012-0105341 (http://www.freepatentsonline.com/y2012/0105341.html, retrieved 10/15/13).
79. Soss, D. (2007). Advances in Force-Based Touch Panels. *Information Display* **12**, 20–24 (http://informationdisplay.org/IDArchive/2007/December/AdvancesinForceBasedTouchPanels.aspx, retrieved 10/15/13).
80. Fihn, M. (2009). Interview with Garrick Infanger from Vissumo. *Veritas et Visus Touch Panel* **3**(7/8), 80–83.
81. F-Origin (2013). *Force-Based Touch-Screen Technology* (http://www.f-origin.com/zTouch0153Technology.aspx, retrieved 10/15/13).
82. NextInput (2013). *Force-Sensitive Touch Technology* (http://nextinput.com/t/ForceTouch, retrieved 10/15/13).
83. Wacom (2013). *Touch Panels Product* (http://wacom.jp/en/products/components/displays/touch/index.html, retrieved 10/15/13).

第3章

用户界面中的声控式交互技术

Andrew Breen, Hung H. Bui, Richard Crouch, Kevin Farrell,
Friedrich Faubel, Roberto Gemello, William F. Ganong III, Tim Haulick,
Ronald M. Kaplan, Charles L. Ortiz, Peter F. Patel – Schneider, Holger Quast,
Adwait Ratnaparkhi, Vlad Sejnoha, Jiaying Shen, Peter Stubley, Paul van Mulbregt
Nuance 通信公司

3.1 引言

基于自然语言理解的语音识别和合成是现代移动通信设备用户界面（UI）不可或缺的部分。近年来，这些技术从配合文本输入、支持有限命令和控制的“附加程序”已经发展成各种主流移动消费设备的核心功能，如语音驱动智能手机系统。有评论甚至把 UI 语音识别和自然语言的理解定义为用户界面的“第三次革命”，第一次和第二次分别是鼠标输入的图形用户界面和触摸输入的触控感知界面。

这些新技术名声大噪的主要因素有两个：一是它们快速改进的性能；二是它们克服现存“收缩桌面”式的移动 UI 固有缺陷的能力。后者主要通过从有声语言输入中精准地推断用户意图。

伴随各种移动设备使用量暴增的是用户对“内容”、功能、服务和应用方面同样急速增长的需求。海量的信息变得愈发难以用现有的可视移动桌面识别、寻找和管理；信息很容易淹没在层级文件夹、几十种甚至几百种应用图标、应用屏幕和各种菜单中。

通常，执行单个触摸屏装置指令需要多个步骤。例如，一个简单的银行转账事项需要用专门的移动应用程序来回切换十几个应用屏幕。

不同设备的特定用户界面中存在很多的变化性，使得可用性问题变得更加严重。现在移动设备有许多种“形态因素”：有大屏和虚拟键盘的平板电脑，有为眼手忙碌而无暇操作提供便捷的车载装置界面，有无键盘无定点设置的电视机，也有各种“可穿戴的”装置（例如智能眼镜和手表）。通过这些完全不同的界面，用户正越来越多地获取相似的服务——搜

索信息、查收邮件、浏览社交媒体、定位导航以及欣赏音乐和视频等。

在这样的背景下，语音识别（VR）和自然语言理解（NLU）代表了一个强大的自然控制机制，它可以穿过多重视觉层次、中间应用或网页。自然语言的表达紧凑地对大量信息进行了编码。当你说“发条短信给罗恩，说我要迟到 10 分钟”就能暗示哪个应用程序应该先启动、要把信息发给谁和发送什么信息，而不用明确地提供所有信息和每个步骤。同样的，你可以给电视下令：“播放昨晚保存的女高音歌曲”，要比使用常规界面、横贯多层菜单结构更简单。这些功能的实现能够创造一个新的 UI：一个可以通过对话与用户互动并提供强大功能的虚拟助手（VA）。

在以上例子中，用户开始操作时无需先点击电子邮件的应用程序图标，只要用语音和自然语言就能找到并操纵资源——无论它们是显示在设备屏幕上还是存储在设备或云端（Cloud）中。这种融入其他服务的方式有效地拓宽了传统界面应用。

通过了解用户的意图、喜好和过往的交流记录，包含了语音和自然语言的界面在解决问题时可以绕过中间搜索引擎结果页，直接定位到认为对用户有用的目的页面上去。例如，某位用户的产品查询将直接在页面中显示他/她平时喜好的购物网站。

换言之，这样一个系统可以直接从结构化数据源或非结构化数据源中提取想要的信息，通过自然语言生成（NLG）来构建答案，然后通过语音合成进行反馈。

最后，那些很难用点选式界面明确说明的指令在语音界面上是容易表达的，例如，写一个以其他事件为条件的通知：“快到咖啡店的时候通知我。”

在符合用户需求的条件下，还可以用其他方式减少一些步骤。用户甚至可以自然地对设备说出自己的需求而无需开启设备。在一种称为“无缝唤醒”的模式下，装置运用节能算法的数字信号处理器（DSP），能够持续地接收到重要事件的发生。当检测到有意义的输入时，装置会激活再处理模块以确定是来自主人的有效命令（用生物计量法确认身份），最后执行命令。

运用自然语言的前提条件是语音识别能在大量的用户和嘈杂的环境中准确的工作。语音识别在过去几年里发展显著，这主要归功于以下几方面：一个更加强大的计算基础（包括专门用于语音识别的芯片结构）；高速快捷的连接能力——接入云计算甚至是最小的移动平台；新算法和建模技术的发展（包括最近兴起的神经网络模型）；利用海量数据库训练强大的统计模型。

语音识别同样也利用了越来越复杂的信号采集技术，例如利用可控的多话筒波束形成和杂音消除运算来提高语音辨别在嘈杂环境里的准确率。在以车内和客厅内为代表的高噪声、多语音源和常有娱乐背景声的环境下，这种处理更有价值。

近期从自然表达中抽取意义的技术发展很快，主要得益于以下三个互补的方法：

- 能从数据中发现规律的机器学习。
- 明确的语言“结构”模式。
- 明确知识表现（本体）的形式，能把已知关系和实体预先编码。

就像在语音识别中一样，这些算法是自适性的，并且都从每次互动中适应、学习。

简洁概括的表达本身是很含糊的，但是人类却可以通过背景环境获取许多信息。同样

的，以算法的方式抽取正确信息要求应用一个通用的模型以及一个能够体现交互背景和历史的表达式，还包括由其他传感器和元数据提供的其他信息形式。在信息不足导致无法消除歧义时，语音和自然语言界面可能会与用户进行对话交流，获取或澄清信息。

对话或会话管理最早是从“系统主导”形式发展起来的，“系统主导”限制用户只能回答某个应用程序（通过视频或者合成语音）设置好的问题。但现在已经发展成更具灵活性的“混合主导”形式，让用户可以积极主动地提供相关信息。最先进的形式推理方式——传统人工智能（AI）的范畴——可以消除每次互动需要的预定义，并动态地推断出目标和计划。

早期的人工智能处理十分生硬，而现在的系统依靠的是既灵活又稳定的方法应对模糊表达。当无法提供准确的回应时，它也会给出最接近的解决方法。这种高级系统的目标就是能够成功地掌握所谓的“元任务”，例如，仅仅只要输入“最后一个会议后在‘吉普赛人私房菜’预订一个餐位，通知汤姆和布莱恩在那里等我”，而不是让用户顺序执行基础的“微”任务，例如确定日期和订桌。

因此，我们认为“语音界面”的宏观内涵实际上就是它是智能系统的重要组成部分，该系统包括：

- 通过多种方式和用户互动。
- 理解语言。
- 能对话和推理。
- 利用语境和用户喜好。
- 拥有专业知识。
- 解决高级任务。
- 在现实环境里具有稳定性。

如图 3.1 所示，该系统的元素通常分布在客户端和云服务上。

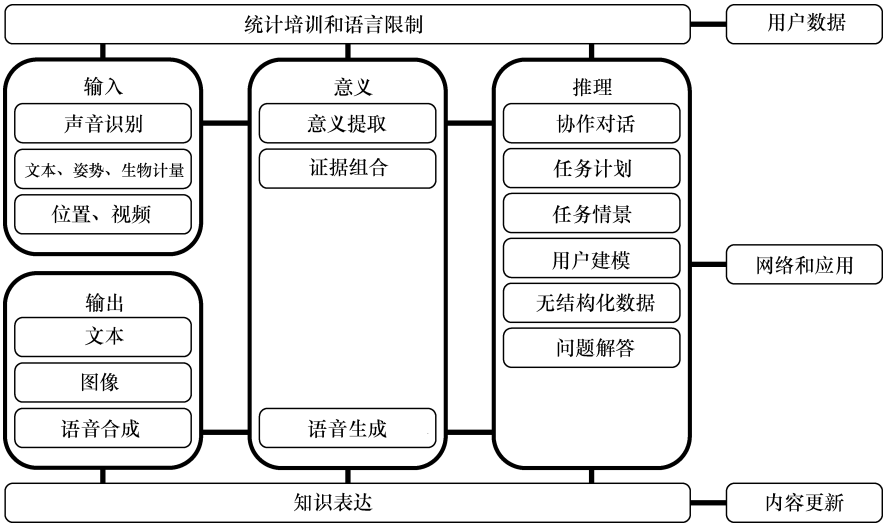


图 3.1　智能语音界面架构

这样做的原因包括优化计算、增加服务的可获取性和处理延迟，以及为用户提供在多对象、多元特性和功能的环境中始终如一的体验。

分布式的结构体系可以进一步使用户数据从多个设备中聚合，这样就可以不断改进服务器、具体设备识别和 NLU 模型。而且，存储在中央存储器里的交互历史能使用户无缝衔接其开始交互的设备与其完成交互的设备。

以下各节将详细描述这些概念和基本技术。

3.2　语音识别

3.2.1　语言的本质

语言属人类独有，能让人不费力地交流复杂的思想和感觉。因此“话音通道”才会被高度优化以促进人类完成交流任务。组成有声话语的小微语言元素叫作音素，它是语言中最小的单位，一旦改变，单词或者表达就会跟着变化。音素的物理表达就是“通话”，但语音信号不只是一系列拼接的声音，像摩尔斯电码。我们的发声器官（舌、下颚、唇）以难以置信的速度和精心的编排在变换着共振结构。我们的声带可以每秒打开和闭合 100 ~ 300 次，生成叫作基频（F0）的信号，它激发声道共振，从而发出一个高频宽的声音（例如 0 ~ 10kHz）。

有时，共振是混乱的噪声在声道收缩时产生的，例如 S 的发音。一个音素的声学表达不仅是不固定的，而且在现实中会受到前一个和下一个预期的音素影响——这种现象称之为协同发音。当说话者根据当前情况和听者的需求调整自己的话语时，其他的变化就会产生。由此导致的语音信号反映了这些在复杂且快速变化的信号中运动的发音器官和声源。图 3.2 展示了一个简短话语的语音谱图。

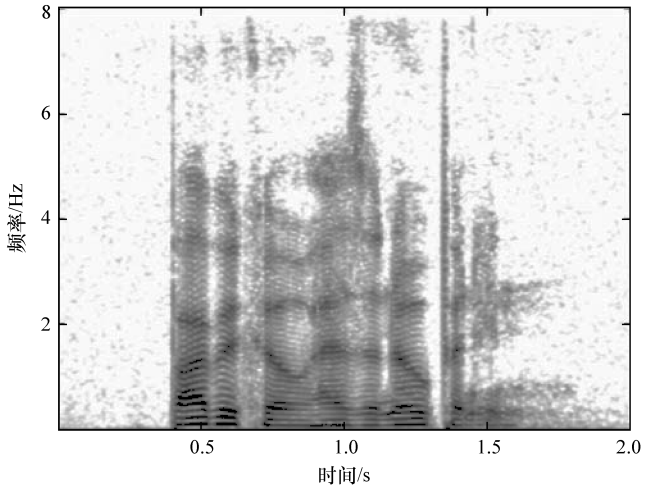


图 3.2　这是短语“Barbacco has an opening”的语音谱图，横坐标表示时间，纵坐标表示频率。黑色部分表示在一个频率范围内的总能量

语音识别的准确性和性能的进步是科学和工程学研究人员共同努力的结果，因此最先进的识别器包括了许多精心优化设计的元件。1990～2010 年间，大多数最先进的系统是相似的，并在逐步地加强和改进。接下来我们要介绍一种“标准”语音识别系统的基本组成部分以及一些最近的发展。

能利用标准语音识别器解决的问题都符合贝叶斯规则（Bayes’ rule）：

$$W^* = \arg\max_{\tilde{W}}(P(\tilde{W}|\bar{O})) \tag{3.1}$$

语音识别的目标是找到词组序列的最可能概率 W^*，假设声学观测集 $\bar{O}$，运用贝叶斯规则，我们可以得到：

$$P(\tilde{W}|\bar{O}) = \frac{P(\bar{O}|\tilde{W})P(\tilde{W})}{P(\bar{O})} \tag{3.2}$$

注意到 $P(\bar{O})$ 和词组序列 $\tilde{W}$ 无关，因此我们想要找到：

$$W^* = \arg\max_{\tilde{W}}(P(\bar{O}|\tilde{W})P(\tilde{W})) \tag{3.3}$$

我们使用声学模型（AM）评估 $P(\bar{O}|\tilde{W})$，并用语言模型（LM）评估 $P(\tilde{W})$。

因此，假设给定语言结构，大多数的语音识别器的目标就是通过声学观测得出的最高组合概率来找到词组序列。

如图 3.3 所示，一个标准语音识别系统图可以很好地反映到这个公式中。

声学概率的评估是由声音前端和一个声学模型处理的，而词组序列的概率评估则是由一个语言模型处理的。找到得分最高的词组序列的代码称为搜索组件。虽然这些模块在逻辑上是分开的，但是它们在语音识别中的应用是高度相互依赖的。

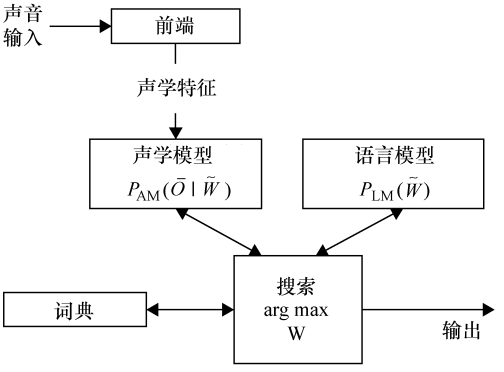


图 3.3　标准语音系统的组成元素

3.2.2　声学模型和前端模式

前端模式：输入的语言被数字化，并转化成一个矢量序列，它可以找到由一个声学前端输入的整体频谱。多年来，标准的前端模式都是用梅尔频率倒谱系数（MFCC）的矢量来表示语言的每一个帧（大概 25ms）[1]。该表达被选择呈现一帧的整个频谱包络，但抑制了基本频率的谐波。最近几年，其他的表达式流行了起来[2]。

声学模型：在一个标准系统里，语言被建模成词组序列，词组则是音素序列。但是声学表达是协同发音的结果，声音和词组里的每一个音素都相互依赖。虽然语境依赖性可以跨越

几个音素或音节，许多系统仍采用“三音子”估算近似音位，三音子即音素受到的左、右语音语境的限制条件。因此，一个词组序列是通过三音子序列的表达式来体现的。这里有许多可能存在的三音子（比如 50^3），当中又有很多极少发生。所以标准的技术就是用决策树让它们聚集起来[3]，然后为聚焦的集合建立模型，而不是针对每个三音子。

当一个单词包含了一个特别的三音子时，声学特征可以建成隐马尔科夫模型（HMM）[4]，见图 3.4。HMM 是简单的有限状态机（FSM），包括状态、转换和转换概率。而且每个状态都与一个含有可能的前端矢量的概率密度函数（PDF）相关。

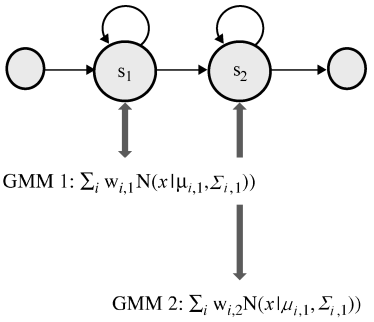


图 3.4　一个简单的 HMM 由状态、概率分布和概率密度函数（PDF）组成。PDF 是高斯混合模型，在假定一个 HMM 状态下估计一个输入帧的概率

PDF 是常用高斯混合模型（GMM）表达式的体现。GMM 是已经分析过的、易受训的 PDF，它能很好地估算任意 PDF 的结构。一个 GMM 是高斯函数的加权和；每个高斯函数可以写作：

$$N(x|\mu,\Sigma) = \frac{1}{\sqrt{\det(\sum)}\sqrt{(2\pi)^n}}\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right) \tag{3.4}$$

式中，x 是一个输入矢量；μ 是一个平均数矢量；Σ 是协方差矩阵。x 和 μ 是长度 n 的矢量，而 Σ 是一个 $n^2 \times n^2$ 方阵，且每个 GMM 是高斯函数的一个简单加权和，即

$$\mathrm{GMM}\left(x|w_{1,n},\mu_{1,n},\sum_{1,n}\right) = \sum_i \left(w_i N\left(x|\mu_i,\sum_i\right)\right) \tag{3.5}$$

3.2.3　使语音对齐隐马尔科夫模型（HMM）的过程

在语音数据流中，各个音素有长有短，因此需要校准和对齐输入帧和 HMM 的各状态，即已知输入语音帧 $\bar{O}$ 和一个 HMM 的状态序列 $\tilde{H}$，一个对齐 A 将单语调帧数映射到 HMM 状态。所以系统需要找到帧数（f）和 HMM 状态之间的最优（即概率最高）对齐 A。

$$P_{AM}(\bar{O}|\tilde{H}) \cong \max_A \prod_f (P(O_f|H_{A(f)})) \tag{3.6}$$

这常用维特比（Viterbi）算法[5]来完成。

对于每个假定的单词序列，系统会从字典中查找每个构成单词的音素的发音，然后用决策树来查找语境中每个音素的三音子。接着，根据三音子的序列，系统会查找 HMM 的状态序列。该假设的声音概率即为输入语音与这些状态最优对齐后的概率。该对齐的例子如图 3.5 所示。

3.2.4　语言模型

语言模型能够计算不同单词序列的概率，并帮助识别系统指出输入话语最可能正确的含

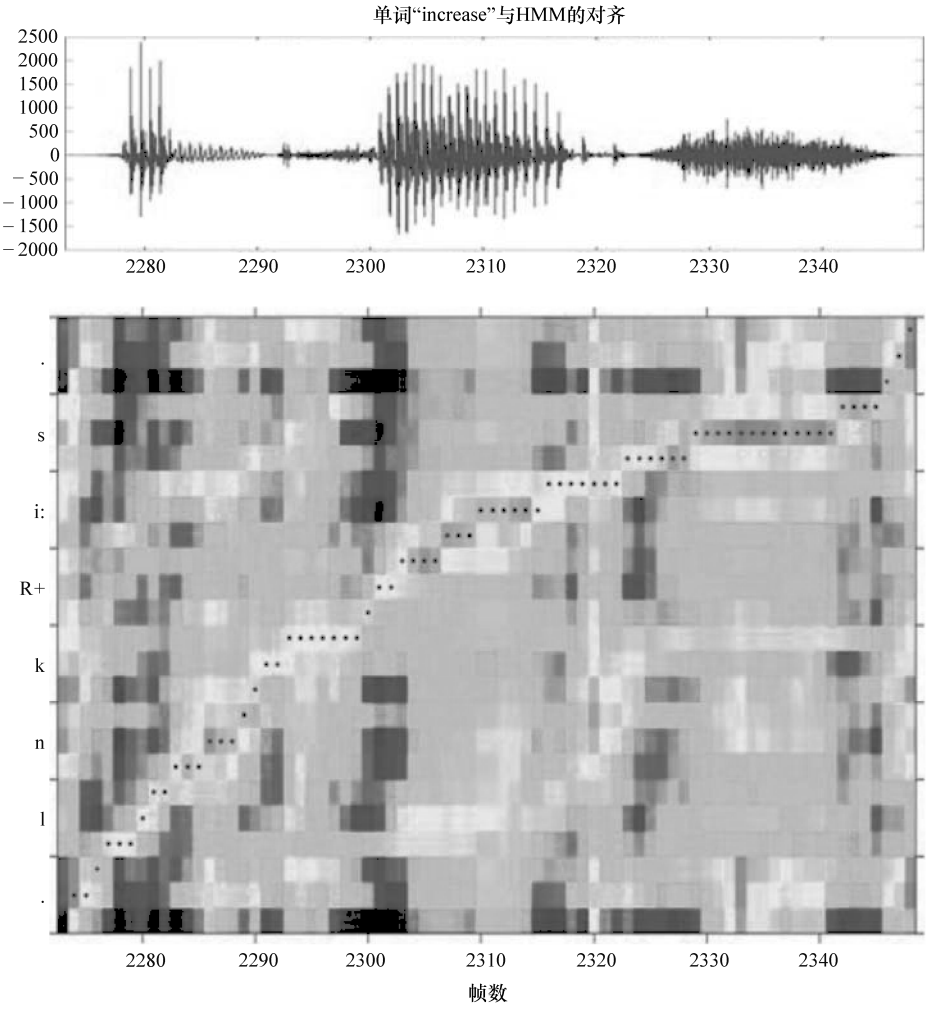


图 3.5　语音信号的维特比对齐（横轴），相对于 HMM 序列（纵轴）；浅色区域指给定帧值下受到 HMM 评估的较高概率；虚线表示对齐

义。运用于语音识别系统中的语言模型可以分成截然不同的两种类型：语法型语言模型和随机型语言模型。

语法型语言模型允许一些单词序列，但并非全部。这些语法往往取决于应用程序，支持与某些特定任务相关的话语，比如预约餐厅或发布电脑命令。这些语法规定了准确的单词序列，用户须按照这些单词序列才能指示系统行为。比如，一个预约系统的语法可能可以识别像“找一家附近的中国餐厅”“七点预订两个人的餐位”，或是“给我看看菜单”。相同的语法将无法识别诸如“辣香肠披萨”“餐厅运营的经济学分析”，或是“无色的绿色思想愤怒的沉睡”。

语法能够辨识的一组单词序列是通过诸如有限状态机或无语境语法的形式语法描述的。这些语法往往以形式体系编写，像语音识别语法规范（SRGS）（见参考文献［6］）。虽然建构简单的范例语法并不难，但是编写一个能囊括用户所有可能输入的语音的语法体系就不简单了。所以，你可能会说“附近的中餐馆”“请找一家附近的中国餐厅”“我想吃中国菜”，或“哪里有卖广式点心的”，所有这些句子的意思是一致的（对一个订餐应用程序来说），但是编写一个能包含所有选项的语法任务却异常艰巨，因为用户的表达总是各种各样的。

随机型语言模型（起初用于脱稿听写）估算了任意单词序列的概率（有些出现的概率会比其他多得多）。这样，“中国餐厅”就是一个合理的概率；“餐厅中国”相比起来的概率就小些；而“附近餐厅中国的一个找”的概率就更小了。编写一个语法型语言模型以覆盖所有可能的英文输入的尝试至今没有成功；因此一般口头命令应用程序更青睐随机型语言模型。人们发现使用随机型语言模型来设计一个实用具体的程序语言模型根本没那么复杂，NLU 处理模式还能针对某个具体的应用来设计。

随机型语言建模的宗旨是计算 $P(\tilde{W})$ 的近似值，内容定义如下：

$$P(\tilde{W}) = \prod_i P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3}, \cdots, w_2, w_1) \tag{3.7}$$

语音识别技术的一个惊人突破是一个简单的近似值算法（三元近似值）就能达到不错的效果：

$$P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3}, \cdots, w_2, w_1) \approx P(w_i \mid w_{i-1}, w_{i-2}) \tag{3.8}$$

三元近似值认为句子中下一个可能出现的单词仅仅取决于前两个单词（并且一个 N 元文法模型是对更长跨度单词序列的一般化概括）。无论从科学还是语言习惯上来说，这都是不正确的：许多话语表达都超过两个单词[7]！但是这个近似值在语音识别上的表现却非常好[8]。

3.2.5　探索：以每秒 1000 个单词完成填字游戏

找出联合优化的单词序列来描述声学观测简直就像是在玩填字游戏，声学得分限制了列数，而 LM 得分限制了行数。但是，直接考查单词序列有太多的可能性（每 10 万个单词可以产生 10^{50} 个十单词的句子）。搜索组件的目标就是找出正确的假设，尽可能少地考查其他假设的可能性。这是通过使用许多试探策略来实现的。其中一个特别重要的策略就是定向搜索，即一个接一个地排查数据帧，使得分接近最佳假设的一组假设“存活”下来。

这需要多少计算呢？在大的词汇处理任务里，就一个典型帧来说，搜索的数据内容可能有约 1000 个可行的假设。为每个可行假设更新得分需要通过计算声学模型（GMM）得分来扩展对齐，这样才能符合当前处于 HMM 状态下的假设，然后进行下一个状态，然后继续升级对齐。如果假设是在词尾（每帧约 20 个词尾假设），那么系统也需要查阅 LM 得分来寻找下一个可能的单词（每个词尾约 100 个新词）。因此，每帧我们需要约 2000 个 GMM、1000 个对齐运算和 2000 次 LM 查询。在一般 100Hz 的帧速下，我们要运算约 20 万个 GMM、20 万次 LM 查询和 10 万次/s 的对齐更新。

3.2.6　训练声学和语言模型

用在声学模型中的 HMM 是经过复杂的训练过程从大型数据集中创建的。语音数据被转录，然后提供给一个运用最大似然目标函数的训练算法。该算法估算声学模型参数，以便能够根据转录内容增大观察训练数据的可能性。这一过程的核心是自展程序，即利用引导指令将一个初始的近似声学模型输入一个改进的版本，通过按照转录内容校准训练语料并反复训练 HMM。该过程被重复多次，以期生成多个高斯混合模型，它们随后经过训练数据的考核并得出其中的高概率模型。

但是，语音识别的目标并非重现声学状态的最有可能单词序列，而是给予正确的单词序列假设比错误的假设更高的概率。这样，各个形式的区别训练就已经开发出来并用来调整声学模型，以减少有关识别错误率的各种方法[9-11]。

产生的声学模型一般有上千种状态、上万种混合模型组件和上百万个参数。标准系统使用“辅导”训练，即使用语音和相关的转录来训练。随着语音数据集的扩充，用尚未转录的或“粗略标注”的数据找到训练方案要花费很大的功夫。

随机的语言模型是经过含有数十亿词汇的大型文本数据库训练而得出的。大型文本数据库从互联网、专业文本数据库和安装的声音识别应用等地方收集。基础的训练算法比在声学训练中使用的要简单得多（基本就是一种计算方式），但是找到好数据、仔细比较数据以及处理未加观察的单词序列需要大量的工程技术。产生的语言模型常常包括数万个到数十亿个不等的 N 元词尾和数十亿个参数。

3.2.7　为特定说话人识别系统调整发声和语音模型

人们的说话方式千差万别。每个人的遣词造句都会受到其生理、口音、所受教育和说话意图风格（如宣读正式文件和日常手机短信的区别）的影响。

由此产生的不同发声可能会使识别特定说话人的语音系统出错，尤其在系统还未经过话语特征组合范例训练的情况下。反之，依照某个说话人模拟的特定说话人系统可能会比一般的语音系统获得更高的准确率。但是，用户不大可能录下上千小时的语音来训练一个声音识别系统。一般非特定的语音模型仅使用单个用户的语音数据，若能将这些模型改编成针对特定说话人的声学和语音模型，使用效果是非常乐观的。

声学模型有很多种编制方法。早期的产品经常使用 MAP（最大后验概率法）训练，它能修改被 HMM 使用的 GMM 的均值和方差。MAP 自适应经常会闹“数据荒”，因为它需要对系统使用的大多数 GMM 使用训练范例。其他更多的高效数据自适应会对所有类别的三音子（如 MLLR、最大似然线性回归[12]）修改 GMM 参数。改变模型或改变输入特征都是可行的。虽然“标准”自适应受到了“监管”（即使用带转录的语音数据），有些形式的自适应目前仍缺乏管制，使用未经查验正确的转录来输入语音数据和识别假设。

语言模型也可以根据用户或任务的不同而进行自适应。自适应既可以是调整单个参数（即根据某个特定的 N 元文法模型调整建构模型的参数，类似于 MAP 声学自适应），也可以

有效地适应参数群（类似于 MLLR）。比如在为一个新领域构建一个语言模型时，可以使用差值加权来合并来自不同语料库的 N 元文法数据。

3.2.8　“标准”系统外的其他系统

我们已经概述了语音识别系统的基本原理。我们提出了许多方法，其中比较重要的几个在表 3.1 中列出。

表 3.1　其他语音识别方案

技术	解释	参考文献
结合 AM 和 LM 得分		
比较 AM 和 LM	代替了等式 II 中的公式，$W^* = \arg\max_{\tilde{W}} (\tilde{PO} \mid \tilde{W})^{\alpha} P(\tilde{W})$ 使用了 $\alpha < 1$，因为 AM 对所有帧的观察结果并非独立	[8]
其他前端		
PLP	感知线性预测分析：找到一个语音输入的线性预测模型，使用感知加权	[2]
RASTA	“相对谱”：用来去除信道中的慢速变化和背景噪声特质	[13]
差值（Delta）和双差值	给定一个帧的前端，扩展矢量至包含求解帧与其前后帧的差值及二阶差值	[14]
降维	异方差线性判别分析（HLDA）：给定 MFCC 或其他语音特征，以及差值和双差值特征，或“叠加帧”（即一系列帧），使用线性转换来减少前端特征的维度	[15]
声学建模		
近似全协方差	在 HMM 使用全协方差矩阵需要大量的运算和训练数据。较早版本的 HMM 使用更简单的协方差矩阵版本，如“单方差”、对角协方差	[16]
声道长度归一化（VTLN）	VTLN 是一项特定说话人技术，它能基于说话人声道长度的不同修改声学特征	[17]
区分性训练		
MMIE	最大互信息估计：一种可以调整声学模型参数的训练方法，它能把正确单词序列相对于所有其他单词序列的概率最大化	[9]
MCE	最小分类错误：一种可以调整声学模型参数的训练方法，它能最小化单词被错误识别的数量	[10]
MPE	最小音素错误：一种可以调整声学模型参数的训练方法，它能最小化“音素错误”，即被错误识别的音素数量	[11]
LM		
反向预测（Back – off）	在经典的 *N* 元文法模型中，不仅很有必要预测观察到的 *N* 元文法的概率，而且还需要预测那些未出现在训练语料库里的 *N* 元文法	[18]
指数模型	指数模型（又称最大熵模型）通过把许多不同的概率估值和其他函数相乘来估算单词序列的概率，并在对数域比较这些估值。它们比 *N* 元文法模型涵盖更长的范围特征	[19]
神经网络 LM	神经网络语言模型是指数模型的延伸，经过输入数据的非线性变换，它使指示函数能够被自动确定	[20]
系统组织		
FST	为了减少冗余运算，最好呈现出音素决策树，它能把各音素映射到三音子中（三音子是构成单词的音素顺序），并把语法转换成有限状态机模型（又称加权有限状态传感器（WFST），然后将它们并入一个大型的 FSM，并优化这个 FSM。把所有的信息编入一个更为统一的数据结构有助于提高效率，但也会在动态的语法或词汇表方面出现问题	[21]

3.2.9　性能

语音识别准确度在近几十年内一直稳步提高。早期的口令系统在 20 世纪 80 年代末创立，一些用户欣然接受并成功地使用了它们，而许多其他用户则发觉其错误率很高并深信语音识别的“时机未到”。2010 年，语音识别性能的些许进步引起了大家的注意，因为纽约时报的科技专栏记者 David Pogue 报道了总体口令的错误率不到 1%[22]。虽然大多数对讲系统并未显示接近该水平的性能，但通过改进算法、增加运算和使用更大的训练数据库等联合手段，其性能仍在逐年提高。事实上在一项识别多个说话人同时说话的特殊任务中，语音识别系统能够表现出比真人更高的语音识别能力[23]。

根据作者近十年来的体验，平均单词错误率在大型词汇口令任务中已经每年减少大约 18%。这意味着获得合格性能体验的未受训用户人口比例在稳步逐年增加。该进步不仅让我们能够面对诸如语音搜索的技术挑战，还能有机会应付更具挑战性的使用环境，比如车内语音控制。最后，准确度的提高意味着语音识别已经成为了解决复杂自然语言处理的有效前端，从而催生出一批崭新的界面程序。

3.3　语音识别的深度神经网络

稳步改进的“标准”语音识别系统由于深度神经网络（DNN）的创立而在近几年间遭到了阻断。深度神经网络是一种人工神经网络（ANN）的形式。ANN 这种运算模型在大脑的刺激下，能够进行机械化学习和模式识别。它们可以被视为相互连接的神经元，经过神经网络获取信息，从而运算出输入数据的数值。

正如其他机器学习方式，神经网络已经被用来解决了许多普通按规则编程难以处理的问题，包括电脑视觉和语音识别。

在语音识别领域，ANN 在 20 世纪 80 年代末和 90 年代初曾一度流行。这些早期的、相对简单的 ANN 模型并未真正意义上超过基于 GMM 的 HMM 和声学模型的成功组合。研究人员利用含有单层非线性隐单元的人工神经网络，以期从声学系数范围中预测 HMM 状态。在这个方面他们还是取得了一些成功[24]。

但是在那时，硬件和学习算法都不足以在大量数据信息中测试含有许多隐层的神经网络；无论是使用含有单一隐层的神经网络，还是使用脱离语境的音素作为输出，两者的性能优势均不足以真正地挑战 GMM。因此，当时神经网络的主要贡献实际在于为 GMM 提供额外的特性，或者说提供了使用 ANN 的“瓶颈”系统来为 GMM 提取额外的特性。ANN 当时在语音识别系统和有限的几个商业产品中取得了一定的成功[25]。

几年前，大多数语音识别系统仍是通过在 GMM 的基础上使用 HMM 来建模 HMM 发射分布的。直到最近，新研究才证明了混合声学模型运用了更为复杂的 DNN，在局部最优环境中测试很少出现“卡壳”，因而能够极大改善小音素识别任务的性能[26]。这些结果后来被应用到一个大型词汇语音搜索任务中[27,28]。从那之后，几个测试组也因为在大型词汇持续

语音识别（LVCSR）任务中使用了深度神经网络声学模型而取得了很大的收获[27]。按照这个趋势，DNN 嵌入系统将很快成为语音识别领域的最新最前沿的技术。

在实践中，用作语音识别的 DNN 是数个多层感知器神经网络。每个网络含有 5 ~ 9 个层，每层 1000 ~ 2000 个单元。尽管 20 世纪 90 年代使用的 ANN 输出的是脱离语境的音素，但是 DNN 使用了数目庞大的绑定状态三音素（像 GMM）。两个模型的比较如图 3.6 所示。

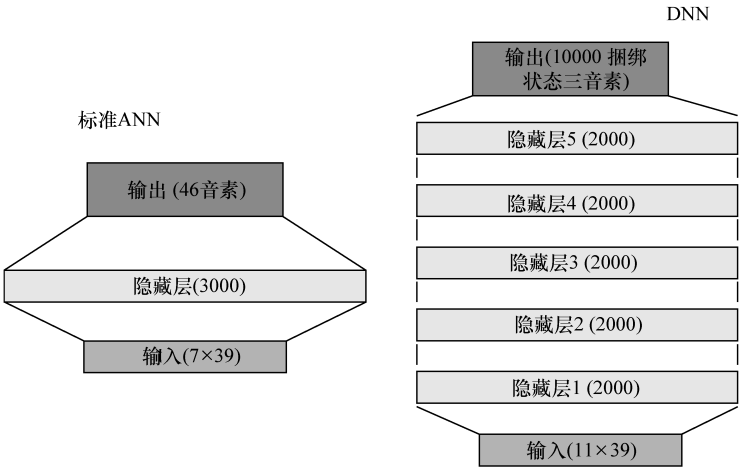


图 3.6　20 世纪 90 年代用于 ASR 的标准 ANN 与现在使用的 DNN 的比较

DNN 经常与局限型玻尔兹曼机器（Restricted Boltzmann Machine）算法一起预训练，并利用标准反向传播进行调试。分段信息通常由现存的 GMM – HMM 系统生成。DNN 训练方案包括许多显著的环节，如图 3.7 所示。

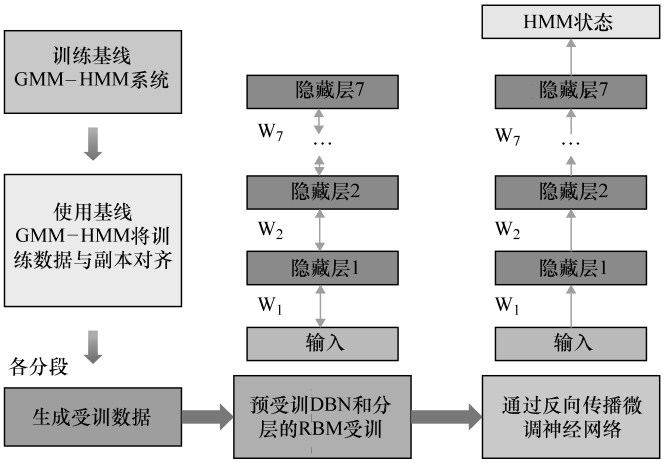


图 3.7　DNN 训练

运转时，DNN 是一个标准的前馈式神经网络，它含有多层反曲形的单元和一个最顶层的 softmax 单元，可以在传统或并行的硬件上高效执行。

DNN 被 ASR 使用的方式有两种：

1）用 DNN 来为 GMM 提取特征（即受限特征）。这可以通过在 DNN 中插入一个受限层并把该层中激活的各个单元用作 GMM 的特征。

2）直接在解码器（DNN – HMM 混合模型）中使用 DNN 的输出（绑定三音子概率）。

第一种方法可以对现有的基于 GMM 的 ASR 系统实施快速改进，错误率减少 10% ~ 15%，但是第二种方法的改进效果更大，较最新的 GMM 系统能常常减少 20% ~ 30% 的错误。

神经网络作为高质量声学模型在近期重获好评的主要因素有三个：

1）更深层次的网络的使用使其更强大，因此深度神经网络（DNN）代替了浅层神经网络。

2）正确的初始化系数和使用更快的硬件使其能够有效训练深度神经网络：DNN 与局限型玻尔兹曼机器算法一起预训练，并使用标准反向传播进行调试；GPU 用于加速训练。

3）使用大量依赖语境的输出单元而不是脱离语境的音素。一个含有大量 HMM 的绑定三音子状态的大型输出层极大地提高了 DNN 的性能。重要的是，该选项使解码算法大体上保持不变。

其他出现在 DNN 训练方案内的重要发现[27]包括：

1）DNN 对滤波组件输出的作用效果比 MFCC 要好得多。实际上它可以应付关联输入特征，比起提前改变的特征，它更偏好使用原始特征。

2）DNN 比 GMM 对说话人的敏感度更低。其实使用特定说话人的方法相较于非特定说话人 DNN，并没有得到很大改进。

3）DNN 在嘈杂语音中性能良好，结合了许多去噪预处理方法。

4）使用标准逻辑函数神经元有一定道理，但可能不是最佳方案。其他单元，如修正线性单元可能更具发展潜力。

5）相同的方法可以用作应用程序而不是声学建模。

6）DNN 结构可以以不同的方式应用于多任务（如多语言）学习，而且 DNN 比 GMM 在抽取某个任务数据和改进相关任务性能方面要有效得多。

3.4　硬件优化

前面描述的算法要求很大的电脑运算资源。芯片制造商越来越意识到语音界面的重要性，因此他们正在开发专门为语音优化的处理器结构和 NLU 工艺，以及其他的输入传感器。

现代用户除了享受桌面电脑和电视机之外会使用不少移动设备（笔记本电脑、平板电脑、智能手机、定位系统），但是往往受到电池续航能力的制约。这些设备本身已经变得越来越复杂，集多种功能于一身，供应商积极参加“军备竞赛”，较量谁能在下一个最畅销的

必备产品中拔得头筹。

虽然用户期待功能的增加，但是他们对电池使用时间的期待却并未降低：笔记本电脑应该可以待机几小时；智能手机可以至少一天不用充电。然而电池是设备的一部分，影响着设备的重量和大小。

3.4.1　低电量唤醒运算

移动设备因此需要减少耗电量。软件可以暂时叫停一些未使用的功能（蓝牙、无线上网、相机、全球定位系统和麦克风），并在需要使用时快速启动它们。设备还甚至可以进入省电模式，让系统处理越来越少的任务。联想到遵守节能星（Energy Star）国际标准的电视机和其他装置，它们通常都设有超过三个状态（开机 – 关机 – 待机）的调节。那么，系统是怎样“唤醒”的呢？用户的一个动作是当前最普遍的控制方式，比如按下设备的开关键。

但是今天的设备装上了各式各样的传感器来实现这个目的。红外传感器能够检测遥控信号；光传感器能够在被掏出口袋后开启；运动传感器可以侦测到动作的发生；相机可以定焦于人；麦克风声音唤醒可以感应到声音活动或一个特殊的短语。

这是通过低能耗、基于数字信号处理（DSP）的“唤醒口令”识别来实现的。它可以使用户对设备发出口令而无需先将其打开，进一步减少了区分用户意图和期待结果的步骤数。比如，英特尔超极本（Intel – inspired Ultrabook）就集成了这些功能，在听到“你好，小龙”后，它能马上被唤醒并聆听用户的命令或听写文字。

至此安全问题开始浮现。电视机响应已知的信号，无论该信号是否为原始信号。任何人用遥控器就能操控它，或者说，任何一个匹配的遥控器即可。客厅虽然通常最多只有一台电视机，可会议室可能会有 20 个人，人人都可能有手机。要是某人在尝试唤醒自己的手机的时候把别人的也唤醒了，那可就太不受待见了！因此，我们需要用个性化设置来增加安全性。运动传感器只会对某些动作有反应。数码相机只会响应某个（些）用户，该技术又称“脸部识别”（Facial Recognition），某个声音唤醒只能感应特殊的用户的专门口令——“语音生物识别技术”。

3.4.2　特定运算的硬件优化

这些传感器都会耗电，特别是在它们运行的时候，主 CPU 运行的程序会耗用大量电能。在开启音频系统的全部功能时，包括多个麦克风、回声消除和波束形成等都会大量耗电。制造商因此需要研制特殊的硬件以减少这些传感器的电力负荷，或依赖通常比主 CPU 的运行速度慢的 DSP，其速度约为 10MHz 而不是 1 ~ 2GHz。

与单一 N 维高斯模型有关联的概率密度函数（PDF）如式（3.4）所示。一个高斯混合模型（Gaussian Mixture Model）总的 PDF 是各个 PDF 的加权总数，有些系统可能有 10 万个或更多的 PDF，可能需要被每秒估算 100 次。优化运算（仅计算“可能的”PDF）和模型估计（如假定协方差矩阵呈对角线形）均应用于减少计算负载。单指令多数据（SIMD）计算机硬件的出现曾是一个重大突破，因为它使这些线性代数算法能够每次处理四个或八个特征。

最近在图形处理单元（GPU）的使用上又取得了进展。一开始 GPU 主要用来加速 3D 电脑图像（特别是游戏），这会大量使用线性代数。GPU 在上述那些方面帮助了 PDF，但有证据表明它运算 DNN 的效果特别显著。

如前面所述，DNN 有许多节点层，每个节点层都是几乎线性处理的结果，该线性结果被应用于每层节点层下方的节点层。有 1000 个节点的层是正常的，通常层数为 5 ~ 10 个，因此有效的应用 DNN 需要计算 5 ~ 10 个矩阵向量乘法，每个矩阵是 1000 × 1000 阶，且该计算每秒进行多次。训练 DNN 要耗费更大的运算资源。近期研究表明，训练非常少量的数据可花费三个月的时间，但是使用 GPU 可以将时间缩短至三天，时间上减少了 30 倍之多[28]。

3.5　稳健语音识别的信号强化技术

在真实场景里的语音识别应用，接收的语音信号通常会夹杂许多干扰有声信号，比如背景噪声、扬声器发声、冲突声音或回响等。在麦克风离说话人较远的时候尤其如此，比如，在车里或家里的应用。最糟糕的情况是干扰的信号甚至超过目标信号，使语音识别器的性能严重降低。语音技术作为人机交互的一项基本高效工具正变得日益重要，这使得在恶劣环境下的系统抗噪能力成为影响语音对话系统的核心因素。

3.5.1　稳健语音识别

抗噪性可以通过调整语音识别过程来实现，或者通过一个专用的语音增强前端。当前的系统通常使用两者的结合。

稳健语音识别的前沿技术通常包括使用诸如 MFCC 或神经网络这样的抗噪特征，并用噪声夹杂的语音数据来训练声学模型，这些数据往往代表了在正常应用中经常出现的各种噪声。但是由于声学环境纷繁复杂，训练不可能涵盖所有的噪声情景。于是人们发明了若干种根据噪声环境快速改编声学模型参数的方法，这些噪声短暂地出现在输入信号中。例如，该技术已经成功地使长距离对话声音在变化的回音环境中保持稳健。

语音增强算法可以大概分成单通道法和多通道法。由于各种噪声来源和环境的具体统计属性，并不存在一个涵盖所有信号和干扰的统一解决方案。根据应用程序，语音增强前端常常结合多种方法。最普遍的是把单一通道噪声和诸如消噪声、空间滤波的多通道技术结合起来使用。

3.5.2　单通道噪声抑制

单通道噪声抑制技术主要是基于频谱加权法。在这种方法中，信号一开始被分解成叠加的数据模块，每个模块时长约 20 ~ 30ms。随后每个模块通过使用短时傅里叶变换（STFT）或合适的解析滤波器组转换成为频域或子带域。接着，噪声信号的频谱内容由衰减系数加权，衰减系数根据估计瞬时信噪比（SNR）函数在频带或子频带中进行计算。选择了该函数的结果是有低 SNR 的频谱内容被衰减，而有高 SNR 的则没有。这样做的目标是为了得到一个免噪语音信号的频谱系数的最佳估值。由于频谱系数得到改进，一个无噪声的时间域信号

就能被合成出来并传输到识别器中。另外，特征抽取可以直接在改进的频谱系数上进行，避免了把频谱系数转回到免噪声的时域中。

目前大量用来计算频谱加权函数的线性和非线性算法已经开发。这些算法主要在基本优化标准以及对语音和噪声的统计特征的假设上存在差异。加权函数的最普遍范例是谱减法、威纳滤波器（Wiener filter）和最小均方误差（MMSE）估算器[29]。单通道噪声抑制方案如图 3.8 所示。图 3.9 显示了在应用所描述的频谱加权系数之后，噪声短语“Barbacco has an opening”的频谱图和增强信号的频谱图。

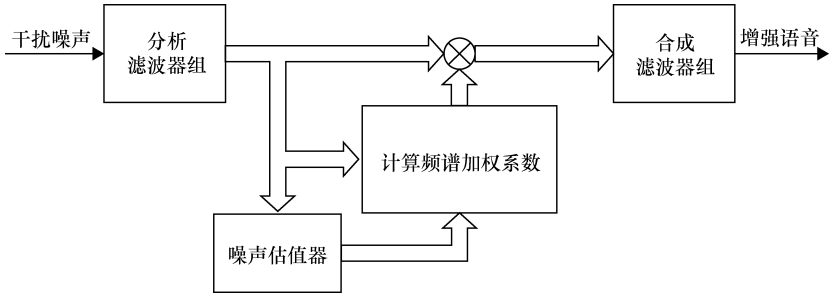


图 3.8　基于频谱加权的单通道噪声抑制方案的框图

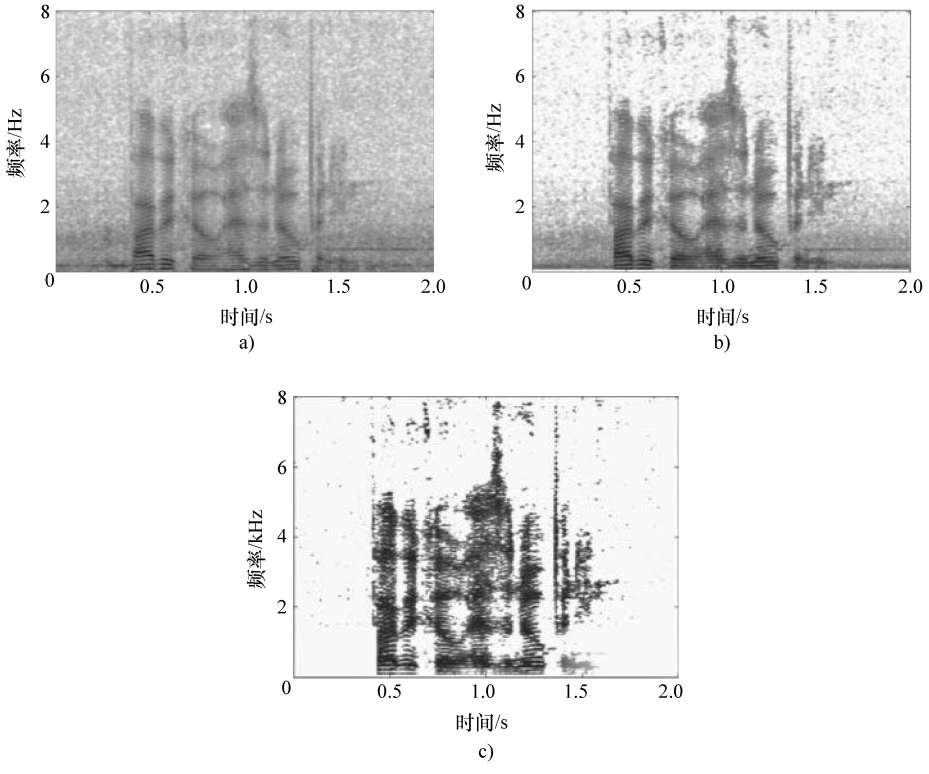


图　3.9

a）噪声的时频分析　b）增强语音信号的时频分析　c）应用衰减系数的噪声语音信号的时频图

单通道噪声抑制算法对诸如像空调风扇、电脑风扇、车内行驶噪声等平稳背景噪声非常有效，但它们却不适合诸如说话或音乐等波动干扰源。在单通道系统中，背景噪声大多只能在语音暂停时被追踪到，因为在嘈杂的语音信号中，声音和干扰叠加所产生的频率通常较高，使单通道的减噪方案主要被限制在时间变化慢速的背景噪声，而这样的噪声在话语活动中变化并不大。

目前已经为克服这个局限提出了若干个优化方案，包括利用显化清晰的语音模型或语音和具体干扰项的时空特征，以实现语音和波动噪声的分离。有效的方法能够减少风扇声、敞篷车的风吹声[30]、高速脉冲噪声或模糊不清的声音。

单通道噪声抑制的另一个缺点是频谱加权技术对声音的固有扭曲，这极大地影响了低信噪比。由于该方法依赖于 SNR 产生衰减，当背景噪声增加时，会有越来越多的目标信号内容被抑制。递增的语音扭曲因此会降低识别器的性能。

3.5.3　多通道噪声抑制

不像单通道噪声抑制，多通道的方法能减少声音扭曲并增加抵抗波动干扰的效力。其缺点是增加运算的复杂性，而且需要有额外的麦克风或输入通道。多通道方法可归为噪声消除技术，主要利用不同噪声参考通道和空间滤波技术，如波束形成法（见下文讨论）。

3.5.4　噪声消除

在相关的噪声参考存在时可以使用自适应噪声消除[31]。这意味着位于主通道（即麦克风）和参考通道的噪声信号是单一噪声来源的线性变换。自适应过滤器用于找到能把参考信号投射到主信号内噪声的转换功能。通过用转换功能过滤参考信号，主通道内的噪声内容估值就可以计算出来。随后，噪声估值从主信号中减除，获得改善的语音信号。自适应噪声消除的原理如图 3.10 所示。因为信号和噪声在麦克风的线性叠加，信号的减除并不会导致任何语音扭曲，只要确保进入参考通道的不相关的噪声内容和目标语音信号的串扰足够小。

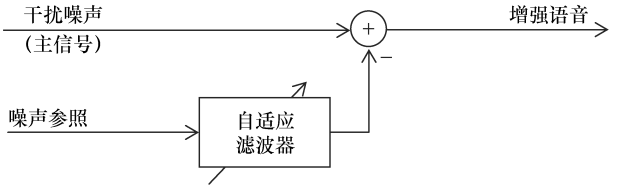


图 3.10　噪声消除器的基本结构

噪声消除的有效性实际上高度取决于是否有适合的噪声参考，而这又因具体的应用程序而异。噪声消除技术在手机上的应用很成功。这里，参考话筒通常安装在离主话筒尽可能远的地方——通常在电话的上方或后部——以减少语音信号泄漏到参考通道中。但噪声消除在车载应用中减少背景噪声方面就没有那么好的效果了，因为强劲的风声和轮胎噪声具有漫分布特性。因此，如果主话筒和参照话筒分开超过几厘米远，关联性就会大打折扣，导致大量语音信号无法避免地泄漏到参考通道中。

3.5.5　回音消除

一个经典的噪声消除应用程序是移除干扰扬声器的信号，这又被称为声学回音消除

(AEC)[32]。该方法是用来移除远端用户使用免提接听电话的回音。在声音识别中，AEC 被用来移除语音对话系统中的提示音或家庭影院、移动设备播放的立体声信号。

与上面描述的噪声消除器相似，扬声器参考声音经过自适性滤波器处理，能够获得扬声器发声在话筒信号内的估值。声音环境可能变化迅速——比如，人们在房间内的移动，这就使自适性滤波器的快速追踪能力对有效移除扬声器噪声至关重要。

归一化最小均方（NLMS）算法因其稳健性和简易性的优点而被广泛使用于调节自适性滤波器的过滤系数。该算法的缺点是，如果干扰信号的音频波动很大，算法的收敛速度就会降低，就像演讲或者音乐的噪声。所以 NLMS 常常在频域或子带域使用。由于在单频率子带域中的频谱动态通常比整个频率范围要低得多，追踪行为可以有显著的改善。另一个在子带域工作的好处是 AEC 与诸如噪声降低的频谱加权技术可以实现高效合并。

3.5.6　波束形成

当话语被一系列话筒组合捕捉到时，产生的多通道信号也会包含关于声音来源的空间信息。这促成了空间滤波技术，如波束形成。该技术能从目标方向中抽取信号同时减弱其他方向的噪声和震动。自适性滤波技术[34]可以把波束形成器的空间特性调整至实际的声场，从而有效抑制了移动的声源。不过，这种自适性波束形成器的方向性取决于话筒的数量，在应用设备中常常因为成本而限制在 2 ~ 3 个。

为了改良方向性，波束形成器可以与一个叫作空间后置滤波器的装置捆绑[35]。该装置是基于应用于降噪的频谱加权技术，不过它使用了自适性波束形成器的空间噪声估值。虽然空间滤波可以显著地减少干扰噪声或冲突扬声器，一旦扬声器没有安装在指定位置，它依然是有危害的。因此必须具备稳健的扬声器本地化系统，特别是在移动的情况下，目的地的方向往往随着用户移动或设备倾斜而发生改变。

实现音效本地化的一个简单方法是选择声音强度最大的方向[36]。若在相对接近设备的地方安装一个扬声器，例如使用平板电脑时，该方法的效果就还可以。但若用于智能电视或其他智能家电时，离设备较远处可能会同时有好几个扬声器。这使得声源本地化无法稳定进行。因此，较好的做法是通过相机追踪用户视线并关注那些面朝设备的扬声器。另一个方法是用手势来提示设备应该注意哪个声源。

如前面所述，声音增强前端常常结合若干种技术以有效应对复杂的声音环境。图 3.11

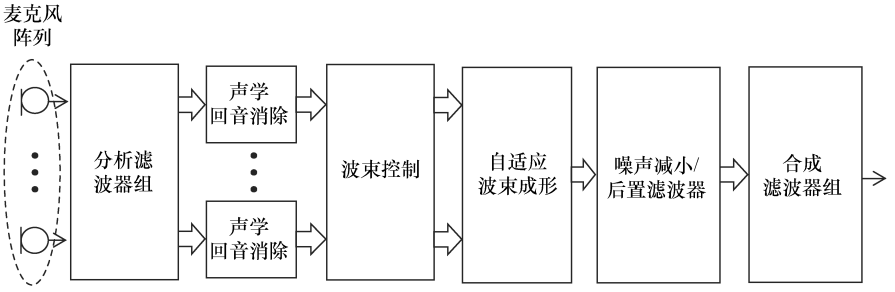


图 3.11　远距离控制电视机的语音增强前端

列举了一个远距离操控电视的语音增强前端。该系统内，回音消除的使用移除了多通道娱乐信号，同时波束形成和噪声降低的功能用于抑制诸如室内谈话和背景噪声等干扰声源。

3.6　声音生物计量

3.6.1　引言

许多移动设备安装的声音驱动型应用需要核实用户身份。这有时是出于安全的需要（比如，用户可以进行金融交易），有时是因为要保证语音命令是由设备的主人发出的。

声音生物计量通过人的声音样本辨识身份。主要使用的商业应用是说话人验证。所要求的身份是通过比较在注册和验证环节的声音样本来验证的。把声音样本和一组多个注册用户进行匹配也是声音生物计量学的一种应用。最后，如果录音包含来自多人的声音数据，比如在代理人和顾客之间的对话中，“说话人分类”从每个用户身上抽取声音数据。所有这些技术都在人机交互中发挥着作用，特别是有安全考虑的情形中。

声音生物计量学将会是移动用户界面的核心组件。传统安全方法主要采用了包括个人身份号码、密码、口令牌等乏味的措施，在与移动设备互动时特别笨拙不便。声音生物计量提供了一种更为自然方便的核实用户身份的方法。它有多种应用，包括诸如查收电子邮件和唤醒移动设备等日常活动。想要实现“瞬间唤醒”，不仅需要用词完全正确，而且必须由机主本人启动才行。这有利于省电和防止未授权的设备介入。其他应用包括手机银行交易和购物许可等验证。

致力于开发和改进说话人验证、身份识别和分类的技术在过去 50 年中取得了不小的进展。虽然早期的技术主要聚焦模板式途径，如动态时间规整（DTW）[37]，但它们已经朝着诸如 GMM（1.5.2 节已经讨论过）这样的统计模型发展。最近的说话人识别技术已经采用 GMM 作为人声模型建构的初始步骤，随后又在冗余属性投影（NAP）[40]、联合因素分析（JFA）[41]和全要素分析（TFA）[42]中应用。TFA 途径产生了紧凑的人声表达式，又称为 I 矢量（或身份矢量）。这些都是声音生物计量学的前沿发展成果。

3.6.2　声音生物计量面临的挑战

其中一个声音生物计量学的主要挑战一直是减少由于错误匹配注册与验证声音而产生的错误率。比如，当人们用手机注册了自己的声音，又在个人电脑上验证网上交易的时候，错误就可能发生。此情况下错误率增加的主要原因是电脑麦克风和用来录制的频道不匹配。这一问题已经得到了研究人员的广泛关注，并能够由 NAP、JFA 和 TFA 途径成功的解决。但是新呈现的应用有必要进行进一步的研究。另一项任务是应对“声音老化”。这是指由于注册和验证的间隔时间逐渐拉长而导致的验证准确率下降[43]。模型自适应调整是一个可能的解决方案，即注册后的模型可以随着验证过程中的数据特点变化而改变。当然，这只能在用户经常介入设备的前提下可行。

声音生物计量的另一项挑战是以最小的声音数据维持可接受程度的准确性。这是商业应用的一项基本要求。在“依赖文本”的说话人识别中——相同的词组必须用来注册和验证——2 ~ 3s（或 10 个音节）通常能够产生足够的准确度。但是如有些在移动设备上使用唤醒词的应用则需要时间更短的话语来验证用户。

虽然把握时间信息和使用定制的背景建模能改进准确度，但这个问题一直是一个挑战。相似的，独立文本的说话人验证——用户能在注册或验证时说出任何短语——通常 30 ~ 60s 就能够产生足够的准确度。但是说话人验证和身份识别性能是经常需要用较短的话语完成的，比如在向移动设备发出声音命令以及与客服中心的代理简短谈话的时候，等等。美国国家标准与技术研究院（NIST）已经赞助了许多包括验证较短话语的说话人识别评估项目[44]，该问题仍是目前研究关注的领域。

3.6.3　声音生物计量的新研究领域

声音生物计量技术自诞生以来虽然已经取得了重大进展，但仍有许多领域值得进一步的探索。应对“欺骗攻击”的措施（用录音回放、声音拼接、声音转化和文本朗读技术）仍显不足。许多类似的攻击手段已经在国际语音大会上讨论[45]。持续的研究致力于评估这类攻击的风险并尝试预防和阻止，主要通过改进生物特征识别策略和语音合成的检测算法。

声音生物计量是未来语音交互系统的一个趋势。虽然语音识别、自然语言理解和文本朗读的开发时间更早，但声音生物计量技术正在以前所未有的速度为商业和政府部门提供服务。它拥有验证身份或定位已注册用户在某一地点的便捷手段，减少了身份盗窃、诈骗钱财和安全威胁等风险。近期在计量算法上取得的突破增加了用户群体并促进了该项技术的广泛应用。

3.7　语音合成

许多手机应用程序不仅能识别并执行用户的有声输入，而且能将语言信息通过文本语音合成（TTS）向用户展示。TTS 有丰富的过往发展经验[46]，许多元素已经得到标准化。如图 3.12 所示，TTS 有两个组件：前端（FE）和后端（BE）处理。前端处理从文本分析中获取信息；后端处理将该信息依照以下两个过程转为声音：

- 首先，它在存有预先分析的语音数据的索引知识库中搜索，找到与前端提供的信息最关联的索引数据（单元选择）。
- 其次，该信息被语音合成器使用，以生成合成语音。

预先分析的数据可能被存为编码语音或一组用来驱动语音产出的模型参数，或两者同时存在。

图 3.12 中，前端被分成两个组成部分：文本预处理和文本分析。“真实世界”中的应用需要文本预处理，这些应用程序中的 TTS 系统应该阐释大范围的数据格式和内容，包括短小、语体特色鲜明的对话提示和长篇的、结构复杂的话语。文本预处理视具体应用而定，

比如，需要阅读从数据库抽取的顾客和产品信息的预处理将与阅读从 RSS 源获取的新闻截然不同。而且，文件可能包含辅助浏览器内可视化阅读的标记或在页码上的标记，比如标题、章节名称等。预处理器必须重新阐释该信息，使其产出能够按照文本的结构表达。

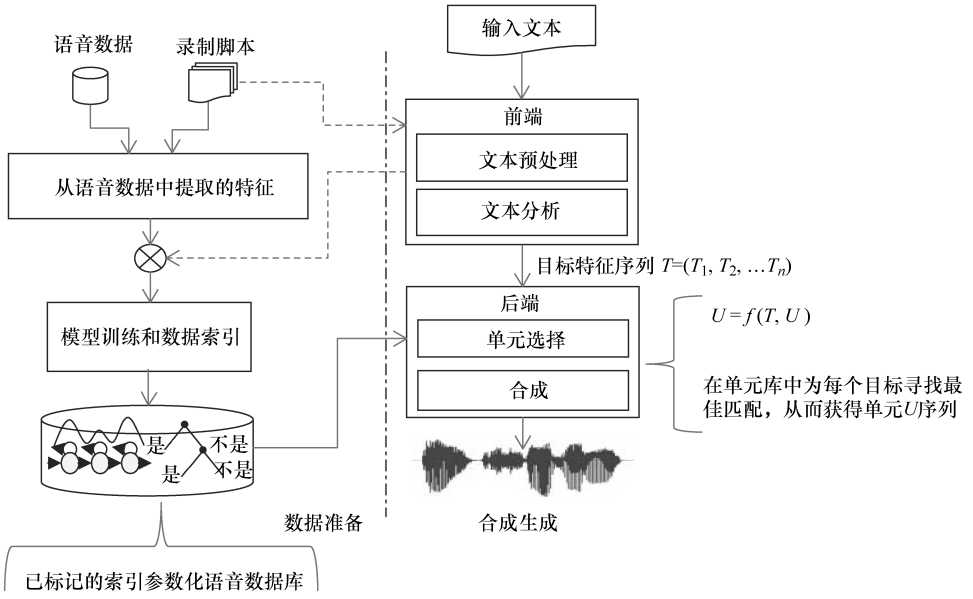


图 3.12　语音合成结构

文本分析可以分成四种处理活动：符号化和标准化，句法分析，韵律预测以及字音转换。符号化有助于合理解析正确的拼字。比如，一个电话号码在写下来后能够被识别，并在阐述的时候会表现出常规的韵律结构。在符号化的过程中时，各字母被分归进了符号组，一个符号就是一串从属于定义类别的字符。一个数字就是一个简单符号的例子，而电话号码就是一个复杂的符号。符号化在像汉语这样的书写体系中十分困难，因为句子是以汉字的顺序书写，汉字之间没有书写间隔。

文本标准化是把正确的拼字转化成扩展的标准化表达式的过程［如$5.00 就被扩展成“five dollars”（五美元）］。该过程是下一步句法分析的前提。句法分析通常包括部分语音和确定稳健的句法结构。这些处理有助于语音发音的筛选和韵律结构的预测[47]。

韵律可以定义为语音的节奏、强调和语调，它是交流说话人意图（如问题、陈述或命令）和感情状态的关键[48]。在声调语言中，字的意义与具体的声调规律之间也存在着关系。韵律预测组件通过使用具有象征意义的信息（如强调模式、语调和换气单位）和参数信息（如音高、振幅和长短轨迹），能够在韵律上表现编写在文本内的深层含义和结构。参数信息可以量化并在筛选过程中或直接在参数合成器中作为一个特征来使用（或两者同用）。

在大多数语言中，字素（即字母）与声音的表达（即音素）是非常复杂的。为了简化筛选正确声音的过程，TTS 系统首先将字素序列转化成音素序列，以便更贴切地表达要发出

的声音。TTS 系统通常结合使用大型发音词典和字素到音素（G2P）规则来把输入转化成一个音素序列。一个发音词典包含了数以万计的词条（通常是词素，但也有成形的单词），每个词条都含有单词发音的语音表达，但有时也有其他诸如词性的信息。发音可以直接从词典中获取，也可以通过结合单词的形态解析和词汇查询来获得。没有哪个词典是完整的，因为新的单词会从语言中持续生成。各 G2P 使用语音学法则来为词汇表之外的单词生成发音。

生成音素序列的最后一步是后词汇处理，也就是影响了连音、吞音、删减和韵母弱化的持续语音生成被应用到音素序列中[49]。根据说话人的调整也可以应用于把词典存储的或 G2P 规则生成的范例发音转化成合乎习惯的发音。

如前面所述，后端包含两个阶段：单元筛选与合成。在两个广泛使用的合成形式中更受青睐的是拼接合成，即由单元索引的选定声音片段有选择的组合一起。诸如基音同步叠加法（PSOLA）这样的信号处理方法可以用来修整衔接处并提供更强的韵律控制，虽然这会导致一定的信号退化[47]。参数合成常用 HMM 合成法，即使用频谱帧和激励参数来驱动一个参数语音合成器[50]。

表 3.2 指出了拼接法和参数法的不同。如表所示，拼接法保证了最大的忠实度，却牺牲了灵活性和规模；参数合成在小规模的基础上提供了很大的灵活性，却牺牲了忠实度。因此，参数方案通常使用在存储空间有限的嵌入式应用中。

单元筛选[51,52]尝试从已生成的数据库中寻找单元 U 的最优序列，数据库中描述了前端为分析句子而生成的目标序列 T 的特征（见图 3.12）。两个试探性获得的成本函数被用来限制搜索和筛选。这些是单元成本（数据库中的单元特征与目标序列中的元素的匹配近似度）和联合成本（附近单元的匹配程度）。通常动态编程用来建构全局中最优单元的序列，以减少单元和联合成本。

$$U = \underset{u}{\operatorname{argmin}} \sum_{n=1}^{N} Unit(T_n, U_n) + \sum_{n=1}^{N-1} Join(U_n, U_{n+1})$$

表 3.2　拼接法和参数法的不同

类别	拼接分析	参数分析
语音质量	质量不等，最好高度自然。通常有好的分段质量，但可能会韵律较差	语音质量一致，但是具有合成“处理的”特点
语料库大小	质量关键取决于声音数据库的大小	训练少量数据时表现流畅
信号操控	低至零	默认信号操控。适合说话人和语体自调整
基本单元拓扑	波形	语音参数
系统空间占用	简单的言语存储编码导致占用较大的内存空间	深度建构语音信号模型导致占用较小的系统空间。系统具有弹性，可以减少系统空间占用
产出质量	质量取决于从单元存储中选择的持续语音的长度。比如，限域系统，往往在选择中趋向于产出较长的存储语音，其合成也更为自然	顺畅稳定，较之前未能看到的上下文来说更具有可预测性能
语料库质量	需要准确标记的数据	可容忍标记错误

在 HMM 选择，目标序列 T 被用来建构一个 HMM，参考来自语境集群的三音子 HMM 的拼接。得出的最优序列的参数矢量可以对下式进行最大化：

$$O = \arg\max_{o} P(O \mid \lambda, N)$$

式中，O 是要被优化的参数矢量序列；λ 是一个 HMM；N 是序列的长度。不同于单元筛选法是基于局部单元成本和联合成本来决定最佳性，统计法则设法构建一个避免突然阶跃变化的最优序列，通过考虑二阶特征来实现[50]。虽然还是未被广泛采用，现在一个新兴的趋势是混合这两种方法[53]。混合法使用状态序列来共同生成参数和单元的候选序列。对于使用哪种方法的决定需要在每一个状态下做出，且基于语言的语音规则和对参数方案强大建模功能的理解。

生成自然合成语音有两大最根本的挑战。首先是表达式，它是 FE 能够辨识和稳健抽取特征的一种能力，抽取的特征与在有声语言中观察到的特征一一对应；伴随相关的是另一种能够查找并标注相同特征的语音数据的能力。一个索引了极少特征的语音数据库会生成较差的单元识别力，而只能生成一组索引特征的 FE 将导致数据库中的单元永远无法用作训练或筛选。换句话说，FE 的表达能力必须匹配索引的表达能力。

第二项挑战是贫乏性，即必须存在足够的声音样本来充分展示 FE 生成的特征表达能力。在拼接合成中，贫乏性意味着系统被迫选择一个匹配不足的声音，仅仅是因为它无法找到充分的近似值。在 HMM 合成中，贫乏性导致产生了训练不足的模型。听觉效果的贫乏性随着语体越发丰富而增加。通过构建能够从高层特征中生成合成声音的语音模型，贫乏性能够在某种程度上因为这些强大的模型而得到缓和。最近，诸如 CAT（集群适应性训练）[54] 和 DNN（深度神经网络［Zen 等，2013］）[55] 这样的技术已经得以应用，通过避免分段造成的贫乏性效果增加，它们能够最优化现有的训练数据。

如表 3.2 所示，拼接法取得的商业成功主要由于高度忠实的合成技术是可行的，只要小心控制好录制语体并确保在构建语音单元数据库时，在重点应用领域有足够的声音覆盖。用相对简单的 FE 分析和简单的 BE 合成是可以取得令人意外的优质成果的。但从技术上来说，这些方法有可能会逐渐陷入困境。虽然这些系统服务于许多传统市场，但它们还是比较昂贵，生成也比较费时。

高度表达个性化代理日益增长的商业需求正不断推动可训型系统的开发。在 FE 方面，统计分类器正在取代规则式的分析方法；在 BE 方面，数据筛选和混合参数系统正在促成灵活性与忠实性的相互结合[53]。想要合成诸如新闻和维基百科词条这样的复杂文本的决心鼓励着开发者思考如何把语义学和语用学的知识灌输到 FE 中，也因此需要考虑如何在 BE 中实现抽象概念与其声学实现的复杂数据匹配[54]。

3.8　自然语言理解

我们已经谈到，语音是与移动设备交流的一种特别有效的方式。用户的语音构成了特定系统运行指令和系统获取相关信息的请求。用户的话语首先经过自动语音识别模块转换成文本。随后已辨识的文本经由自然语言理解（NLU）模块处理后，语义从声音中抽离出来。在真实场景中，已辨识的文本可能会有错误，因此通常的做法是输出一个备择假设的

“top – N”列表，或一个结果网格，这样 NLU 就能探索其他的备选答案。意义的准确抽离对系统执行正确指令或获取目标信息十分必要。

NLU 模块的复杂性取决于系统提供给用户的各种功能和预设的用户语言变化。现在许多口语对话系统受限于能够执行的任务范围，并需要有限的、可预测的语音输入来完成那些任务。例如，餐厅的预订系统就要求在标准指令模板内填写一系列数据（餐厅名，时间，就餐人数）。同样的，一个电视界面可能只需某个电影或节目的规格参数就能决定其播放的频道并在屏幕上播放。

有些系统有高度的系统主导性。它要求完整具体的问题来对号入座，并且期待回答仅限于所问的问题。例如，针对提问“你想要给谁打电话?”，如果可识别的文本匹配得上一个已知的姓名，数据库搜索可以填上这个电话号码。或者，针对问题“你的航班是哪天的?”，一个常规的短语就会被用来匹配用户表达日期的多种方式。

3.8.1　混合主导对话

根据参考文献［56］，Walker 和 Whitaker 指出更为自然的交流会显示出混合主导性。在人际对话中，说话人可能会提供所问问题之外的额外信息，或让听话人改变当前执行的任务。因此，能够在混合主导设置下运行的对话系统必须对提供含有限定条件之外的信息话语有所准备。餐厅系统可能会问“你想在哪里吃?”，此时若用户仅关注了时间，他就可能回答“我们想预订 7 点用餐”。该回答与餐厅预约系统内的其中一个问题相关，但却不是系统所期待的针对该问题的回答。这就要求一个 NLU 组件能够对更为复杂的输入进行解码和阐释，而不仅仅是简单直接回答的短句。

用户给予设定问题的直接回答方式当然有很多种。例如，用户可以就问题“你想在哪里吃?”给出各种各样的描述餐厅特点的回答，如下第一列所示：

用户话语	预设 – 填值对
泰国罗勒	名字：“泰国罗勒”
一家印度餐厅	菜式：“印度菜”
一家在圣弗朗西斯科的餐厅	地点：“圣弗朗西斯科”
我想去一家米其林星级的意大利餐厅吃饭，在圣弗朗西斯科，明晚 8 点	菜式：“意大利菜”
	评级：“米其林星级”
	地点：“圣弗朗西斯科”
	日期：“明天”
	时间：“晚上 8 点”

系统需要满足一个能够对应某个特定餐厅的高级预设，但尽管所提供的信息来自低预设 – 填值集合，这些信息仍可间接缩小可能的选项范围。要注意的是，应答可能按不同的顺序给出，分别对应不同的预设，表现出语言的自然变体。最后一个表达就是一个满足了多个预设的应答，且应用了相当复杂的自然语言描述。目标预设往往根据特定的领域进行设定；它们通常在后端数据库内对应列名。依照一组预设 – 填值对，应用逻辑可以从后端数据库中检索结果。

NLU 模块的任务是把第一列的话语映射到第二列的预设推论中。如果要高度准确地确

定含义，NLU 模块就必须处理语言表达和顺序的变化。NLU 的一个简单策略是按照填值满足预设的模板样式配对话语：

询问	格式模板
泰国罗勒	［名字］
一家中国餐厅	一家［菜式］餐厅
一家在圣弗朗西斯科的中国餐厅	一家［地点］［菜式］餐厅

在此简易方法中，每个短语都要求有自己的模板。更为复杂的格式配对会使用常用词组、去情景化语法，或以更清晰的语言形式体系来编写规则，使少数的规则能处理大多数的变体。但不管怎样，这些方法都需要解决配对中的语言模糊性问题。

模板或规则为可能出现的实体或关键词组提供了语境。命名实体识别（NER）任务经常是一个单独的处理步骤，它能挑选出可识别的意向实体（如例子中的餐厅名字和菜式）的子串。像参考文献［57］那样的机器学习途径通常用来进行命名实体检测。这些技术已经用于处理配对中的表达变化和语义歧义，但是它们需要大量的话语范例与正确的预设 - 填值对组合。组合好的话语随后被转化成 IOB 符号，其中每个单词都分到了以下其中一种标签：

标签类型	描述
I	满足预设（Inside a slot）
O	预设之外（Outside a slot）
B	预设开始（Begins a slot）

I 和 B 标签有与其相关的预设名。一个经过 IOB 标注的话语范例如下：

IOB 标签	O	B 菜式	O	O	B 地点	I 地点
话语	一家	意大利	餐厅	在	圣	弗朗西斯科

该 IOB 标注的话语包括了为训练机器学习算法的训练数据。此时的任务可以看作是一个序列分类的问题。序列分类的一个一般方法是单独预测序列中的各个标签。对每个单词来说，分类器需要把基于周边单词和之前标签的特征结合起来，以最佳估算出当前标签的概率。一个在概率框架内合并数据的可行方法是条件最大熵模型，如参考文献［58］所示：

$$p(a_i \mid b_i) = \frac{1}{Z(b_i)} \prod_{j=1\cdots k} \alpha_j^{f_j(a_i,\, b_i)}$$

式中，a_i 和 b_i 分别是 i 位置单词的标签和有效语境。$f_j(a_i, b_i)$ 标明了从有效语境中抽取的解码信息的特征，它们通常包含一些以前的标签、当前的单词和一些周边单词。α_j 是模型的参数，它们有效地衡量了估计概率过程中各个特征的重要性。随后某个搜索程序（如 Viterbi）会被用来寻找最大概率的标签序列。

针对每个可能满足预设的回答训练数据并不理想。而且，含有明显单词的特征并不会直接概括全部相同的单词。正因如此，机器学习方法常常使用外部字典。若一个单词或词组在字典中是已知值，模型就可以把该值作为一个特征。参考文献［57］和之前的参考文献［59］一起共同使用最大熵模型来合并语境特征以及来自外部资源的特征，如字典。总的来

说，根据同现关系统计（如参考文献［60］所示），单词能被自动分级，且基于这些分级单词的特征能够改进产生的模型的概括能力，如参考文献［61］所示。

最近的神经网络方法如参考文献［62］尝试利用自动生成的单词与连续向量空间的配对，假定相似的单词应该会“相近”。该方法内的特征就可以直接使用这些向量表达式的特定坐标了。

如参考文献［63］所述的条件随机场（CRF）是另一个序列分类模型，它能为整个标签序列生成一个单一概率，而不是每次一个标签。参考文献［64］是 CRF 应用于命名实体检测的一个例子。

3.8.2　预设和填值技术的局限

一个移动助手可以仅仅依靠 NER 算法找到可以满足行为模板的答案，就能成功地执行诸多任务。

填值作为一组对于后端数据库内元素的单独限制，系统往往把它们的连接词（如，“菜系：意大利”和“地点：圣弗朗西斯科”）作为一个附加在合理输入词条（“吉普赛人私房菜”“巴巴可”）上的限制而从后端抽取。若用户使用更灵活的话语或更概括的条件互动，则该基础的自然语言理解形式将无法胜任。

思考一下“一家有现场音乐表演的意大利餐厅”和“一家没有现场音乐表演的意大利餐厅”的区别。虽然都提到了相同的特征，但是由于介词的不同，它们描述的是完全不同的两类餐厅。NLU 必须要辨别出介词表达的不同关系，辨别出“没有”是一个对预设值的消极限制，而不是指特定的餐厅的集合。诸如“有”或“没有”等修饰语以及其他介、连词常常在传统信息获取或搜索系统中被视为无用词，但移动助手的 NLU 必须要格外注意这类单词。

自然语言也会通过话语中特殊单词的顺序来设定意义。“一家有卖好红酒的意大利餐厅”并不会与“一家有卖意大利红酒的好餐厅”混淆，虽然肯定有很多餐厅都能符合两种描述。这种情况下，NLU 必须把单词的顺序转换成特定类别的语法关系或相依性，并要考虑到英语的形容词通常在名词前修饰。这种关系在以下的相依性图示中会表现得更为明显。

图 3.13 表明了依存关系分析器的输出，这是 NLU 处理过程中作用于命名识别器结果的一个环节。依存关系分析器检测单词之间的意义关系，如该例子中的“意大利”就是“餐厅”的修饰语，“有着”对餐厅加以限制，最后“好”修饰的是“红酒”。

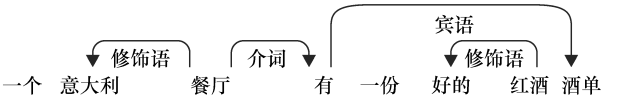


图 3.13　相依性图示 1

依存关系分析器也在所有从句中检测关系，查找一个事件和参与人以及他们的具体角色。图 3.14 显示的主语和宾语的标注限制了要搜索哈利被罗恩所救而不是相反关系的电影。编码了相依性的语法规则可以非常复杂，而且诸多方面重叠。这是在命名实体的语义模糊识别之外的另一个可能理解，如图 3.15 所示。

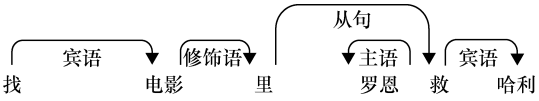


图 3.14　相依性图示 2

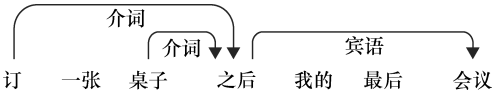


图 3.15　相依性图示 3

按照英语的语法规则，“之后”这一介词短语可以修饰“预订”或“餐位”。第一种情况的理解是要求当天晚些时候预订，在会议之后。第二种可能性更大的理解是现在应该就完成预订，以便晚些时候能有位置。依存关系分析器可能会更青睐某个语法规则，但最可能的意思是结合对话系统中其他可参考的信息，比如，在一个能考虑到特定事项模型的过往行为或一般餐厅预订规则常识的 AI 和论证模块中。

机器学习方法也已经为依存关系分析器[65,66]进行了定义。至于命名实体识别，由于受到标注有依存关系的大型语料库的驱使，该任务被划分成一个分类问题。有一项技术会考虑到话语内所有单词的可能依存关系并选取有最大化扩展项的概率树，即训练数据评估后分数最高的依存关系集合[67]。其他技术则从左向右逐渐处理句子，估计每个会最佳配对训练数据的行为点[68]。这些行为能够为下一个单词引入一个新的依存关系或将下一个单词暂存至一个栈而以后决定。

还有的语法分析器通过大规模手工编写语法生成依存结构或其相等结构[69-71]。它们根据语言的基本理论，通常得出含有更多语言学信息的表达式。而且，它们并不需要构建昂贵的已标注语料库，因此不受限于语料库的该特征。但是，它们可能会比数据统计的分析器消耗更多的计算资源，而且要求更多的语言学专业知识来开发和维护。这些因素都会决定哪种分析模块在特定移动设备配置中更有效。

依存结构生成了连接句子中各个单词的关键语法关系。但是想让系统理解单词的含义并转化成正确的系统执行还需要进一步的处理。许多单词含有 NUL 元件需要识别的多重或不相关的含义。鉴于移动设备能执行的任务的能力，通常只能执行一项内容。英文“book”这个动词本身就有多重含义（“预订”和“关押入狱”），但是对于预订服务设备来说唯一的可能就是第一项含义。区分英文单词“play”的含义则要多下点功夫：

- 谁打（played）塞雷娜·威廉姆斯？
- 谁扮演（played）詹姆斯邦德？

同样的单词在第一个问题中表达“打比赛”的意思，在第二个问题里是扮演的意思。意思的选择取决于宾语的类别。若宾语是一名运动员，则第一种含义成立；若宾语是一个影视角色，则第二种含义成立。去歧义处理取决于命名实体识别（查找命名）、指代消解（查找名称指代的对象）、语法分析（给予对象语法关系）。此外，去歧义还取决于本体推理：后端知识元件知晓塞雷娜·威廉姆斯（Serena Williams）是一名网球运动员，网球运动员属于运动员类别，该类别只和表达“打”含义的宾语匹配。

去歧义的推理不仅仅依靠查询类别的名称。限定和非限定描述的类别信息也同样需要用来确定含义，比如：

- 谁打（played）赢了法网公开赛？

- 谁弹的（played）史特拉第瓦里？

第一种情况需要了解哪些对象是体育赛事的参与者，即运动员；第二种情况需要知晓什么是史特拉第瓦里（Stradivarius），即一种提琴管弦乐器。随后该信息就能被传入一个从本体意义跨越到推理演算的模块，比如，提琴是发声物体，进而与“play”为“演奏”意义的宾语实现匹配。

这些例子适用于 RDFS[72]，一种与 RDF 连接的小型本体语言；RDF 是“资源描述框架”[73]，代表了在语义网（Semantic Web）中的实体对象的简单信息。RDFS 许可各类对象的表达式（塞雷娜·威廉姆斯指一个人，也指一名网球运动员）、此类别的概括关系（如网球运动员属于运动员）和归类到不同的逻辑关系中（“打赢”的主语是人）。

去歧义也会需要更复杂的推理链，包括组合多个对象或描述的信息。这些更为复杂的情况可能需要功能更强大的本体语言，如 W3C OWL Web 本体语言[74]。OWL 延展了 RDFS 在定义类别方面的能力（如定义一个人是男性），并提供局部归类（如一个人的孩子是人）。本体推理器是综合知识表达式和推理能力的具体案例，它们不仅能解决更为隐含的歧义，而且还可以为更为灵活的对话互动（见 3.10.6 节）进行策划和推理。这些需要有能够执行更为复杂的逻辑演绎任务的能力（比如一阶谓词逻辑），比相对简单的基于本体的推理耗费更高的运算成本。

用户输入的某些单词可能会根据它们的语境取义。指代会话中前述对象的代词和其他描述就属于这种情况。若系统指出一个满足所有用户要求的某个餐厅，用户可能进一步提问“那有不错的红酒吗？”随后系统必须识别（通过一个名为回指消解（anaphora resolution）的过程见——Mitkov[75]）句中的指示代词“那里”指的就是该餐厅。用户甚至可能会问“红酒品种怎么样？”，该问题没有包含明确的代词，但还是能理解为所指餐厅的红酒品种。限定描述（“红酒品种”）和指定餐厅的联系取决于本体意义指出的该餐厅的部分信息和属性。

有些单词和表达并没有涉及会话中前述的对象，而是直接指向在对话中的客体或发生的某些情况。指示代词（这，那，那些）和其他所谓的指示词(现在，昨天，这里，那里）就属于此类。如果对话发生在用户开车的时候，该用户可能会指向某个餐厅并问，“那家餐厅有好的红酒吗？”这种情况的对话系统必须识别用户正做出一个手势指示，辨明手势所指的是一家餐厅，并向 NLU 元件提供信息以便其将合适的对象信息赋予用户所指的“那家”餐厅。

当用户问到“这附近有没有什么好的餐厅？”或者“有什么在接下来一小时内会播放的电影吗？”，其他方面的对话情景（如当前位置和时间）必须同样考虑在内。这些例子说明了对话系统必须能够管理和对接来自不同渠道的多模态信息。能够处理这些多模态信息并使其同步化的是一个由 W3C 本体语言建议的名为“可扩展多模态注释标记语言”（EMMA）的工具[76]。

3.9　多轮对话管理

上述的 NLU 模块形式可以满足单轮对话系统的需求，即用户交流在单一话语结束后完成。但在多轮对话系统中，NLU 必须在问题、陈述和行为系统的场景以及前述话语中理解用户指示。这需要系统能够识别并追踪用户的整个对话意图。

把用户意图的空间分成对话意图和领域意图[77]是一个有用的方法。对话意图表明了子对话想要阐明、纠正或开启一个新的话题的开始，它是领域独立的。领域意图表明了用户想要通知系统或要求某个特定的系统行为。Young（1993）[77]的研究认为，两种类型的意图都需要建模并通过一个复杂的多轮对话过程追踪。

有一个意图追踪的方法叫对话状态追踪[78]。每个用户的话语首先由 NLU 模块处理以（通过分类）找到对话意图（告知、询问、纠正）和领域意图（播放、录制电影，预订餐位），并从话语中抽取预设 - 填值对。从当前话语中抽取的信息（包括模型不确定性的概率）反馈到一个动态模型（如一个动态贝叶斯网络）中作观察用。然后根据系统在当前话语前的信念状态，通过贝叶斯信念修正来移除或减少不确定性。

系统 您想在哪里吃？
用户 圣弗朗西斯科的一家意大利餐厅。
系统 我找到几家圣弗朗西斯科的意大利餐厅，它们是……
用户 其实我更想在今晚 7 点去一家中国餐厅。
系统 我找到在圣弗朗西斯科的几家中国餐厅，今晚 7 点它们都有餐位。它们是……

在这个例子中，为了正确理解用户最后的话语，对话状态追踪器区分了具有纠正意图的话语并覆盖了前述话语中提到的菜式种类。因此，该系统能够把用户最后的话语中提到的菜式、日期和时间预设与其最初提到的地点信息结合。这样的系统结构颇具吸引力，因为它能够处理语音识别产出/NLU 传递途径内在的不确定性和歧义。

尽管追踪话语意图对处理有声对话的自然流量是十分必要的，但是识别领域意图对系统理解用户的最终目的并采取措施也同样不可或缺。用户的领域意图往往很复杂，类似于一套以自上而下的方式组织的 AI 方案[79]。因此，包括从“与或”任务网络[79]到概率层级 HMM[80]的各层级结构都收到了根据复杂意图建模的指示。尽管稳定概率建模也会在预设和填值之间徘徊，但对复杂的意图进行稳定概率建模还是会要求更为清晰明确的、能结合概率和逻辑构建的表达式。这种复合建模方法是当前 AI 研究的活跃领域[81]。

根据对话状态，系统必须调整预期并找到一个合适的回答。像 RavenClaw[82]这样的对话管理器已经用来引导控制流量，使系统有足够的提示信息而得以完成任务。对话管理器在混合主导场景中必须使用 NLU 模型来检测任务在意外的对话时点发生变化。复杂的对话还要求一个错误矫正策略。

因此，对话中自然语言的理解需要与对话管理策略密切合作。正如前面所述，话语复杂性的范围可以从完全匹配已知数据列表的简单单词或短语，一直延伸到提供额外信息的开放性话语，或在任意时点命令转换任务的要求。准确的 NLU 模块综合使用训练数据和手动设定的语言材料来处理语言变体，包括字典、语法和本体意义。NLU 面临的其中一项挑战是恰当的理解话语、单词或短语的不完整信息。如果系统刚刚问到“您想什么时候出发?”，“早上 9 点”的回答就会被理解成在机票预订对话中填写的起飞时间，而预设“您想什么时候到达?”则针对的是对到达时间的提问。对话管理器把握着对话状态并能提供能够简化阐释话语碎片任务的对话语境信息。

早期提出的一个与 NLU 元件交流语境信息的简单建议是让对话管理器预测一系列的语言环境，从而能够帮助 NLU“理解”用户的下一组话语[83]。如果系统已经询问：“您想要什么时候离开?”那么对话管理器就能提供陈述式的前缀“我想在……时候离开”，以拼接用户阐述的任意前端信息。如果用户的回答是“早上 9 点”，则在连接之后的结果就是一个完整的、可阐释的和有意义的句子——按照正常的语法结构来说。在一个混合主导的场景中，用户并不局限于给予系统问题一个直接或最简的回答，因此话语管理器能够提供一系列可能的前缀并期待其能够涵盖用户的指令：

- ［我想］“早上 9 点”［离开］
- ［我想在］“周二早上 9 点”［离开］

该方法的主张是有一小组语式能够为自然的、有意义的用户回答提供环境；如果用户针对这个问题回答“波士顿”而不是一个时间或日期，这对机器乃至人来说都是十分诧异和费解的。当然，用户可能会选择根本不回答这个问题并提供关于旅行的其他信息，或设置转向另一个任务。那样的话，自然话语将是一个完整的句子，且对话管理器可以根据落空的语言环境做出预期：

- ［］“我想坐飞机去波士顿”

这是一种对话管理器和 NLU 元件共同合作的方法，用以决定用户下一话语段的含义。对话管理器能够根据对话的当前状态输入预期对象，通过一种能够简化整体系统的方式传送给 NLU，同时产生更为恰当的对话行为。

NLU 的输出模块能提供对话管理器需要的信息，以使其能够决定用户的意图和预期（比如，寻找附近的餐厅，看电影，订机票，或仅仅是想知道第一任美国总统的信息）。对话管理器还能考虑到系统的功能（比如，获取地方电视台节目，操纵 Netflix 上的视频或获取实时交通信息以及导航驾驶）、用户的行为和偏好，以及过往的交互体验。

如果用户的意图和预期得到满足，系统就仅会执行合理的领域活动。否则，其任务就是按照一个对话策略[84]，找出“接下来要说什么”，以便从用户处获得更多信息并最终满足用户的需求。一旦“说什么”的问题得以回答，自然语言生成（NLG）模块则将广泛应用并能够回答“该怎么说”的问题（即决定和用户交流的最佳方式）。

虽然对话管理器是如此根本的一个有声对话系统的元件，研究和运营单位对其的定义和功能还存在不同的理解。但是，人们一致认为对话管理器应该至少包含两个交际系统的基本

方面，即追踪对话状态和决定下一行为。

实施这两种功能的方式有许多。大多数商业系统和研究单位主要依赖于某些形式的有限状态机（FSM）[84]。该 FSM 方法要求对话中的每个变化都要被明确地表现为网络中两种状态的转换，并假定用户输入能够被系统提示局限或指挥。这意味着对话管理器并不灵活且无法处理突发的情况。让更为复杂的系统采用这个方法并不现实，因为它不得不完全明确在每个话轮的所有可能选项。而且，这种方法使得任何程度的混合主导变得几乎不可能实施。

上述缺点导致了“功能模型”方法[85-88]的问世。这其实是传统 FSM 的拓展。传统 FSM 允许有限状态机启用任一分类，旨在在每个状态实行主观决策，并对过渡数据假定任意复杂的先决条件。这些延展功能使系统能够接受过于具体的用户话语，这些话语以混合主导的形式存在。相对的，信息状态修改法[89,90,91]使用框架或树形结构作为控制机制，并为意外的用户话语留存空间。但是，任何这些系统处理的对话都通常是满足预设值的类型。系统仅会在指定任务的某个参数缺失的情况下询问用户问题。

为了处理更为复杂的任务，包括协作解决问题、智能助手和辅导对话，对话系统常常与规划技术一起实施[91,92]。最近，使用机器习得方法（更具体来说是强化学习（RL）法）的数据系统已经成为当前研究的重要技术。这些方法把对话策略建成一个顺序决策过程模型，称为“部分可观察马可夫决策过程”（POMDP）。Frampton 和 Lemon（2009）[93]综述了针对在有声对话系统中应用 RL 技术的科研进程。

这些方法为开发人员提供了精确严谨的数据导向优化模型，而不是依赖于专家和机构的策略。它们还有可能对隐蔽的状态进行归纳，对未知的情景进行调试，但由于需要大量的训练数据和稳定的技术来构建策略优化使用的状态空间、奖励功能和目标功能，这些方法也饱受诟病。还有，对于如何使该系统内的习得规律获得自然用户的本能理解并在需要的情况下加以修改，这一点的认知是广泛缺乏的。此外，解决 POMDP 问题的复杂性往往限制了对话系统表达式的丰富性。

最新研究开始关注上述问题，比如使用分层的 RL 来减少状态空间的大小[94]。另一项策略就是从一个小的数据集合来学习一个模拟的环境，使其能够使用 RL 技术而无需广泛的人与人对话数据[95]。

近几年，有声对话系统的商业应用已经获得了大批用户的青睐，这主要得益于其在移动设备上的安装运行。但是，这些系统仍然缺乏许多重要的功能。它们非常擅长执行某些基于满足预设条件的对话，但往往不是重返至一般网络搜索，就是搜索具体领域的某项服务（餐厅、电话联系人、电影）。

目前基本不存在能够扩展系统行为的合作筹划或解决问题的能力，也因此无法处理复杂或突发性的话语，因为有些话语需要通过与用户进行一系列灵活的、多话轮的互动才能进一步澄清理解，提炼内容，需要考虑到对话发生的情景和时代背景。话语研究人员正不断试验新的技术和系统，以期能够在更广阔的领域和场景支持更为自然有效的对话中介行为。随着整合机器学习、人工智能和推理、用户界面设计以及自然语言理解的新技术的不断开发，具备上述功能的系统将在不远的将来出现。

3.10　规划和推理

本节将探讨在未来对话系统开发中发挥越发重要作用的深层处理过程。这些包括深层语义分析，语篇表达式，语用学与知识的表达式和推理。下面首先以一个为例来指出技术上面临的一些挑战，随后简要概述相关领域的前沿发展。

3.10.1　技术挑战

考虑以下与未来自动虚拟助理（VA）的模拟对话：

（1）鲍勃：最后一次会议之后在“吉普赛人私房菜”预订一个餐位，通知汤姆和布莱恩到那里见我。

（2）VA：不好意思，那里直到晚上 9 点都没有位置了。你想让我帮你看看晚上 6 点半附近其他的意大利餐厅吗？

（3）鲍勃：你能找一下哪家餐厅有不错的红酒吗？

（4）VA：“巴巴可”有一个位置。它在金融区，但是去那里的路程时间差不多。

（5）鲍勃：好的，就那里吧。

话轮（1）在预约时间上的语义是模糊的：预订是在会议之后执行还是现在就执行？模糊消解问题需要系统建构常识：预订应该越快越好，否则餐位可能会被占满。自动的假设同样需要正确阐释这次交流：所述的“最后一次会议”是今天的最后一次会议而不是昨天的。

此处的推理是一次对一般交际规则的申诉[96]。人们总是尽最大的可能交流必要的信息；若一个人想预订明天的晚餐，他就应该会说。但是，这不过是又一个可以作废的假设，肯定不会每次都成立——因为该对话之前已经讨论过明天的计划了。同样，对介词短语“最后一次会议之后”的阐释必须要以同样的逻辑处理，因为明天或当天之后也能满足该条件。

当然，以上关于晚餐计划时间的推理只是估计；日程编排器需要更多确切的信息。系统当然也可以直接询问时间，但是一个真正有效的助理应该尽可能地努力“满足预设”；这里，系统应该尝试就晚餐的最佳时间创建一些合理的期待值。为了这个目的，VA 可以尝试根据过去的行为模式来进行推理；它可能知道鲍勃下午 5 点才下班，而且他通常走之前会花 30 分钟处理电子邮件。这些信息将存储在一个包含用户偏好和愿望的“用户一般行为模型”中，如后面所述。此外，系统应该根据任何可选地点的餐厅考虑行程时间。最后，汤姆和布莱恩的身份必须确认。同样的，该信息将能够存储在“用户朋友和联系电话模型”中。

值得强调的是，该话语表现的重要原则——对话系统在与用户交涉时必须能够考虑到各种可能的情景因素。这些因素不仅包括谈话的历史记录，如前面所述，还有许多对话内容之外的用户及对话发生的情景。正确的系统反应根据情景的不同而变化，如日期，用户所处地点，当前或预期的交通情况，用户最近听的音乐或看的电影。对话系统必须接收和阐释来自各类不同的传感器信号，并保持对话记录和过往的事件、行为等。

在话轮（2）中，我们看到对餐厅的初始搜索达到了明示要求但没有满足暗示预期。一项诸如“没有”或“我找不到你要的餐厅”的敷衍回答显然没有什么用。例子中提供了一个实用且有效的回答，简单解释了失败的原因，然后系统提出另一项建议。其他可能的建议可以通过放宽一些次要的限制而进行查找；在该例子中，餐厅类型和晚餐时间的条件被放宽了。该活动应由对话管理模块负责，指导系统和用户共同许可可执行的限制条件，同时尽可能获取用户模型来捕捉相对重要的条件。

在话轮（3）中，出现了文献所指的“间接言语行为”[97-99]，该概念将在 3.13.3 节中详细解释。如果按照字面理解，针对此句可以直接回答“是”或“不是”，但是都不尽人意。该句实则是一个执行行动的间接要求，该命令暗指预订餐位。就对话管理而言，应该注意到用户自己已经间接回答了话轮（2）中的问题。此处再次违反了表达简洁清晰的一般交际原则。既然用户没有不同意，也就意味着他给出了一项间接的确认和一个新的条件——是否有不错的红酒。这些限制必须要在某个时点加以集中并展开搜索。这就需要阐释当前话语时考虑到前述话语的相关限制。话轮（3）意味着该要求应该被阐释为寻找一家有卖好红酒的意大利餐厅。此时，许多数据库和网页可以执行该项搜索。

随着对话进展到第四轮，VA 告知用户它已经解除了其中一个先前的限制（“附近的餐厅”）而保留了其他条件，比如“相同的行程时间”“意大利餐厅”和“今晚”，这是通过启用与话轮（2）一样的处理过程实现的。然后鲍勃确认了话轮（5）中的提议从而使对话结束。VA 现在可以前往一个合适的网页来预约餐位并给汤姆和布莱恩发送邀请了。但是，得力助手的职责还没有完成。它必须有一贯且前摄的行为，始终监控突发事件（比如迟到）的发生并提供最大可能的帮助。

鉴于存在的这些技术挑战，下面将综述文献概述的最为常用的若干解决方案。

3.10.2　语义分析和语篇表达

多数设有虚拟助手的对话系统都需要执行一般性的行动。从语义和语篇层次的分析来看，有一种方法是将发生的事件具体化并将句子结构映射到一阶逻辑表达，表达式中的恒定标记代表了具有多种属性的（如“杀害”）特定事件（如肯尼迪遭枪击）及其与其他时间的联系。以话轮（3）的话语片段“你能帮我找一家餐厅”为例。以下公式能够表达这一转换：

$$\exists e1\,\exists e2\,\exists x(surface_{request(e1,e2)} \wedge agent(e1,Bob) \wedge agent(e2,find) \wedge object(e2,x) \wedge restaurant(x))$$

这可以说成是 $e1$ 代表了鲍勃是行为人的表层请求（见后面所述），且该请求是关于 $e2$——一个由虚拟助理（有时又称个人助理）执行的“查询”类事件。查询事件的对象是 x，也就是一家餐厅。该表达式的优势在于可以把一个事件的额外属性通过一个显化的方式串联在一起：比如，可以添加额外的条件，像餐厅是意大利的：意大利（x）。但是有一个问题，从逻辑的角度来说，增加这样一个条件必须重写整个公式，因为限制条件必须在存在量词的范围中出现。

语篇表述理论（DRT）已经提出了一个解决方案，即把语篇演变的动态模型保存在能

够扩增新信息的结构中。图 3.16 展现了一个完整话轮（3）的语篇表述结构（DRS）。图 3.16 的方框内列出了一组对应变量 *e*1、*e*2、*x* 和 *y* 的参照标记，接着是含有那些变量的一组条件。该结构能够随着新信息的出现而扩增，然后根据需要被转换成推理所需的一阶逻辑形式（如右侧所示）。

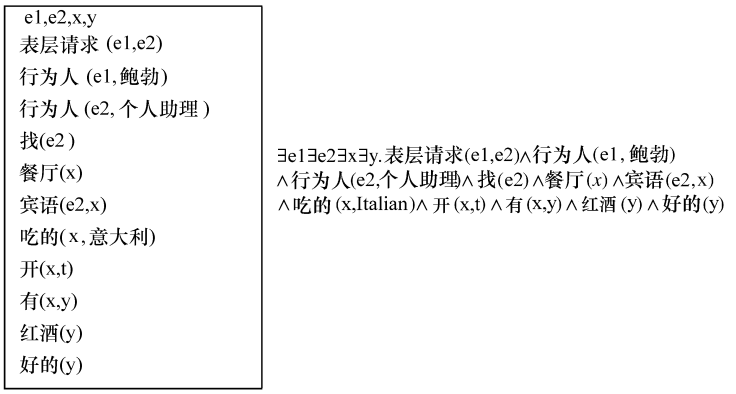


图 3.16　语篇表述结构和依据“你能帮我找一家有卖好红酒的意大利餐厅吗?”提问的一阶逻辑陈述

3.10.3　语用学

在分析我们的目标用户——虚拟助理交互的过程中出现了许多语用学问题。哲学家 Grice[96] 以精练的语言概括了说话人在会话过程中遵循的交际法则。这些法则描述了说话人提供的信息应该：

1. 真实；
2. 充分（但不冗余）；
3. 关联；
4. 清晰、间短并避免模糊隐晦。

如何用运算方式捕捉这些原则特征当然是一大挑战。这些法则表达的是最佳的默认行为，而行为人可以违反，也可以在交际过程中执行这些法则[100]。这些法则也反映了语言的效率性，即交际内容比话语内容要丰富得多。如在例子话轮（3）中，话语的理解需要结合语境才能正确地识别用户的要求是要找一家有卖好红酒的餐厅（在计划的时间和日期）。

言语行为理论的应用也是语用学的一个核心议题。关于此理论最佳的解释是，话语是能够以某种方式改变世界的行为（具体来说是其他行为人的信念和意图），而不是一个基于事态的真实值。以我们的虚拟助手为例，言语行为必须转化成 VA 所表现出的为用户服务的意图或承诺。意图往往不会显化，于是必须推测。上述例子就体现了一个需要推理的交际过程；可能读者并没有留意，用户在对话中没有在任何一处提及他或她想要在餐厅吃晚饭。这可能是一个不起眼的细节，但如果虚拟助手决定要在一家仅在白天提供饭菜或仅在晚上提供酒水的餐厅，这个选择就恐怕难以满足用户的需求了。

在话轮（3）的初始 DRS 中，话语被阐释成了一个查找餐厅的请求，但通过“意图识

别”过程，该请求被转化成了图 3.17 所示的结构。这体现了预订餐厅的请求以及话语实际的意图。计划识别过程的实施从查找行为开始反向进行，并推理：如果用户想找餐厅，可能是因为他想去餐厅，而他想去餐厅的一个可能前提是有预订的餐位。通过虚拟助手的帮助，最终系统收到了为用户预订餐厅的任务。

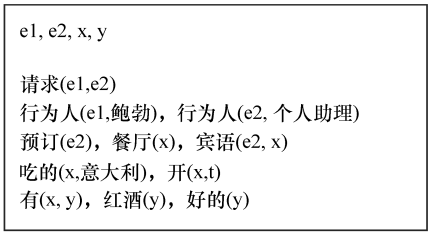


图 3.17　“你能帮我找一家有卖好红酒的意大利餐厅吗?”提问的意图结构

除了采用了逻辑法的计划识别[91,92]，目前已经开发的概率法也能清晰地处理关于系统理解用户大脑思维的不确定性问题[80]。

3.10.4　对话管理协作

目前已经提出的管理对话和提供帮助的方法层出不穷。大多数的方法都是基于观察到的对话常常会包含某些关联任务。那么，一个对话就是一次两个行为人之间的协作活动：行为人参与任务，平等互换信息，从而能使双方共同完成任务。主要的方法可以按照以下方式分类：

- 主从式：互动根据预设好的任务层级（又称为“任务列表”）被追踪和管理，有时存在明确的言语行为。
- 规划式：将协作建立成共同规划过程，该规划是一个复杂的结构，其中的任务列表不需要事先制定而是在运行时构建。
- 学习式：尝试习得对话互动原型。

我们将展开谈一下规划式方法，因为它为建构协作模型和支持虚拟助手提供了有效途径。

对话理解的规划式模型构建了一个行为人（无论是自然人、电脑系统或是一个团队）的信念、预期或意图。无须考究哲学内涵，此处只要把信念视为特定情景中捕捉到的用户持有的信息就足够了。一个行为人的想法可能出现偏差，但虚拟助手的责任恰是要检测用户可能持有的错误信息（如错误地认为某家餐厅在城市的另一侧）。在对话情景中，行为人的想法可能涉及最为明显的对象特点或正在讨论的对象（如我要去的“意大利餐厅”是“吉普赛人私房菜”）。

预期能够反映用户的偏好（比如，比起中餐用户更青睐意大利餐，或是用户更喜欢在可能的情况下走高速）。意图体现了行为人的责任。比如，系统可能负责保证用户按时到达预订的餐厅。因此它会一直监控用户在实现意图之前的进展状态。

虚拟助手的一项重要任务是依靠任务目录库内的信息，帮助用户分析用户已经交代给自己的高层级意图或规划。一旦规划得以读取，生成的一组潜在选项（如果选项多于一项）就能被分析，而其中回报（或效用，如让用户按时到达餐厅）较高的选项将会被选为行动。

3.10.5　规划和再规划

规划的过程会使用前面提到的任务列表，该表会为其他执行行动的方法编码。在我们的

模拟对话中，假定有三种预订餐厅的方式：去“订台网”预约，去餐厅主页预约，或直接打电话给餐厅。然后用户和系统将共同在那些目录中填写信息或以某种方式列举。每个行为人都有互补的能力和职责。接下来，对话将记录下各方为推进目标的实现而做出的贡献。

任务列表包括对各个组成节点的逻辑限制。它们可以是预设完毕并针对某个特定目标的深层结构，或是可以在策划过程中构建的浅层分解。图 3.18 展示举例中关于后者的几个例子。从这些元素中构建一个复杂结构能够增强系统的灵活性，以应对在规划过程中的突发事件。

图 3.18　目录结构

3.10.6　知识呈现与推理

系统内的各种知识，比如餐厅类型、行程时间、酒水和餐厅菜单等都存储在知识表达式中。一阶逻辑作为一个非常清晰的知识表达式，仅是诸多选项中的一种。为了方便处理，其他逻辑还包括基于本体意义描述的与“语义网”有关的各个逻辑。这些对区分单词含义有特别显著的效果，这在 3.8.2 节中已经探讨过。在某些情况下，时间关系、默认知识和限制条件可能会需要专业知识表达式，而这些可能需要专业的推理器。

纵观本节，我们一直在讨论用户偏好模型。偏好与行为人的预期直接相关。偏好可以由用户提供，由虚拟助理检测，或是通过观察获得。这些偏好可以通过定性法表达，或者通过按照功效的定量法表达。比如，我们可能有一个简单的用户模型陈述用户喜欢毕兹咖啡。但是，更好的陈述可能是含有比较关系的表达，比如比起星巴克更喜欢毕兹。这些又叫“其他情况不变”偏好，因为它们主要用来获取其他条件不变的情况下的一般性数据。但是例外的情况需要处理。比如，比起美国滴滤咖啡，有人可能更喜欢浓咖啡，从这里可能得出的结论是这个人更喜欢星巴克而不是有着故障咖啡机的毕兹，这就是例外条件。由于这些偏好均从用户身上获取，因此必须检查它们的一致性，因为用户可能突然提到相反或不一致的偏好。如果如此，系统应该与用户通过对话交谈以解决该问题。

3.10.7　监控

我们延续上述情景如下：

（1）［VA 注意到已经是下午 5 点半了，而鲍勃还没有离开办公室，这样他无法按时到达餐厅。MA 通过 TTS 打电话给鲍勃。］

（2）VA：鲍勃，你要迟到了。我应该变更预订时间吗？

（3）鲍勃 > 行。我大概 30 分钟以后出发。

（4）［VA 重新规划：巴巴可晚些就没有位置了。根据餐厅偏好，VA 在另一家类似的餐厅执行了预订。］

（5）VA：巴巴可无法保留你的餐位，所以我在卡帕尼纳帮你预订了一个餐位。它们的红酒不错。我会告知汤姆和布莱恩。这样行吗？

（6）鲍勃：可以，谢谢。

（7）［VA 发信息给汤姆和布莱恩，重新建立监控］

在之前的情景结束时，VA 实施了监控（有条件意图）来确保计划能够顺利完成，这包括餐厅的选择、预订餐位和出席晚餐。该意图是相对于 VA 的信念。有条件意图应该在每一个步骤都检查，观察其是否受到某些变化因素的影响。在这个例子中，如果 VA 逐渐认为鲍勃没有在预计时间结束会议，那么它就应该建立起需要重新规划预订活动的意图。

在接下来的情景中，上述猜测果然发生了：VA 注意到鲍勃还没有离开会议。它查看了餐厅并发现预约无法延长。因此，基于互助的合作关系，基于鲍勃对晚餐的要求和他的偏好，它开始寻找其他餐厅：餐厅的地址、菜肴的类型和红酒。它找到了一个替代对象：卡帕尼纳。虽然没有在同一个区域，但是行程时间相同。它于是舍弃了地点的限制条件，继续执行任务，预约了餐位并通知布莱恩和汤姆最新的变更。

3.10.8　推荐阅读文献

本节仅粗略综述了相关的研究领域。有兴趣的读者可以在以下参考文献中获取补充信息。

（1）Davis（1990）[101]很好地梳理了常识推理、一阶逻辑和诸如物化和其他事件表达式等技术。此外，一年两次举行的“常规推理逻辑范式”大会会刊将为读者提供该领域的最新发展[102]。

（2）Kamp 和 Reyle（1993）[103]以及 Gamut（1991）[104]介绍了 DRT 技术。一个相关的方法是分段语篇表达式理论（SDRT）[105]。语用学领域的介绍可以参考 Levinson（1983）[106]，包括对格莱斯会话法则的详细讨论[96]。言语行为理论详见参考文献［97，99］。规划识别主题的年度会议期刊对其他方法进行了很好的综述。

（3）关于知识表达式领域的介绍详见参考文献［107，101］；该领域内包括默认推理技术等更新的发展详见年度 KR 会议论文集[109]。参考文献［107］讨论了关于偏好表达式的效用理论。已经开发了一系列用于表示偏好的工具[110]。

（4）关于对话处理规划式方法的原创论文包括参考文献［111，112］。对话处理的信息式方法详见参考文献［113］。对话处理技术近期的发展详见参考文献［114］。

3.11　问题解答

在与上述虚拟助手互动的过程中注定会出现一些普遍问题。例如，用户可能会问：“这附近是否有素食餐馆？”，从而决定在何处订餐。回答这个问题只需要系统执行找到相关信息的任务行为，但是这类的问答（QA）互动是制定和优化任务的重要组成部分。

问题解答作为独立的任务有很长的历史，经历了自然语言界面到数据库的转变[115]，再到支持智能分析[116]，乃至最近 IBM 的 Watson/DeepQA 系统可以成功地在“危险边缘”问

答游戏中胜过人类冠军![117]但如上所述，问答在更广泛的口语对话应用中扮演一个自然的角色。例如，一个用户可能想要在预约吃饭时间之前问到电影开始的时间，或者系统可能自行决定它需要这个信息，以此来核对用户的时间选择是否和他的其他计划相冲突。成功的回答一个问题需要在关键步骤中涉及所有的自然语言内容（识别、句法分析、含义以及推理）：

- 问题分析：提问人想知道什么？
- 定位与问题相关信息。
- 确定答案以及答案的证据。
- 提交答案信息给提问人。

这些步骤必须要依照提问人的需要、偏好和生活来对意义和回答进行导向。“还有时间去看电影么？”这样的问题要求了解提问人的日程。

3.11.1　问题分析

问题分析可能不止与确认问题内容有关，而且会涉及问题提出的原因。“这儿附近是否有提供素食的餐馆？”技术上是一个是非问题，但应提供的有用信息应该是列出合适餐馆。“有用答案”取决于提问人的意图（如果他们只想要确认上一条信息，那么“是”就是合适回答，而无需对信息全部描述）。确认提问人的意图取决于对话内容、领域和对世界的知识（此情况下，需要了解提问人的地址）。

问题分析通常要确认问题中关键项和其之间的关系。关键项（实体）通常是名词，而关系则可以是主要谓语（提问者意图的主要信号）或对答案起到限定。例如，对于“这附近是否有素食餐馆？”这一问题，意图就是符合提供素食的限定，且距离提问者当前地址较近的餐馆列表。此处，“餐馆”“素食”和“这”是三个实体。

一般的简单问题可以通过制定句型（“＜事件＞的时间？”）来解决，但是这种办法无法满足各种变化较多和不太常见的语言句型。语法分析（见 3.8.2 节）是常见方法，但语言微妙处的构建很难准确分析，因此通常需要用数据实体和意图检测来补充（见 3.8.1 节）。已知实体的词典（例如电影明星名单，药品名，书籍标题，政治人物等）也可以有效定位各种常见实体，尤其对于特定领域。

3.11.2　寻找相关信息

一旦确定了意图和实体，我们便去找与意图相关的信息。例如，对于“这附近是否有素食餐馆？”，我们可以寻找餐厅的数据库，依照提问者位置寻找商务黄页，或进行一般互联网搜索。从这些结果中，我们可以编辑符合限定的（“菜单上有素餐”且“在当前位置附近”）餐馆列表。

有些信息仅仅存在于固定结构的形式中（表格或数据库，例如棒球队史上运动员信息），而其他种类的信息仅仅存在于非固定结构的形式中（自然语言文档，例如电影情节概要）。获取固定结构的信息需要精确的问题分析（提问者或许不了解数据库设计者如何创建字段名/域名，因此问题分析需要将问题的语义映射到数据库字段）。而通过搜索（在因特网或在特定源文件中）获取非固定形式信息要求的精度相对低，但会导致产生更多潜在回

答，因此更难选择正确的答案。

3.11.3　解答与依据

大多对问答的研究关注“事实陈述型”回答，这种回答简要地涵盖于某个文档的单一位置（例如秘鲁首都是哪儿？“利马是秘鲁首都”）。而对于非事实陈述型的回答研究较少[118]，尤其是回答所需依据来自多处的情况（例如，同一文档中不同部分的文字段落、不同文档中的文字段落、结构化数据和未加工文本的组合）。

支持事实陈述型和非事实陈述型回答的依据，可能因为用户问题语言不同而变化巨大。这个问题引起了近来对于掌握释义信息来驱动文本推理方面的研究兴趣[119]。非事实陈述型回答在针对内部复杂结构问题时（天为什么是蓝色的，但在早晨和傍晚又常是红色的?），或在针对简单问题提供高质量回答时（何时我可以在胭脂咖啡厅订桌：如果是两人桌，则在晚上 7 点钟，一张四人桌则在晚上 7 点半，四人以上桌则在晚上 8 点半，或者今天胭脂咖啡厅不营业）会出现。

3.11.4　呈现答案

集齐回答所需的依据后，系统必须找到一种方法将答案呈现给用户。这关系到如何策略性地确定呈现多少依据，以及技巧性地决定如何呈现出最好形式[120]。这些决定的做出依赖回答媒介（手机屏幕上的话语、文字）提供信息。然而，对于对话应用中植入的问答，背景目标（谈论中的问题）对于策略决定有巨大影响。在产生自然语言的策略层面，分散的文章篇章需要以连贯自然的方式拼接在一起。而结构性数据的自然语言回答必须从数据中获取或者通过文本搜索来确定能够呈现和支持问题的文本片段。

由于机器学习系统能够从对其行为的反馈中潜在得益和改进，因此捕捉用户对系统信息的反馈是十分必要的。但是，如果给出的反馈太过不自然和唐突（例如只是通过喜欢/不喜欢回答表示），那么机器学习就可能不会成功。相反，应该检测一些反应成功和失败更细致的线索作为反馈内容（例如，用户重复/重新表达问题，放弃任务，完成任务所需要的步骤数[121]）。

3.12　分布式语音交互架构

对于执行上节所述任务，用户对各类操作设备抱有越来越高的期待。尽管这些设备的显示形态因素和处理器能力各不相同，但都具有同样强大的局部计算和显示能力，并且都能够连接网络。用户对于跨设备（智能手机、平板电脑、超级本、汽车、可穿戴设备和电视机之间）的操作统一性和互动连续性的期待也有所增加。例如，用户可以从他们的智能手机、手表或眼镜上询问“凯尔特人队的比赛结果如何?”，并在其到家后命令智能电视“播放这场比赛”。

要达成这样的连续功能和互动模式，语音界面的框架需要能够跨设备、跨云端灵活操控。这使设备的计算能力、可用性（联网不成功的情况下）和延迟都得到了优化，并实现了用户个人喜好和互动历史的跨设备应用。为解决“播放这场比赛”这个指令，电视交互

界面将会获得云端用户档案和对话历史，使用期间引用这些信息，并在进程完成后上传到服务器，从而让用户在与其互动的下一个设备中也可使用这些信息。

3.12.1　分布式用户界面

如电话、电视机等设备通常可以作为其他移动设备的枢纽使用，这些移动设备运算能力较低，但可以给用户提供更有效更直观的界面，补充或代替枢纽设备的用户界面。例如，智能手机可以链接智能眼镜、手表，或者无线耳麦，从而把用户和设备之间传输信息的任务发布出去，使得这些外围设备用起来更高效、自然、顺手。

3.8.2 节中讨论过，多样化的模态可能会重叠和互补。假设用户输入导航目的地后，设备显示出一张标识出所有名为“春田”的小镇地理位置图。用户手指向正确的目的地，并确认“这个”。通过手势识别可以确认目标的位置，而使用语音命令来表达指令的性质以及测定手指指示的时刻。

这些设备生产商争相为自己的平台占据市场主导地位，导致用户每天都要分配自己的注意力来与不同的设备交互。这就增进了对“可移动”体验的需求，即追求在功能、互动模式、用户偏好以及与不同设备的互动历史方面的连续性。

一份通用型的设备用户档案将很快成为必不可少的信息。该存储的档案将包含用户本人的基本情况，例如喜欢的音乐风格或新闻类型，短期使用的诸如航线和酒店预订等相关信息，还有语音识别的相关信息，如声学模型。

设备使用声音生物计量来识别用户，获取基于服务器的用户档案，使用期间更新这些信息，并在进程完成后上传到服务器，从而让用户在与其互动的下一个设备中也可使用这些信息。

用户可能用智能手机搜索餐馆位置，这一行为就会被存储在他/她的用户档案中。此后，在一台有较大屏幕的台式电脑上，他/她可以选择一个地点。上了他/她的车后，这个位置信息就会被车载导航通过用户档案提取出来。同时，因为最近对话的语境可以通过用户档案调出，只需一句“开去餐馆”这样简单的指令就足够指引导航系统了。

在这个模型中，人机界面设计不再聚焦于任何特定机器的互动。机器的角色变成了第二位的，它成为了链接用户与数据和服务的通用多设备策略的许多实例之一。

这种抽象界面的常见构建就是“虚拟助手”，它可以链接用户与信息和服务。这样的助手可利用基于服务器的用户档案，其中集合了用户喜欢的所有设备上的相关资料。助手的具体形式可以是拟人的，以此展现出相同的 TTS 声音、说话方式和视觉外观，从而确保跨设备对话能够形成连续的个人化体验。用户不再关注与硬件的互动，而是直接或通过虚拟助手，转向与所需的信息进行对接。互动模式从使用“人－机”界面转向“人－服务”界面。

正如用户界面可以通过多种互动环节分布到多种互动设备上，不同的服务也可以通过多种设备和资源分布。同样的，硬件可以转移成为用户界面群功能和数据的后台，这个分布过程依照的是领域，而不是设备。

考虑如下元素：一款自动音乐播放器可以获取汽车硬盘自动点唱机、一张 SD 存储卡、一台相连的手机以及互联网服务中的音乐。用户体验设计师可以利用一个同类群内所有可用

的资源来设计体验内容。因此，一个指令“放点儿爵士乐”会生成从所有资源中挑出相符音乐这一行为结果。

3.12.2　分布的语音及语言技术

对于语音及语言用户界面，大致上要考虑以下因素来确定处理的位置：

- 平台能力：中央处理器，内存以及功率情况。
- 联网能力：网速，稳定性，带宽，联网额外花费，例如数据包限制。
- 语音识别和理解的应用领域所需要的模型类别和规模。例如，在不同语境下，是有10万个城市名需要识别，还是只有用户联系人列表中的几百个名字要识别？

以下设备类型是不同平台变化范围的一些例子：

- 个人电脑：充足的CPU和内存，持续供电。经常连接因特网。本地领域：命令运行软件和电脑，文本听写。
- 手机和平板电脑：有限CPU和内存，电池供电。经常连接因特网，联网可能更贵且不稳定（例如，信号覆盖消失）。
- 车载电脑：有限CPU和内存，持续供电。经常连接因特网，联网可能更贵且不稳定。
- 电视机：有限CPU和内存，持续供电。经常连接因特网，但并非所有用户都会将电视机联网。
- 云端服务器：广阔的CPU和内存资源，可同时应对多项互动。连接因特网以及其他大数据资源。

越来越多的联网促成了混合架构的发展。这些混合架构模糊了传统内置设定和基于服务器设定的界限，并且促成了对多种个人设备功能和领域的期待，例如，信息搜索、媒体播放、语音输入。

在考虑如何在分布式构架中分配任务时，备受推崇的做法曾经是“在数据本地进行处理”，而这种做法随着联网带宽的增长，已经并非绝对必要，但仍旧是良好的指导方针。假设自然语言或对话部分完全在远端服务器运行，则用户界面的一致性也是重要方面。如果数据连接中断，用户可能容易理解数据连接服务就像网络搜索一样被中断，但是他们可能不清楚，设备的自然语言对话能力也随之不可用了。

在语音与语言用户界面中，植入的“自带”语音识别通常通过处理语音命令来操作指定设备，这可以通过使用语法分析型命令及控制类别识别器，或小数据语言模型（SLM）来达成自然语言处理。然而，当前的移动平台在试图识别装载数万城市名的大预设列表的SLM语音时达到了极限。这样，该任务只能通过基于服务器的识别器来完成。很多情况下，在植入平台和服务器上同时进行识别是一种好办法，通过比较结果的置信度，然后选择最优解，从而避免低置信度自带语音识别结果对服务器语音识别的触发而引起的延迟。

自带识别器上的其他任务还有唤醒词语检测，并结合声音生物计量来分别启动设备并验证用户，利用语音行为和终点检测来分割语音和进行语音识别特征提取，从而保证只需往识别服务器上传输语音特征而非整段语音。

用户档案对于存储对话决策相关的个人偏好、说话人特征、本地语言声学和语言建模都有益处。用户档案还存有生物计量信息，可以确认用户的身份，从而授权某种服务或资料获取。若用户档案可以在任意设备上获取，则作用最大。但即使设备中断网络连接，比如当车过隧道时，用户档案也应该继续发挥作用。这个问题可以通过云端主人用户档案配合本地设备的同步复制档案解决，或者通过把手机作为档案的中心枢纽，因为手机是陪伴用户时间最久的设备。

在服务器上存储这样的档案的另一个优点是这一系列的档案可以组成一个包含广泛信息的独立实体，并且允许从用户群体或部分群体中获取数据。有的新闻服务可能有兴趣从所有连接到档案群的记录中找到热门话题，然后进行关键词搜索。有的音乐网店可能会查询档案群寻找加利福尼亚州 18 ~ 25 岁男性最喜欢的歌手。

通常，不同用户的各个档案相互连接，例如，用户 A 和 B 互为彼此电子邮件通讯录中联系人，或通过社交网站有联系。如果这个信息存储在用户档案中，这一组跨区相连的档案群就可以允许用户的虚拟助手进行提问，例如，“我现在要去的城镇有没有我的朋友，或者有没有朋友的朋友在那里?”或者“我的朋友们都在听什么音乐?”基于服务器的识别和日志，当在用户档案中存储数据时，隐私和数据安全是设计和操作服务器基础设施的关键。

最后，在输出方面，TTS（文本转语音）和语言生成通常在用户设备上运行，除非高品质声音所需的内存比本地内存大，或者整个应用程序受托管且服务器解决方案更方便建立和维护。

3.13　结语

语音驱动的 NLU 交互界面涵盖了广泛的设备，包括了手机、平板电脑、电视机、汽车和信息咨询台。用户与它们的交互已经成为了每日例行的活动。这些界面使安装在设备上的复杂功能变得更简单自然。相对于发出一系列细微的命令，用户可以以越来越自然的语言表达他们的综合意图，而由系统决定需要执行的步骤。这种自然语言的互动在许多环境中正在变得愈发实用：街道上、汽车内、客厅里以及新的装置上。

所有这些新的功能都指向一个问题：如何能最佳的把自然语言理解植入今天的视觉界面呢？有一系列多样的途径，包括“虚拟助手”这类选项。2013 年是该技术的丰收之年，我们见证了诸如苹果 Siri、三星 S - Voice、声龙 Assistant 和谷歌 Now 的诞生，以及市场上将近 60 种的类似产品。

虚拟助手可以被视为是单独的个体，它可以进行对话，还有自己的个性。它可以阐释用户的输入，并协调用户、设备本地用户界面和一系列（非）自带识别应用。某些情况下，助手甚至可以用自己的 UI 对获得的信息进行重新整理，从而承担起了放大和过滤网络信息的任务。

另外一种设计也许可以称为“环境 NLU”，它保持了本地设备的外观和应用界面的使用体验，但嵌入了语境敏感的 NLU。通过与该界面对话，用户可以获取信息，也可以打开并控制熟悉的应用。系统在需要完成多轮讨论或消除歧义时与用户对话。相对于在事件中占据主导，这类助手的特点是低调、高效和灵活。可能的情况下，它会基于一个单一话语指令完成任务而不会限制用户可以获取信息的渠道。它的目标并不是帮助用来解决现有 UI 的短板，而是致力于成为一个改进 UI 的内在组成部分。

不管哪种，语音和语言理解现已被视为一个新的基础元件——能获取和控制位于设备或云服

务的无形资产，为传统视觉 UI 增加了一个新维度。往后若干年，随着工程师不断地对体验结构进行更新再造，我们必将目睹这些新维度的积极扩展和对当前“缩小桌面”现象的快速修改。

语音革命的进程因为多种组件技术的持续发展而不断推进，近几年在许多系统只是“简单运转”的领域进行着持续的推广和改进。性能上的不断突破主要归因于许多互补领域的改进，包括：

- 语音识别技术，特别是 DNN。
- 信号获取增强。
- 改进的 TTS 和声音生物计量建模。
- 结合结构性方法的意义抽取和机器学习。
- 对话互动、概率规划识别、知识呈现和推理。
- 问题解答。

许多因素促成了这些进步：

- 可用运算能力的提高，包括特殊目的的计算设备。
- 可用训练语料库的规模。
- 数据统计建模的改进。
- 数以千计人的多年的研发努力。

除了这些进步之外，我们仍旧面临着许多挑战。或者积极地说，我们期待在未来数年取得进一步的发展。构建有着深层理解人类语言能力的对话代理人既是挑战也是我们的承诺。

致谢

作者想要感谢对本章做出贡献的参与者：Dario Albesano，Markus Buck，Joev Dubach，Nils Lenke，Franco Mana，Paul Vozila，Puming Zhan。

参 考 文 献

1. Davis, S., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**(4), 357–366.
2. Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America* **87**, 1738.
3. Bahl, L., Bakis, R., Bellegarda, J., Brown, P., Burshtein, D., Das, S., De Souza, P., Gopalakrishnan, P., Jelinek, F., Kanevsky, D. (1989). Large vocabulary natural language continuous speech recognition. *International Conference on Acoustics, Speech, and Signal Processing, 1989* (*ICASSP-89*).
4. Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2), 257–286.
5. Ney, H., Ortmanns, S. (2000). Progress in dynamic programming search for LVCSR. *Proceedings of the IEEE* **88**(8), 1224–1240.
6. Hunt, A., McGlashan, S. (2004). *Speech recognition grammar specification version 1.0*. W3C Recommendation. http://www.w3.org/TR/speech-grammar/.
7. Chomsky, N. (2002). *Syntactic structures*. Mouton de Gruyter.
8. Jelinek, F. (1997). *Statistical methods for speech recognition*.: MIT press.
9. Bahl, L., Brown, P., De Souza, P., Mercer, R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '86*.

10. McDermott, E., Hazen, T.J., Le Roux, J., Nakamura, A., Katagiri, S. (2007). Discriminative training for large-vocabulary speech recognition using minimum classification error. *IEEE Transactions on Audio, Speech, and Language Processing* **15**(1), 203–223.
11. Povey, D., Woodland, P.C. (2002). *Minimum Phone Error and I-Smoothing for Improved Discrimative Training*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP).
12. Leggetter, C.J., Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language* **9**(2), 171–185.
13. Hermansky, H., Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* **2**(4), 578–589.
14. Furui, S. (1986). Speaker-independent isolated word recognition based on emphasized spectral dynamics. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'86*.
15. Kumar, N., Andreou, A.G. (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication* **26**(4), 283–297.
16. Sim, K., Gales, M. (2004). Basis superposition precision matrix modelling for large vocabulary continuous speech recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP'04)*.
17. Lee, L., Rose, R.A. (1998). Frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing* **6**(1), 49–60.
18. Kneser, R., Ney, H. (1995). Improved backing-off for M-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95*.
19. Chen, S.F. (2009). Performance prediction for exponential language models. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
20. Kuo, H.-K., Arisoy, E., Emami, A., Vozila, P. (2012). *Large Scale Hierarchical Neural Network Language Models. INTERSPEECH*.
21. Pereira, F.C., Riley, M.D. (1997). 15 Speech Recognition by Composition of Weighted Finite Automata. *Finite-state language processing*, 431.
22. Pogue, D. (2010). TechnoFiles: Talk to the machine. *Scientific American Magazine* **303**(6), 40–40.
23. Hershey, J.R., Rennie, S.J., Olsen, P.A., Kristjansson, T.T. (2010). Super-human multi-talker speech recognition: A graphical modeling approach. *Computer Speech & Language* **24**(1), 45–66.
24. Bourlard, H.A., Morgan, N. (1994). *Connectionist speech recognition: a hybrid approach*. Vol. 247. Springer.
25. Gemello, R., Albesano, D., Mana, F. (1997). *Continuous speech recognition with neural networks and stationary-transitional acoustic units*. International Conference on Neural Networks.
26. Mohamed, A., Dahl, G.E., Hinton, G. (2012). Acoustic Modeling Using Deep Belief Networks. *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1), 14–22.
27. Hinton, G., Li, D., Dong, Y., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *Signal Processing Magazine* **29**(6), 82–97.
28. Dahl, G.E., Dong, Y., Li, D., Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **20**(1), 30–42.
29. Loizou, P.C. (2013). *Speech enhancement: theory and practice*. CRC press.
30. Hofmann, C., Wolff, T., Buck, M., Haulick, T., Kellermann, W.A. (2012). Morphological Approach to Single-Channel Wind-Noise Suppression. *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC 2012)*.
31. Widrow, B., Stearns, S.D. (1985). *Adaptive signal processing*. Vol. 15. IET.
32. Breining, C., Dreiscitel, P., Hansler, E., Mader, A., Nitsch, B., Puder, H., Schertler, T., Schmidt, G., Tilp, J. (1999). Acoustic echo control. An application of very-high-order adaptive filters. *Signal Processing Magazine* **16**(4), 42–69.
33. Haykin, S.S. (2005). *Adaptive Filter Theory, 4/e*. Pearson Education India.
34. Griffiths, L.J., Jim, C.W. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation* **30**(1), 27–34.
35. Wolf, T., Buck, M. (2010). A generalized view on microphone array postfilters. *Proc. International Workshop on Acoustic Signal Enhancement, Tel Aviv, Israel*.
36. DiBiase, J.H., Silverman, H.F., Brandstein, M.S. (2001). Robust localization in reverberant rooms. In: *Microphone Arrays*. Springer. 157–180.

37. Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing* **29**(2), 254–272.
38. Furui, S. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Transactions on Acoustics, Speech and Signal Processing* **29**(3), 342–350.
39. Reynolds, D.A., Quatieri, T.F., Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing* **10**(1), 19–41.
40. Solomonoff, A., Campbell, W.M., Boardman, I. (2005). Advances In Channel Compensation For SVM Speaker Recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP '05)*.
41. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing* **15**(4), 1435–1447.
42. Dehak, N., Kenny, P.J., Dehak, R., Ouellet, P., Dumouchel, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* **19**(4), 788–798.
43. Mistretta, W., Farrell, K. (1998). Model adaptation methods for speaker verification. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*.
44. Speaker Recognition Evaluation (2013). http://www.itl.nist.gov/iad/mig/tests/spk/.
45. Evans, N., Yamagishi, J., Kinnunen, T. (2013). Spoofing and Countermeasures for Speaker Verification: a Need for Standard Corpora, Protocols and Metrics. *SLTC Newsletter*.
46. Klatt, D.H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America* **82**(3), 737–793.
47. Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press.
48. Ladd, D.R. (2008). *Intonational phonology*. Cambridge University Press.
49. Ladefoged, P., Johnstone, K. (2011). *A course in phonetics*. CengageBrain.com.
50. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1999). *Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis*.
51. Hunt, A.J., Black, A.W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96*.
52. Donovan, R.E. (1996). *Trainable speech synthesis*, PhD Thesis, University of Cambridge.
53. Pollet, V., Breen, A. (2008). Synthesis by generation and concatenation of multiform segments. *INTERSPEECH*.
54. Chen, L., Gales, M.J., Wan, V., Latorre, J., Akamine, M. (2012). Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training. *INTERSPEECH*.
55. Zen, H., Senior, A., Schuster, M. (2013). *Statistical parametric speech synthesis using deep neural networks. International Conference on Acoustics, Speech, and Signal Processing, ICASSP-13*. Vancouver.
56. Walker, M., Whittaker, S. (1990). Mixed initiative in dialogue: An investigation into discourse segmentation. *Proceedings of the 28th annual meeting on Association for Computational Linguistics*.
57. Florian, R., Hassan, H., Ittycheriah, A., Jing, H., Kambhatla, N., Luo, X., Nicolov, N., Roukos, S., Zhang, T. (2004). A Statistical Model for Multilingual Entity Detection and Tracking. *HLT-NAACL*.
58. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics* **22**(1), 39–71.
59. Borthwick, A., Sterling, J., Agichtein, E., Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *Proc. of the Sixth Workshop on Very Large Corpora*.
60. Brown F, deSouza V, Mercer RL, Pietra VJD, Lai JC. (1992). Class-based n-gram models of natural language. *Computational Linguistics* **18**, 467–479.
61. Miller, S., Guinness, J., Zamanian, A. (2004). Name tagging with word clusters and discriminative training. *HLT-NAACL* 337–342.
62. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* **12**, 2493–2537.
63. Lafferty J, McCallum A, Pereira FC. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*.
64. Finkel, J.R., Grenager, T., Manning, C. (2005). *Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.
65. Ballesteros, M., Nivre, J. (2013). Going to the roots of dependency parsing. *Computational Linguistics* **39**(1), 5–13.
66. Kübler, S., McDonald, R., Nivre, J. (2009). *Dependency parsing*. Morgan & Claypool Publishers.

67. McDonald, R., Pereira, F., Ribarov, K., Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.
68. Nivre, J. (2008). Algorithms for deterministic incremental dependency parsing. *Computational Linguistics* **34**(4), 513–553.
69. Riezler, S., King, T.H., Kaplan, R.M., Crouch, R., Maxwell III,, J.T., Johnson, M. (2002). Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.
70. Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering* **6**(1), 15–28.
71. Callmeier, U. (2002). *Preprocessing and encoding techniques in PET. Collaborative language engineering. A case study in efficient grammar-based processing*. Stanford, CA: CSLI Publications.
72. Brinkley, D., Guha, R. (2004). *RDF vocabulary description language 1.0: RDF schema*. W3C Recommendation. Available at http://www. w3. org/TR/PR-rdf-schema.
73. Manola, F., Miller, E., McBride, B. (2004). *RDF primer*. W3C recommendation; 10, 1–107.
74. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (2009). *OWL 2 Web Ontology Language primer*. W3C recommendation **27**, 1–123. http://www.w3.org/TR/owl2-primer/
75. Mitkov, R. (2002). *Anaphora resolution*. Vol. 134. Longman, London.
76. Johnson, M. (2009). *EMMA: Extensible MultiModal Annotation markup language*. http://www.w3.org /TR/emma/.
77. Young, S.R. (1993). *Dialog Structure and Plan Recognition in Spontaneous Spoken Dialog*. DTIC Document.
78. Williams, J.D. (2013). *The Dialog State Tracking Challenge*. SIGdial 2013. http://www.sigdial.org/workshops /sigdial2013/proceedings/index.html.
79. Ferguson, G., Allen, J.F. (1993). Generic plan recognition for dialogue systems. *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics.
80. Bui, H.H. (2003). A general model for online probabilistic plan recognition. *IJCAI*.
81. Domingos, P., Lowd, D. (2009). Markov logic: An interface layer for artificial intelligence. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **3**(1), 1–155.
82. Bohus, D., Rudnicky, A.I. (2009). The RavenClaw dialog management framework: Architecture and systems. *Computer Speech & Language* **23**(3), 332–361.
83. Bobrow, D.G., Kaplan, R.M., Kay, M., Norman, D.A., Thompson, H., Winograd, T. (1977). GUS, a frame-driven dialog system. *Artificial intelligence* **8**(2), 155–173.
84. Pieraccini, R., Huerta, J. (2005). Where do we go from here? Research and commercial spoken dialog systems. *6th SIGdial Workshop on Discourse and Dialogue*.
85. Pieraccini, R., Levin, E., Eckert, W. (1997). AMICA: the AT&T mixed initiative conversational architecture. *Eurospeech*.
86. Pieraccini, R., Caskey, S., Dayanidhi, K., Carpenter, B., Phillips, M. (2001). ETUDE, a recursive dialog manager with embedded user interface patterns. *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001 (ASRU'01)*.
87. Carpenter, B., Caskey, S., Dayanidhi, K., Drouin, C., Pieraccini, R. (2002). *A Portable, Server-Side Dialog Framework for VoiceXML*. Proc. Of ICSLP 2002. Denver, CO.
88. Seneff, S., Polifroni, J. (2000). Dialogue management in the Mercury flight reservation system. *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems – Volume 3*. Association for Computational Linguistics.
89. Larsson, S., Traum, D.R. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering* **6**(3–4), 323–340.
90. Lemon, O., Bracy, A., Gruenstein, A., Peters, S. (2001). The WITAS multi-modal dialogue system I. *INTERSPEECH*.
91. Rich, C., Sidner, C.L. (1998). COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction* **8**(3–4), 315–350.
92. Blaylock, N., Allen, J. (2005). A collaborative problem-solving model of dialogue. *6th SIGdial Workshop on Discourse and Dialogue*.
93. Frampton, M., Lemon, O. (2009). Recent research advances in Reinforcement Learning in Spoken Dialogue Systems. *Knowledge Eng. Review* **24**(4), 375–408.

94. Lemon, O., Liu, X., Shapiro, D., Tollander, C. (2006). Hierarchical Reinforcement Learning of Dialogue Policies in a development environment for dialogue systems: REALL-DUDE. *BRANDIAL'06, Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*.
95. Rieser, V., Lemon, O. (2011). *Reinforcement learning for adaptive dialogue systems*. Springer.
96. Grice, H.P. (1975). Logic and conversation. *Syntax and Semantics*, Vol. 3: Speech Acts, 41–58.
97. Searle, J.R. (1969). *Speech acts: An essay in the philosophy of language*. Vol. 626. Cambridge University Press.
98. Austin, J. (1962). *How to do things with words (William James Lectures)*. Oxford University Press.
99. Cohen, P.R., Perrault, C.R. (1979). Elements of a plan-based theory of speech acts. *Cognitive science* **3**(3), 177–212.
100. Lenke, N. (1993). Regelverletzungen zu kommunikativen Zwecken. *KODIKAS,/ CODE* **16**, 71–82.
101. Davis, E. (1990). *Representations of commonsense knowledge*. Morgan Kaufmann Publishers Inc.
102. Commonsense Reasoning (2013). *Commonsense Reasoning ~ Home*; http://www.commonsensereasoning.org/.
103. Kamp, H., Reyle, U. (1993). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*. Springer.
104. Gamut, L. (1991). *Logic, Language and Meaning, volume II, Intentional Logic and Logical Grammar*. University of Chicago Press, Chicago, IL.
105. Lascarides, A., Asher, N. (2007). Segmented discourse representation theory: Dynamic semantics with discourse structure. *Computing meaning*, 87–124. Springer.
106. Levinson, S.C. (1983). *Pragmatics (Cambridge textbooks in linguistics)*.
107. Russell, S.J., Norvig, P., Canny, J.F., Malik, J.M., Edwards, D.D. (1995). *Artificial intelligence: a modern approach*. Vol. 74. Prentice Hall, Englewood Cliffs.
108. KR, Inc. (2013). *Principles of Knowledge Representation and Reasoning*. http://www.kr.org/.
109. Baader, F. (2003). *The description logic handbook: theory, implementation, and applications*. Cambridge university press.
110. Boutilier, C., Brafman, R.I., Domshlak, C., Hoos, H.H., Poole, D. (2004). CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res.(JAIR)* **21**, 135–191.
111. Grosz, B.J., Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics* **12**(3), 175–204.
112. Allen, J. (1987). *Natural language understanding*. Vol. 2. Benjamin/Cummings Menlo Park, CA.
113. Traum, D.R., Larsson, S. (2003). The information state approach to dialogue management. *Current and new directions in discourse and dialogue*, 325–353. Springer.
114. SIGdial: *Special Interest Group on Discourse and Dialog* (2013). http://www.sigdial.
115. Woods, W.A., Kaplan, R.M., Nash-Webber, B., Center, M.S. (1972). *The lunar sciences natural language information system: Final report*. Bolt Beranek and Newman.
116. AQUAINT (2013). *Advanced Question Answering for Intelligence*. http://www-nlpir.nist.gov/projects/aquaint/
117. Ferrucci, D.A. (2012). Introduction to This is Watson. *IBM Journal of Research and Development* **56**(3.4), 1:1–1:15.
118. Surdeanu, M., Ciaramita, M., Zaragoza, H. (2011). Learning to rank answers to non-factoid questions from web collections. *Computational Linguistics* **37**(2), 351–383.
119. De Marneffe, M.-C., Rafferty, A.N., Manning, C.D. (2008). Finding Contradictions in Text. *ACL*.
120. Demberg, V., Winterboer, A., Moore, J.D. (2011). A strategy for information presentation in spoken dialog systems. *Computational Linguistics* **37**(3), 489–539.
121. Diekema, A.R., Yilmazel, O., Liddy, E.D. (2004). Evaluation of restricted domain question-answering systems. *Proceedings of the ACL2004 Workshop on Question Answering in Restricted Domain*.

Further reading

Allen, J., Kautz, H., Pelavin, R., Tenenberg, J. (1991). *Reasoning about plans*. Morgan Kaufmann San Mateo, CA.

Allen, J.F. (2003). *Natural language processing*.

Chen, C.H. (1976). *Pattern Recognition and Artificial Intelligence: Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence, Held at Hyannis, Massachusetts, June 1–3, 1976*. Acad. Press.

Dayanidhi, B.C.S.C.K., Pieraccini, C.D.R. (2002). *A portable, server-side dialog framework for VoiceXML*.

Graham, S., McKeown, D., Kiuhara, S., Harris, K.R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology* **104**, 879–896.

Kautz, H.A. (1991). A formal theory of plan recognition and its implementation. In *Reasoning about plans*. Morgan Kaufmann Publishers Inc.

Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Perrault, C.R., Allen, J.F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics* **6**(3–4), 167–182.

Roche, E., Schabes, Y. (1997). *Finite-state language processing*. The MIT Press.

Schlegel, K., Grandjean, D., i Scherer, K.R. (2012). Emotion recognition: Unidimensional ability or a set of modality- and emotion-specific skills. *Personality and Individual Differences* **53**(1), 16–21.

第4章

视觉传感与肢体动作交互技术

Achintya K. Bhowmik
美国英特尔集团

4.1 引言

视觉在我们与现实世界的交互中占据主导地位。虽然我们人类拥有的其他知觉感知与处理能力，例如触觉、语言、听觉、嗅觉和味觉等，也是使我们能够在日常生活中了解周围世界并做出反应的重要组成部分，但其中最为重要、最常用到的感知处理能力，是利用人类视觉系统接收并处理光学信息的能力，它让我们感知并了解了周围的世界。

图像显示器已经是绝大多数电子设备中的首要人机交互设备。我们日常用来计算、沟通、娱乐的电子设备，都是通过图像显示器以视觉信息的形式将系统输出与呈现给用户。显示器上视觉内容的控制与交互仍然是当前研发的热门领域，通常可以参考以“人机交互”或“人机界面”为题出版的文献。

正如第 1 章中所述，在通过显示器和系统实现人类交互的方式中，早期获得商业成功的实现方式，大多是利用远程电视显示和电脑鼠标操作这类非直接操作。随着近年配备触摸屏的显示器越来越多，同时为触控操作优化的软件应用程序及用户界面也越来越为人广泛接受，显示器正迅速成为能够接收直接人类操作的双向交互设备。不过，由于基于触控操作的系统本质上属于二维输入设备，它会将人类与显示器上内容的交互限制在设备平面来进行。而我们人类拥有一套包含双目成像与推理方案的 3D 视觉传感系统，能够在 3D 世界中通过视觉进行感知和交互。如果能够用上这套先进的视觉处理装置，那么一定能够显著扩展交互显示与系统的功能范围。这类显示屏和系统配备有拟人视觉感知与推理技术，能够“看到”并“感知”视觉显示屏前方 3D 空间内的人类动作，使得人类交互体验更为生动、自然、直观、拟真。

图 4.1 显示了交互式显示器的原理框图，为突出显示基于视觉的人类界面与交互，该图经过少许修改。此流程从通过实时图像获取捕捉用户动作开始。图像子系统将由场景产生的

光线变化转化为代表2D或3D视觉信息的电子信号，然后发送给计算子系统。专门设计用于从图像序列中抽取含义的软件算法便能识别用户动作，例如手部姿势、面部表情或双目凝视。此后这一信息会作为用户输入提供给应用层，后者会据此在计算硬件上执行各类处理函数以生成系统回复。最后，显示子系统以光线的形式产生视觉输出，形成能够通过用户眼睛和视觉传感系统感知的图像。

在本章中，我们主要说明基于视觉交互的技术和应用的基本原理和最新进展，重点讲述视觉传感和处理技术，同时还会涉及实现系统智能以启用自动推理与识别用户行为的算法方法。在下一节中，我们会讨论图像获取方法，涵盖2D和3D两方面的图像技术。在之后的三章中会详细说明3D传感技术。了解图像获取技术的概述之后，我们会介绍肢体动作交互技术，包括在交互式显示器上进行应用的姿势建模、分析和识别方法。最后，我们还提供了自动识别面部表情技术的最新进展的综述。

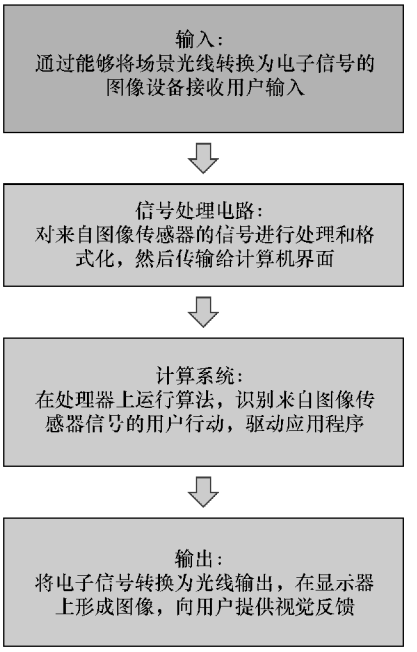


图4.1　交互式显示器系统架构的原理框图，主要突出基于视觉的传感和交互

4.2　图像技术：2D和3D

以能够捕捉2D图像的数码摄像头为代表的图像传感设备如今几乎随处可见，越来越多的设备无论尺寸大小都会连同图像显示器一起配备此类传感设备。摄像头如今是大多数移动设备的集成元素之一，包括手机、平板电脑和笔记本电脑等，同时，也有越来越多的一体化桌面电脑与新式平板电视等设备开始配备摄像头。虽然越来越多的此类系统开始广泛采用图像设备，但它们的应用大多数仅限捕捉数码媒体（例如供打印或在显示设备上查看的图像及视频）或视频会议，而不是基于视觉的用户交互。

传统图像传感和获取设备会将3D场景的视觉信息转化为2D数组，将现实世界原本3D空间中的点作为离散的2D点映射在图像平面上（像素），同时会赋予其一系列数值，以反映其对应主色彩的亮度水平（像素值）。从3D世界中的视觉信息生成2D图像的过程，可以利用透视投影技术的齐次矩阵形式数学地描述为

$$[x'] = [C][x] \tag{4.1}$$

式中，[x] 代表的是3D世界中的点；[x′] 代表的是2D图像上的转化点；[C] 表示的摄像头转化矩阵，含有对应于摄像头的旋转和翻译及透视投影矩阵等一系列矩阵[1]。

然而，由于经过这一转化过程所成图像中的像素只保留了原本 3D 空间中的部分信息，因此经过处理之后，无法从捕捉获得的 2D 图像中真实恢复原有的 3D 信息。如何从单灰度图像重构 3D 平面是广泛研究的课题之一，且正不断取得重大突破[2, 3]。然而，基于单一 2D 图像传感设备进行实时交互的应用实施方面，其适用范围依然有限，且属于计算密集型应用。

人类视觉系统包含双目成像方案，能够感知景深，从而让我们能够自如漫步 3D 世界并与之交互。类似地，如果具有复杂交互方案的丰富人机界面任务能够在获取像素的色彩值之外，还能利用 3D 图像传感设备捕捉像素的景深或距离信息，则效果将更为优异。利用实时 3D 图像进行交互的应用程序开始日渐流行，尤其是在常见于客厅中的游戏和娱乐家用机系统，以及个人电脑的 3D 用户界面方面[4,5]。虽然眼下可以捕捉 3D 视觉信息的方式有许多，但占据主导地位的还是投影式结构光、立体 3D 成像法，以及飞行时间法成像技术这三种[6]。我们会在第 5 ~ 7 章中深入剖析这些 3D 传感技术。

以基于结构光的 3D 传感方法为例，这种方法会将数道具有固有图案或“结构”的光束（通常是红外线）投影到对象的物体或场景上。光线原有的图案会因物体或场景的形状发生变形，然后会使用图像传感器来进行捕捉。最后，会利用这一投影光学图案的变形来确定景深映射及物体和场景的 3D 几何形状。这一方案的概念示意图如图 4.2 所示[7]。在第 5 章中，Zhang 等人会详述结构光 3D 成像技术和应用的基本原理和最新进展。

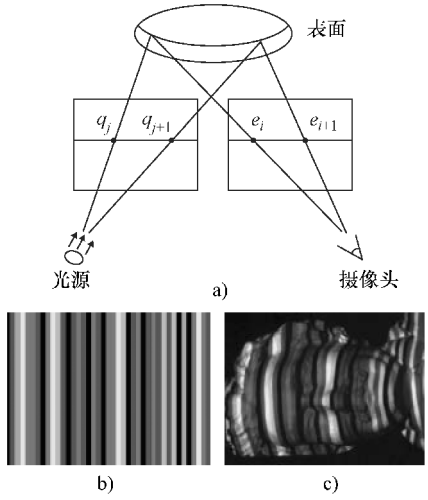


图 4.2　投射式结构光 3D 影像捕捉方法原理。a）照射图案投射在场景上，所得的反射图像经摄像头捕捉。某个点的景深由图案与影像之间的相对变形计算得出。b）投射条状图案示例。在实际应用中，通常会使用红外光，且图案会更为复杂。c）条状图案经 3D 物体反射后所得的捕捉图像。来源：Zhang，Curless and Seitz，2002. 转载已获 IEEE 许可

基于立体图像的 3D 计算机视觉技术试图模拟人类视觉系统，使用两台并排放置且经过校准的成像设备同步捕捉场景图像，之后会从双眼视差中提取出每个点的景深并映射到对应的图像像素上。这一技术的基本原理如图 4.3 所示，其中 C_1 和 C_2 两个摄像头聚焦焦距为 f，在各自的图像平面中的位置 A 和 B 处形成 3D 世界中点 P 的图像。

这一简单示例中，摄像头平行放置且经过校准。由图可知，垂直于两个摄像头中心连线的物体距离，与双眼视差成反比：

$$\text{depth} = fL/\Delta \tag{4.2}$$

用于确定双眼视差及从立体图像中获取景深信息的算法属于研究热点，不断有学者提出

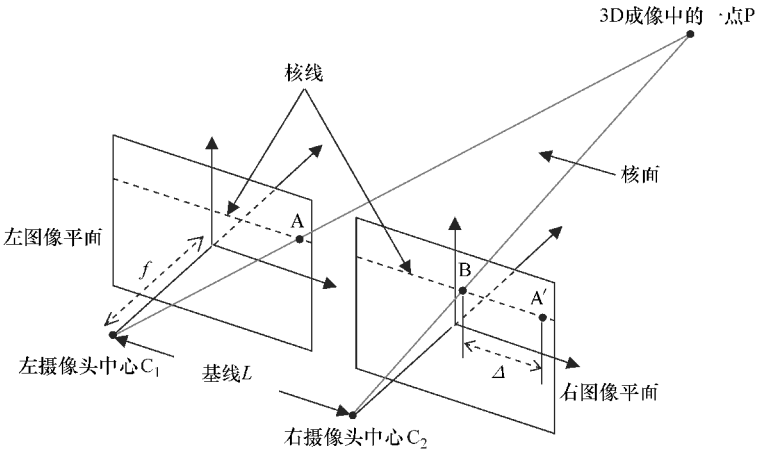


图 4.3　立体 3D 成像方法基本原理。以两台对齐且经过校准的摄像头的简单情况为例，两个摄像头的光学中点分别为 C_1 和 C_2，两者之间基线距离为 L。3D 世界中的点 P 经左右两个摄像头成像分别得点 A 和 B。右图像平面上的点 A′对应左图像平面上的点 A。B 和 A′核线之间的距离称为双眼视差 Δ，此值可知与点 P 到基线之间的距离（或景深）成反比

新的进展[8]。在第 6 章中，Lazaros 会详细介绍立体成像系统和算法发展。

飞行时间法 3D 成像方法利用调制红外光来照射物体和场景，计算光从成像设备出发后经物体或场景反射后回到光源的往返时间（常采用相移测量技术[9]），测出物体各点的距离，由此获得景深映射。这套系统通常具备全场范围成像能力，包括已调幅的照射源和图像传感器阵列。

图 4.4 说明了将反射光学信号的相移转换为点的距离的方法。反射信号如虚线所示，已经相对原发出信号之间有了 ϕ 的相移。该信号有所衰减，且检测设备也接受了部分背景信号，此处假设背景信号不变。在此配置下，可以计算出反射此信号的物体的距离为

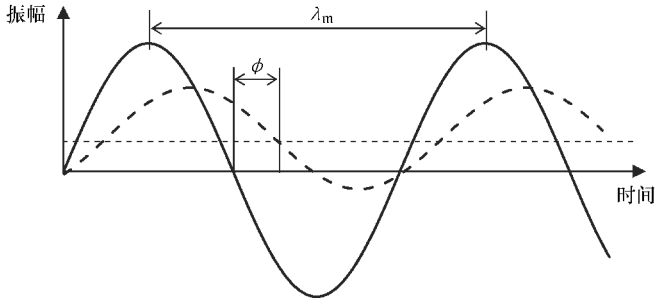


图 4.4　利用飞行时间法测量技术实现 3D 成像的原理。实线绘制的正弦曲线表示的是经光源照射在场景上的已调幅红外光线，虚线绘制的曲线则是成像设备检测到的反射信号。此处所示的反射信号已有所衰减，与发出信号之间有一个角度为 ϕ 的相对相移，且假设背景信号不变。距离或景深映射可利用相移和调制波长来确定

$$d = (\lambda_m/2) \times (\phi/2\pi) \tag{4.3}$$

式中，λ_m为光学信号的调制波长。在第 7 章中，Nieuwenhove 会说明飞行时间法景深成像及涉及交互式显示和系统应用的系统设计。

一般地，3D 成像设备的输出为距离图像（也称为景深映射），通常还会带有场景对应的彩色图像。相关示例如图 4.5 所示，其中景深值经过调整作为 8 位图像进行显示，由此可见距离传感设备越近的点就越亮。如图 4.1 的原理框图所示，利用任一种 3D 传感技术来生成图像及距离或景深信息是实现交互式显示的第一步。下一步是使用能够识别实时人类行为的算法，并利用这一数据进行输入。在下一节中，我们会介绍实现用于交互式应用的肢体语言识别任务的方法。

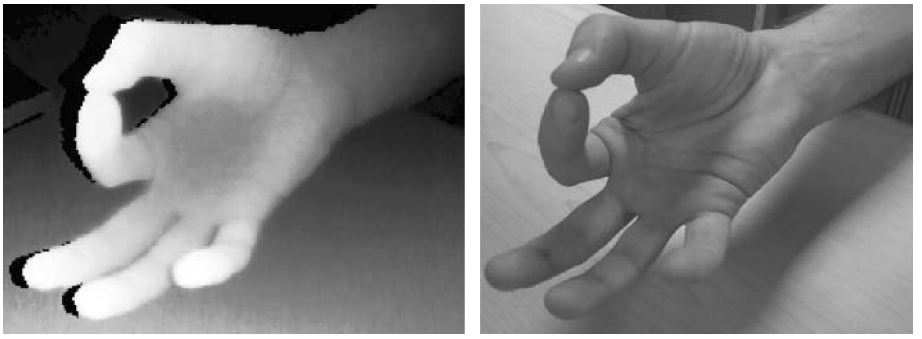


图 4.5　3D 成像设备的输出。左图为在 3D 传感设备前一只手的距离影像，又称为景深映射。景深映射上的灰度值随着图像上的点远离传感器而降低，较近的物体亮度较高。右图为对应的彩色图像。注：两幅图分辨率不同，且经过缩放和裁剪

4.3　姿势交互

在人与人之间的交互过程中，我们会大量使用手指、手掌、头部及身体其他部位、面部表情以及目光形成的姿势与动作，就算交流的主要方式是语言的情况下也仍是如此。相较而言，基于鼠标、键盘甚至触摸屏的传统电脑输入界面只能提供有限的交互体验。因此，如何利用电脑视觉技术在人机交互中添加姿势识别功能也是研究热点之一，许多学者心血倾注其中，就是为了能让用户体验更为自然、高效。

采用姿势识别系统的总体目标在于让电脑通过识别站在交互式显示屏前方的人类执行的姿势和动作，从而自动理解人类动作、指示和表达。

早期的 3D 空间中，人类姿势识别系统是通过基于布满传感器的穿戴设备（如手套）实现。在市面上出现设计用于提供手部姿势、位置和方向信息反馈的数据手套商品之后[10]，这一方法在人机交互研究者之间颇为流行。数据手套与主机相连，佩戴者能够驱动交互式显示屏上的 3D 手部模型，在 3D 环境中实时摆弄物体。关于基于手套的研究和发展，已经有一系列的综述文献在此方面做了深度调查[11, 12]。

虽然基于手套的方案体现出了 3D 交互的高效性与泛用性，但如果想要让更多的消费者

接受这种操作方式，则需要寻找一种不需要在身上穿戴这类跟踪设备的无标记实现方案。近期在先进高性价比小型图像设备方面的发展，高级图案识别算法，以及强大的计算资源，使得基于电脑视觉的自然姿势识别方案的实现成为可能。

近年来，已有许多研究致力于使用电脑视觉技术（基于实时捕捉和 2D 及 3D 图像序列分析的建模和统计方法）进行人类姿势识别，在这方面有大量的文献详细说明了研究结果，同时还有大量的综述与调查[13-16]。广义上来说，这类算法可以分为两大类：使用人类手掌与躯体形状及骨骼模型的基于 3D 模型的技术，以及使用从手掌或身体其他部分视觉图像中获取的 2D 灰度图像序列或低等级特征的基于视图的技术。这两种方法各有其优劣之处，最优化的解决方案应是各取两者优势而成的混合方法。图 4.6 说明了基于视觉的姿势识别流程基本算法的步骤和流程。

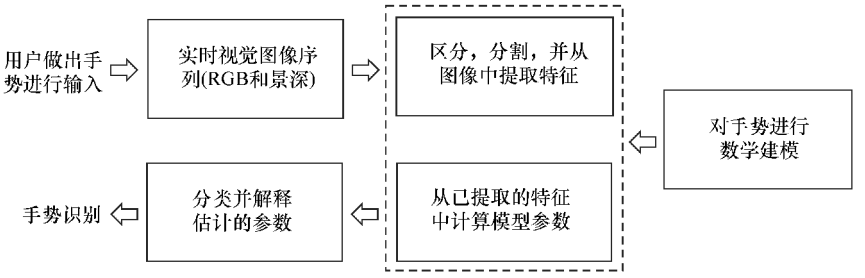


图 4.6　基于视觉的姿势识别流程基本算法的步骤和流程的框图

该流程从实时获取用户的视觉图像开始，在 3D 空间中为姿势提供输入。接收到输入的视觉图像之后，第一步是将目标物体局部化并分割成多个部分，例如用于识别姿势的手部或用于识别表情的面部。传统基于 2D 摄像头的系统会使用色彩或动作线索来进行图像分割，不过这种方法会因为背景色彩及环境光条件多变且无法预测而错误率较高。使用如前一节中介绍的 3D 传感摄像头，就能多提供一种重要的线索（例如手与成像设备之间的距离，可供初始检测及跟踪使用），同时还能根据景深进行分割。

Van den Bergh 和 VanGool 的实验可作为例证之一。他们的实验说明了在实时姿势识别的应用方面，采用景深摄像头进行手部分割取得的效果较基于色彩的方法更好[35]。如图 4.7 所示，采用色彩概率方法在手部与面部重叠的时候无法准确区分手部与面部，但在加入基于景深的阈值移除面部之后，就能够获得准确的分割结果。

在完成目标物体的识别和分割之后，便会从图像中抽取特定的特征，例如轮廓、边缘，或是诸如指尖、面部、肢体剪影之类的特殊特征。通常会专门为目标姿势开发一种数学模型，其中会包括该姿势的时间和空间属性，加入一系列参数之后形成建模。在特征检测与提取流程之后，会利用从图像中提取的特征计算这一模型中的各个参数。最后，通过对在分析步骤中估算出来的模型参数进行分类和解释，识别出用户做出的姿势。

基于 3D 模型的方法早期研究主要致力于找到适用于 3D 手部或躯体模型的运动学参数，使得模型的 2D 投影几何图形能够准确符合对应的基于边缘的图案[18]。简单来说，只有保

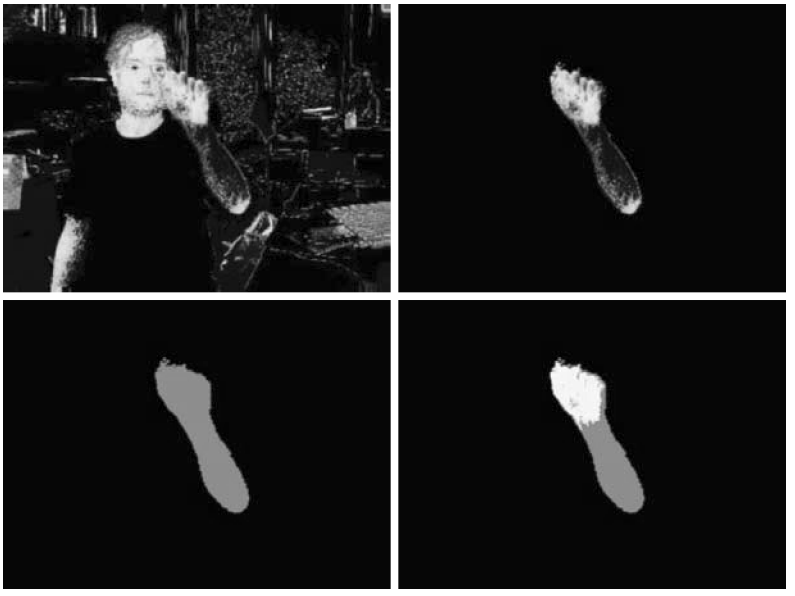


图 4.7　将景深和彩色图像用于手部区分的示例。左上：从 RGB 图像中获得的肤色概率。左下：加入阈值移除包括脸部在内的背景之后的手部景深图像。右上：仅包含利用阈值景深图像确定的前景像素中的肤色概率。右下：合成结果，显示分割出来的手部。
来源：Van den Bergh & Van Gool 2001。转载已获 IEEE 许可

证关节模型的外观与所捕捉的图像相似，才能确保 3D 模型的参数一致。在这类建模过程中，所使用的模型可分为立体模型与骨骼模型两类。立体模型基本上就是用一系列彼此相连，直径、高度有异的圆柱体来代表人类手部或躯体。这一类模型的匹配训练目标就是确定这些圆柱体的参数，使得 3D 模型能够对应上所记录的图像。

与之相对，骨骼模型则包含基于关节角度和线段长度的参数。不管是哪一类模型，使用基于生理学的约束条件来限制自由动作的角度范围，使之与人体解剖学一致，有助于限制分析空间。图 4.8 展示了同一种人类手势用不同模型模拟出来的情况[13]。

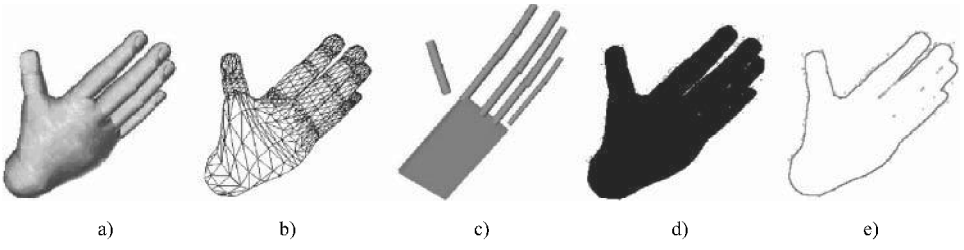


图 4.8　同一种手势的不同手部模型：a）3D 纹理立体模型，b）线框 3D 模型，c）3D 骨骼模型，d）2D 剪影，e）2D 轮廓或边缘。
来源：Pavlovic，Sharma and Huang 1997。转载已获 IEEE 许可

最近，Stenger 等学者使用了一种基于卡尔曼（Kalman）滤波的方法来估计人类手掌的姿势，这种方法能够将 3D 手掌模型的 2D 投影与从手掌图像中提取出来的轮廓之间的几何偏差降到最低[19]。虽然基于 3D 模型的方法计算量巨大，但这种方法在人机交互中的泛用性已取得广泛认可[13]。

随着低能耗、低成本的景深传感摄像头的出现，也有许多学者开始研究更为高效及稳定的方法。例如，Melax 等学者最近报告了一种计算方面更为高效的方法，能够实现手部的 3D 模型与通过 3D 成像设备获取的景深图像或 3D 点阵云按帧匹配，同时能够添加基于生理学确定的约束条件来追踪手部及个别手指的动作，就算偶有遮挡也能实现[20]。正如图 4.9 所示，该方法能够在配备了实时 3D 传感设备的交互式显示系统可视地表示 3D 空间中物体的操作。类似地，也有人报告了利用景深摄像头捕捉的实时距离数据来稳定跟踪人体姿势的方法[21]。

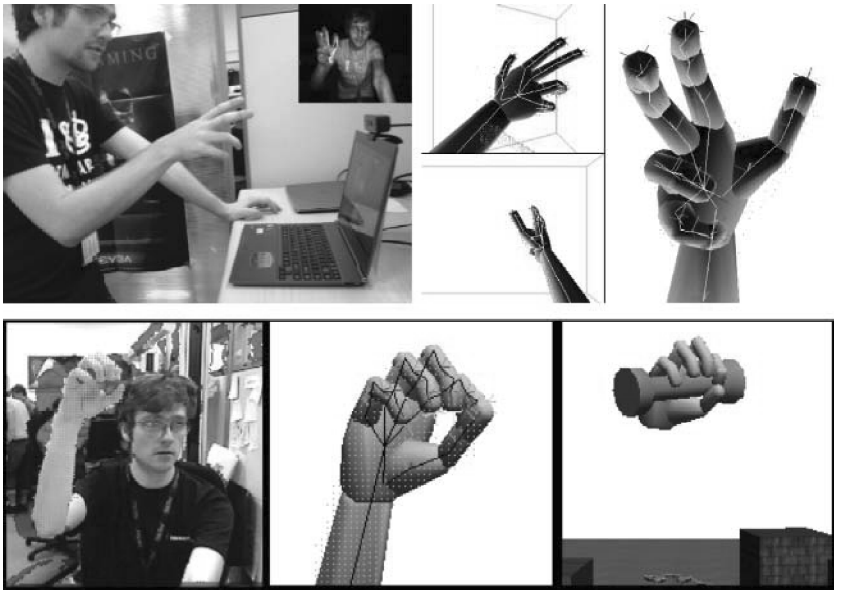


图 4.9　一种基于 3D 关节模型的手部骨骼跟踪技术，能够在 3D 空间中对物体实现细致多样操作。加入生理学约束之后，手部的 3D 模型能够匹配由景深传感设备获取的景深映射或点阵云

虽然基于观察（又称为基于外观）的方法有文章报告其计算量比基于模型的方法要小，但普遍认为这种方法不像基于 3D 模型的技术达到的效果那样能够普遍适用，因此也发现了一些人机交互方案的应用方面相对受限的地方。不过今年来有越来越多的人报告了采用这种方法的喜人进展。

整体来说，这种方法会预先定义一系列代表各种姿势的模板，然后将这些模板与视觉图像或特征进行比较。大多数早期研究重心主要放在相对简单的情况，例如使用从通用物体识别方案改进而来的算法对静态手势进行识别。然而，要想实现自然的人机交互，仅实现静态姿势的识别还远远不够，必须能够识别动态姿势才能够真正了解人类动作的意图。在此共识

下，从事这方面的学者报告了许多基于统计建模的方法和图像处理及规律识别的技术，以实现人类姿势、手势以及动作的自动识别，包括使用系列训练数据的主成分分析法、隐马尔可夫模型、卡尔曼滤波、粒子滤波、条件密度转移（“Condensation”）算法、有限状态机技术等[22-29]。

要将基于姿势的交互集成到实际应用中，就需要仔细考虑物理交互的人为因素方面，才能保证用户体验的舒适、直观。具体姿势的含义解读和内涵表示属于纯粹主观方面的认识，同样的意图，不同的人可能就会用不同的姿势进行表达，而就算是同一个人，根据时间和场合的不同，其使用的姿势也会有所差异。不少研究者已经对人类动态行为进行了详尽分析并做了报告[30, 31]。

进行姿势识别研究的一种有效方法，便是先理解人类动态行为并对其进行建模，然后接着根据姿势动作开发用于识别用户行为的算法。这就是基于隐马尔可夫模型方法背后的原理所在：将人类行为作为一个大型的心理或意图状态集合，每个个人控制特征和状态间转变的统计学概率都代表了一种状态[22-25]。简单来说，我们下一刻要做的事情，是我们现在这个时刻行为的平滑转变，由于紧接着当前动作之后通常都会有一系列未来动作可供选择，所以这一转变就可以用一个统计学概率来表示。所以，理解人类利用姿势进行交互的意图，就包括对当前手部与手指姿势的识别，以及对于接下来可能动作集合的预测。

以 Pentland 的成果为例，这个研究的模型基础是，皮质处理的基本元素能够用卡尔曼滤波进行描述，且各元素之间可以相互联系，构成更大的行为集合[25]。依照这一假设，他们将人类动态行为描述为一个以卡尔曼滤波表示的动态模型集合，彼此之间以马尔可夫概率转变链相连，并说明了这一技术在根据汽车驾驶员最初的准备动作预测其行为方面的成果。同样的方法也适用于普通的人机界面和交互。图 4.10 便是对此模型的概念说明，表示了用一个之间以概率转变相连且彼此互异的多状态马尔可夫动态模型对一个特定人类行为链进行描述的图示。通常来说，每个状态下面也会含有一系列的子状态，表明人类行为潜在的数学模型的复杂性。对用户行为的识别，加上对用户后续动作的成功预期与预测，能够提高系统对人类姿势输入的反应速度，让交互体验更为流畅。

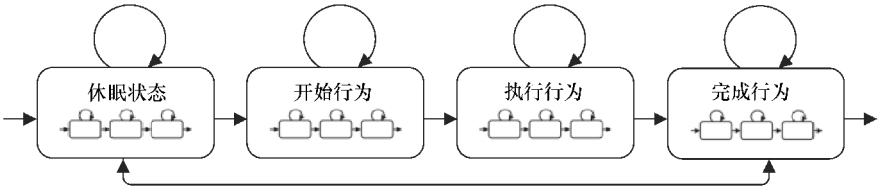


图 4.10　使用马尔可夫动态模型对人类行为链进行解释的简化概念示意图。
各个主要状态之间通过状态间转化概率相互连接。每个状态中包含若干相互连接的子状态。
顺畅的姿势动作识别系统，应包含对当前姿势及预测的解析，以及对下一行为集合的预测

如今致力于人机交互研究的学者除了继续发展传感技术和识别算法之外，也在继续研究适用于交互式显示屏和系统的肢体动作界面及分类[32, 33]。作为人类间交流的一部分，我们

通常会做出某些非特定的姿势，要连同当时所说的口头语言才能了解其中意图，例如说话的时候在空间中挥手或摆手指。从本质上面来说，这类肢体动作并没有严格定义，通常是在无意识的情况下做出，且其中含义会因人和环境而异。同时，我们经常也会做出某些刻意的姿势，意在表达特定的交流内容或指示，这些行为有可能是独立做出的沟通行为，也可能是用来强调口头交流的内容。

从实施人机交互的角度来看，此领域研究中更为注重的是后面这一类的姿势表达。以推进人机交互朝着人与人之间自然沟通交流方向为目标的 Quek 等学者在总结了现有人类姿势解析研究成果之后，从广义上将姿势实现分为两大类：姿势比对与姿势摆弄[32]。姿势比对方法会使用一系列抽象的静态或动态手势或姿势当作字典，然后系统的设计思路为，记录所执行的手势或姿势，然后将之与这一预先定义的姿势库进行比对，找出最接近的匹配条目。从实际结果来看，这个方法能够实现的，只有人们在现实世界日常生活交互过程中每天使用的各类姿势里的一小部分子集。

另一方面，姿势摆弄的方法则会让用户利用手部或肢体动作在交互式显示屏上操控虚拟物体，用户的实时操作会影响到屏幕上显示的物体的移动。虽然基于姿势摆弄的方法较姿势比对来说灵活性更高，但 Quek 和 Wexelblat[32, 33]同时也指出了当前的实现方法与我们在对话和自然交互中的行为比较时出现的不足。虽然专门为特定系统或应用做过优化的实现方法在这些系统或应用上的表现还算理想，但研究的目标始终还是开发出更为灵活的方法，能够让人类使用自然、直观的日常交互姿势与交互式显示屏上的内容进行交互成为可能。

虽然早期的研究和实现方法大多数只着眼于彩色图像的获取、分析和解释方面，但在追加使用景深摄像头提供的范围数据之后，也能够得出相对更为稳定、有效的算法路径[34-37]。最近在 3D 传感、建模和推理算法以及用户界面方面的进展，清楚表明我们能够在不远的将来离达到自然交互的目标更进一步。

我们在日常与其他人类进行交流的时候，除了利用手、手指和其他肢体做出姿势之外，我们还会广泛使用眼神、面部表情。在第 8 章中，Drewe 对眼神跟踪技术、算法以及使用基于眼神交互的应用进行广泛说明。

在电脑视觉领域，对于人类面部及面部表情的检测和识别也是学者广泛研究的方向之一[38-45]。在第 10 章中，Poh 等学者对作为多模态生物识别技术一部分的面部识别技术进行了综述。在面部检测和识别之外，我们还必须强调面部表情在多模态人类交流过程中的重要性，这是因为通过面部姿势表达出来的情感，能够增强或改变通过话语或手势传达出来的交流含义。这也难怪我们在交谈过程中为什么总是试图直视对方，因为这样才能准确了解对方的意图。

针对自动解析人类面部表情的算法研究的早期成果是根据传统的 2D 图像和基于灰度的分析得出，近年来已有一系列针对这一方面的详细综述文章发表可供参考[40, 41]。最近，在利用 3D 模型和通过 3D 传感设备生成的点阵云数据方面也有所进展[42-45]。例如，Wang 等学者报告了一种提取原始 3D 面部表情特征，并在经过非个人相关的表情识别方法处理后用来对表情进行分类的方法，他们的成果如图 4.11 所示[42]。

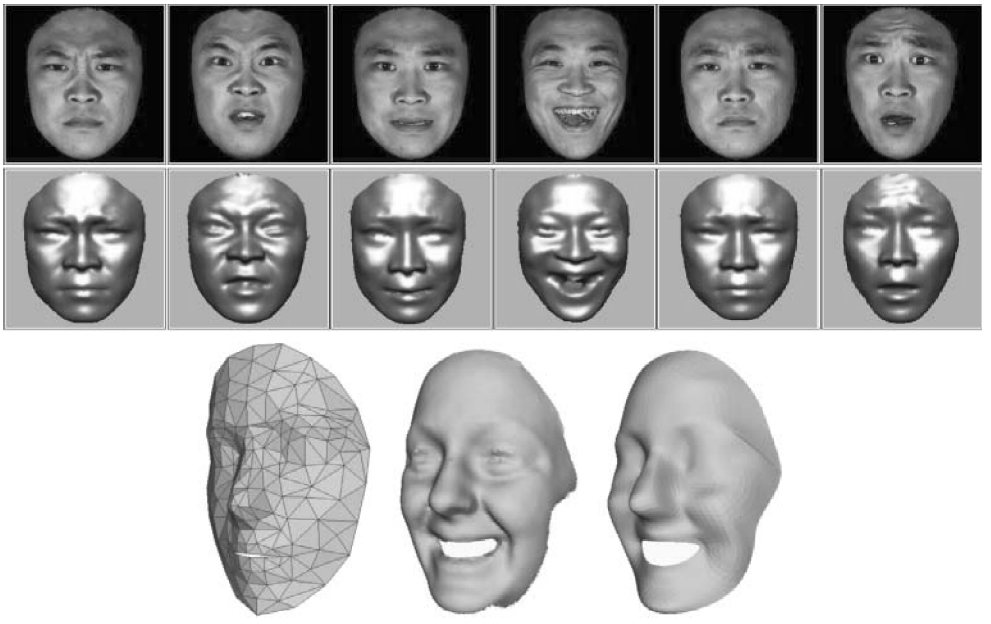


图 4.11　上图是由 Wang 等开发的 3D 面部范围模型示例，共有 6 种表情：愤怒、厌恶、恐惧、开心、悲伤，以及惊吓（从左到右）。上图第 1 行是纹理模型，第 2 行则是对应的着色模型。来源：Wang，Yin，Wei and Sun 2006。转载已获 IEEE 许可。下图是 Mpiperis 等在面部及表情识别中使用的 3D 面部模型方法。左侧是最初的 3D 网格模型。中间是使用 3D 成像技术捕捉的人类脸部 3D 表面。右侧是 3D 网格模型贴合至 3D 表面后的效果。来源：Mpiperis，Malassiotis，Strintzis 2008。转载已获 IEEE 许可

另外一个例子则是 Mpiperis 等学者报告的能够同时达成无关个人的面部表情识别及无关表情的面部识别的基于 3D 模型的方法[43]。如图 4.11 所示，这种方法将可变形的面部表面网格模型贴合至通过 3D 传感器从人类面部获取的 3D 点阵云上。原本的网格模型设计为中性，且在获得需研究的面部的 3D 点阵云，并将这一网格模型贴合上去之后，它也能符合所表现出来的表情。然后这一面部表情会通过数学方法进行识别，在先前确定的集合中寻找对应项目。如今世界上许多实验室还在继续对算法进行研究，也不断取得进展，若能将面部表情理解与手部及身体肢体语言识别技术相结合，就一定能在用于未来的交互式显示器与系统上的基于视觉的人类界面方案带来重大补强。

4.4　结语

能够理解人类在 3D 环境中以自然方式表达的行为和指令并对其进行回应的电脑，早已出现在科幻先锋的畅想之中。例如在 2002 年上映、由史蒂夫·斯皮尔伯格执导、广受好评的美国科幻电影《少数派报告》中，就描绘了一个 2054 年的未来世界，其中已有通过立体界面进行操作的电脑，用户在面前的 3D 空间中通过手势与显示屏上的多媒体内容互动。虽

然这只是对彼时未来技术的大胆预想，但如今的科学家和工程师却已经开发出了相应的技术和系统，或许能够提前数十年实现这一梦想。实际上，现实世界的实现手段已经更为简洁，随着电脑视觉技术的发展，将来就能与电影里面一样，不穿戴任何手套或其他设备直接通过 3D 肢体动作与电脑进行交互。

在本章中，我们综述了基于视觉的 3D 传感和交互技术的发展。虽然电脑视觉方面的研究早期成果大多数都是通过对利用 2D 摄像头获取的灰度图像进行分析而得，最近 3D 成像技术的发展使得景深映射和 3D 点阵云的高效和实时获取成为可能。另一方面，姿势识别算法方面的研究，无论是基于 3D 模型的方法还是基于图像或特征的方法，都在近年取得显著进展。这些成果，再加上人类动态行为的理解与建模方面的研究，让在交互式显示屏前的 3D 空间进行自然人机交互进一步成为现实。能够“感应”到我们手指触控的交互式显示屏已经无处不在，如今再加上先进的视觉传感和识别技术，让新一类能够“看见”并“理解”其面前的 3D 空间中用户行为的交互式显示屏的研发成为可能。

参 考 文 献

1. Trucco, E., Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall.
2. Saxena, A., Sun, M., Ng, A. (2008). Make3D: Learning 3-D Scene Structure from a Single Still Image. *IEEE Transactions on Pattern Analysis, Machine Intelligence*.
3. Chen, T., Zhu, Z., Shamir, A., Hu, S., Cohen-Or, D. (2013). 3-Sweep: Extracting Editable Objects from a Single Photo. *ACM Transactions on Graphics* **32**(5).
4. Microsoft Corporation (2013). www.xbox.com/en-US/kinect. Retrieved Nov 16, 2013.
5. Intel Corporation (2013). www.intel.com/software/perceptual. Retrieved Nov 16, 2013.
6. Bhowmik, A. (2013). Natural, Intuitive User Interfaces with Perceptual Computing Technologies. *Inf Display* **29**, 6.
7. Zhang, L., Curless, B., Seitz, S. (2002). Rapid Shape Acquisition Using Color Structured Light, Multi-pass Dynamic Programming. *IEEE Int Symp 3D Data Proc Vis Trans* 24–36.
8. Brown, M., Burschka, D., Hager, G. (2003). Advances in Computational Stereo. *IEEE Trans Pattern Analysis, Machine Int* **25**, 8.
9. May, S., Droeschel, D., Holz, D., Fuchs, S., Malis, E., Nuchter, A., Hertzberg, J. (2009). Three-dimensional mapping with time-of-flight cameras. *Journal of Field Robotics – Three-Dimensional Mapping, Part 2*, **26**(11–12), 934–965.
10. Zimmerman, T., Lanier, J., Blanchard, C., Bryson, S., Harvill, Y. (1987). A hand gesture interface device, Proceeding of the Conference on Human Factors in Computing Systems. *Graphics Interface*, 189–192.
11. Sturman, D., Zeltzer, D. (1994). A survey of glove-based input. *IEEE Computer Graphics, Applications* **14**(1), 30–39.
12. Dipietro, L., Sabatini, A., Dario, P. (2008). A Survey of Glove-Based Systems, Their Applications, *IEEE Transactions on Systems, Man, Cybernetics, Part C: Applications, Reviews* **38**(4), 461–482.
13. Pavlovic, V., Sharma, R., Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans Pattern Analysis, Machine Intelligence* **19**(7), 677–695.
14. Derpanis, K. (2004). *A Review of Vision-Based Hand Gestures*. Internal report, Centre for Vision Research, York University, Canada.
15. Mitra, S., Acharya, T. (2007). Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man,, Cybernetics – Part C: Applications, Reviews* **37**, 311–324.
16. Wu, Y., Huang, T. (1999). *Vision-Based Gesture Recognition: A Review, Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pp. 103–115, Springer-Verlag.

17. Garg, P., Aggarwal, N., Sofat, S. (2009). Vision Based Hand Gesture Recognition. *World Academy of Science, Engineering, Technology* **25**, 972–977.
18. Rehg, J., Kanade, T. (1994). Visual tracking of high DOF articulated structures: An application to human hand tracking. *European Conference on Computer Vision B* **35–46**.
19. Stenger, B., Mendonca, P., Cipolla, R. (2001). Model-based 3D tracking of an articulated hand. *IEEE Conference on Computer Vision, Pattern Recognition* **II**, 310–315.
20. Melax, S., Keselman, L., Orsten, S. (2013). Dynamics Based 3D Skeletal Hand Tracking. *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics, Games*, 184–184.
21. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S. (2012). Real Time Human Pose Tracking from Range Data. *Lecture Notes in Computer Science* **7577**, 738–751.
22. Yamato, J., Ohya, J., Ishii, K. (1992). Recognizing human action in time-sequential images using hidden Markov model. *Proceedings of IEEE Computer Vision, Pattern Recognition*, 379–385.
23. Kobayashi, T., Haruyama, S. (1997). Partly-hidden Markov model, its application to gesture recognition. *IEEE International Conference on Acoustics, Speech, Signal Processing* **4**, 3081–3084.
24. Yang, J., Xu, Y., Chen, C.S. (1997). Human action learning via hidden Markov model. *IEEE Trans on Systems, Man,, Cybernetics Part A: Systems, Humans* **27**(1), 34–44.
25. Pentland, A., Liu, A. (1999). Modelling and Prediction of Human Behavior. *Neural Computation* **11**, 229 – 242.
26. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing* **50**(2), 174–188.
27. Isard, M., Blake, A. (1998). Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision* **29**(1), 5–28.
28. Bobick, A., Wilson, A. (1997). A State-based Approach to the Representation, Recognition of Gesture. *IEEE Transactions on Pattern Analysis, Machine Intelligence* **19**, 1325–1337.
29. Imagawa, K., Lu, S., Igi, S. (1998). Color-based hands tracking system for sign language recognition. *Proceedings of IEEE International Conference on Automatic Face, Gesture Recognition*, 462–467.
30. Aggarwal, J., Cai, Q. (1999). Human Motion Analysis: A Review. *Computer Vision, Image Understanding* **73**(3), 428–440.
31. Gavrila, D.M. (1999). The visual analysis of human movement: a survey. *Computer Vision, Image Understanding* **73**, 82–98.
32. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., McCullough, K.E., Ansari, R. (2002). Multimodal human discourse: gesture, speech. *ACM Transactions on Computer-Human Interaction* **9**, 171–193.
33. Wexelblat, A. (1995). An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction* **2**, 179–200.
34. Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J. (2013). A survey on human motion analysis from depth data, Time-of-Flight, Depth Imaging Sensors, Algorithms,, Applications. *Lecture Note in Computer Science* **8200**, 149–187.
35. Van den Bergh, M., Van Gool, L. (2011). Combining RGB, ToF cameras for real-time 3D hand gesture interaction. *IEEE Workshop on Applications of Computer Vision*, 66–72.
36. Kurakin, A., Zhang, Z., Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. *European Signal Processing Conference,* 1975–1979.
37. Liu, X., Fujimura, K. (2004). Hand gesture recognition using depth data. *Proceedings of the Sixth IEEE international conference on Automatic face, gesture recognition*, 529–534.
38. Yang, M., Kriegman, D., Ahuja, N. (2002). Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis, Machine Intelligence* **24**(1), 34–58.
39. Tolba, A., El-Baz, A., El-Harby, A. (2006). Face Recognition: A Literature Review. *International Journal of Signal Processing* **2**(2), 88–103.
40. Fasel, B., Luttin, J. (2003). Automatic Facial Expression Analysis: a survey. *Pattern Recognition* **36**(1), 259–275.
41. Pantic, M., Rothkrantz, L. (2000). Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis, Machine Intelligence* **22**, 1424–1445.
42. Wang, J., Yin, L., Wei, X., Sun, Y. (2006). 3D facial expression recognition based on primitive surface feature distribution. *IEEE Conference on Computer Vision, Pattern Recognition* **2**, 1399–1406.
43. Mpiperis, I., Malassiotis, S., Strintzis, M.G. (2008). Bilinear Models for 3-D Face, Facial Expression Recognition. *IEEE Transactions on Information Forensics, Security* **3**(3), 498–511.
44. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M. (2008). A High-Resolution 3D Dynamic Facial Expression Database. *The 8th International Conference on Automatic Face, Gesture Recognition,* 17–19.
45. Allen, B., Curless, B., Popovic, Z. (2003). The space of human body shapes: reconstruction, parameterization from range scans. *Proceedings of ACM SIGGRAPH* **22**(3), 587–594.

第5章

实时3D传感与结构光技术

Tyler Bell，Nikolaus Karpinsky，Song Zhang
爱荷华州立大学机械工程系

5.1 引言

随着近期3D计算方法的不断进步，动态实时3D传感在许多领域已经变得至关重要（如制造业、医学、电脑科学、国土安全和娱乐）。3D传感器也开始成为在显示器上进行3D互动的常用工具。微软公司的Kinect就是一个很好的例子。

3D传感技术在过去数十年里不断进步，最近几年更是进展飞速。许多技术得以开发，包括飞行时间法、立体观察、时空立体观察、结构光、数字全息术和数字条纹投影。每种科技开发的初衷都是满足特定需求，并在专门应用领域发挥绝佳作用，但是总的来说没用一种单一技术能与3D传感的巨大需求相匹配。Zhang（2013）[1]编制了一本集合了各主要3D传感技术的手册，这为工程师选择适当的技术以满足特定需求提供了灵感。

不管是在科学研究领域还是在工业实践领域，结构光方法都有望成为最重要的3D传感技术之一。实时3D传感在最近几十年实现并成为一项主流技术，是因为今天的个人电脑具备强大的计算能力，可以满足实时传感对于计算的高要求。甚至平板电脑也具备了满足这样需求的计算速度。

实时3D传感通常是指以至少24Hz的速度获取、处理和再现感知到的3D数据。虽然面临着严峻的挑战，但是过去几年中不断发展的扫描技术，包括微软公司的Kinect和一些由美国爱荷华州立大学开发的技术已经解决了这些挑战。有趣的是，几乎所有的实时3D传感技术都是光学方法，这意味着现场捕捉不需要与传感器物理接触。然而，因为都是基于光学原理的，所以这些系统很难传感到具有某些光学特征的表面（比如发光的、透明的或者纯黑色的）。

在这些结构光方法中，数字条纹投射（DFP）技术较为独特，因为其结构化模型呈现正弦并由激光干涉仪产生。相较其他的结构光技术，DFP技术已经被证明具有压倒性优势，并

在众多学科中广泛使用[2,3]。

只有在实际应用中有用武之地，实时 3D 传感技术才能进一步发展。人机交互就是 3D 传感技术的一项重要应用，与计算机交互对响应速度要求很高，因此它在本质上是实时的。大多数高分辨率的实时 3D 传感技术不仅能捕捉黑白纹理，还能捕捉带有方向性光的纹理。这在人机交互应用领域可能还不理想，因为在应用中需要纹理的自然色。本章将展示我们研发的运用近红外（NIR）光进行 3D 传感和实时同步捕捉自然色纹理的技术。

由于数字视频投影仪的速度上限，典型的结构光方法的速度可以达到 120Hz[4]。此外大多数视频投影仪是非线性的，没有非线性校准和修正，难以生成高质量的相位。虽然很多非线性的校准技术已经开发出来了[6-11]，也通过实践证明可行，但我们发现问题没那么简单，因为投影仪的非线性伽马实际上会随着时间改变。

平方二进制散焦技术提出的初衷是克服传统的 DFP 技术的局限性[12]。平方二进制散焦技术只需要 1 位二进制结构化模型，而不是 8 位的灰度模型。于是在定位远离投影仪焦平面的物体时，正弦条纹模型就自然融合在一起了。该技术因为只用到两个灰度值，所以不被投影仪的非线性影响。此外，因为只需要 1 位结构化模型，二进制散焦技术大大降低了数据传输率，从而使得大于 120Hz 的 3D 形状测量速度成为可能。利用数字光学处理（DLP）发现平台，Zhange 等学者（2010）[13]成功地开发出实现数万赫兹的 3D 形状测量速度的系统。本章也将对我们在超高速 3D 传感方面取得的进步进行说明。

本章综述了结构光技术的原理。应该强调的是，本章提到的大部分技术已经在会议论文集或期刊上发表。本章绝不是实时 3D 形状测量技术的详尽调查。它关注的是我们过去数年一直在探索的技术，并仰赖于先前的研究出版物[1,4,13-15]。

5.2 节综述了结构光方法，并总结了过去几年里开发出来的结构光模型。5.3 节探讨了对结构光系统的校准问题。5.5 节举例说明了如何运用结构光方法进行 3D 传感。5.6 节揭示了实时 3D 传感人机互动的潜在应用领域。5.7 节讨论了我们近期在使用二进制散焦超高速 3D 传感方面的研究。5.8 节对本章进行了总结。

5.2　结构化图案汇编

光学 3D 传感方法由于其无创性而被广泛使用，在这种情况下，不能用物理方法测量被捕获的表面。立体视觉技术使用两个相机从不同视角捕获两个 2D 图像；这是模拟与人类视觉相同的过程。景深信息通过三角测量来恢复，这可以通过知道两个摄像机之间的对应点来完成。在识别两个二维图像之间的相应对以恢复景深的情况下，如果物体表面没有强烈的纹理变化，则立体视觉技术难以达到高精度。例如，该方法不能从两个均匀的白色平坦表面获得任何景深信息，因为每个纹理中的纹理看起来大致相同。有关这些技术的详细讨论可以在本书的另一章中找到。

结构光系统在某种意义上是相似的，但是不使用两个 2D 相机，而是使用一个投影仪和一个相机。投影仪投影某些类型的编码结构化图案以轻松确定对应关系。对于结构光系统，

结构化图案设计是决定最终可实现的分辨率、速度和准确性的关键因素之一。本节回顾了几个广泛使用的结构化图案编码。

5.2.1　2D 伪随机汇编

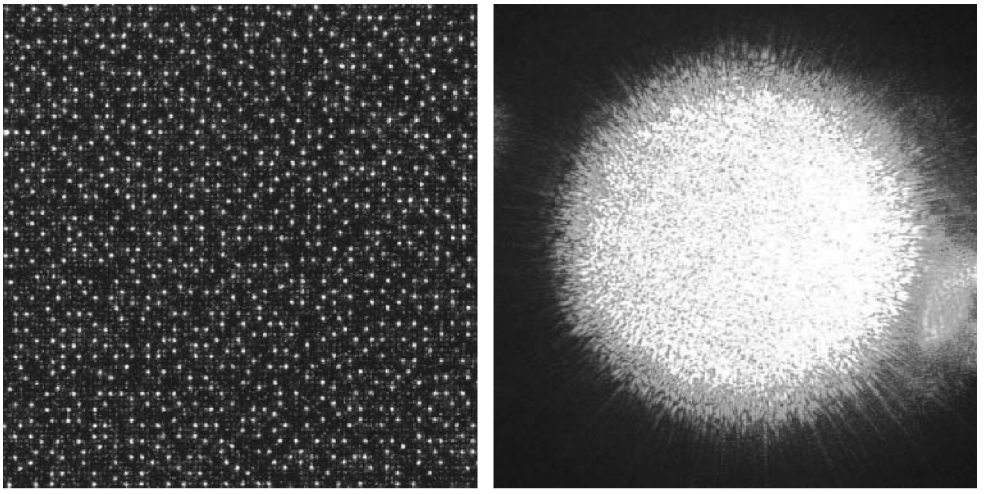
结构光技术已经被广泛研究并应用于计算机视觉、机器视觉和光学测量等领域。结构光技术与之前提到的立体观察方法相似，除了其中一个摄像头换成了投影仪[17]。要在摄像头和投影仪的像素之间建立点对点的映射，一个自然的方法是对被投影的画面进行编码，过程中通过整个画面的 x 和 y 轴上像素都是唯一的。也就是说，每一个像素可以被上面的信息所标记[17]。运用生成伪随机图案或通过使用激光源产生的自然散斑图案等方法已经得以研发[18]。图 5.1 展示了用于 3D 传感的图案。

在伪随机二进制阵列的方法中，$n_1 \times n_2$ 阵列通过一个伪随机序列解码，以确保 $k_1 \times k_2$ 在阵列的任意位置上的任意核都是唯一的。这个伪随机序列解码的 $n_1 \times n_2$ 阵列是通过本原多项模 n^2 的方法派生的，用数学表达就是

$$2^n - 1 = 2^{k_1 k_2} - 1 \tag{5.1}$$

$$n_1 = 2^{k_1} - 1 \tag{5.2}$$

$$n^2 = 2^n - 1/n_1 \tag{5.3}$$



a)　　b)

图 5.1　伪随机图案示例。a）微软 Kinect 使用的伪随机图案；b）伪随机图案伴随镭射光的自然出现。来源：Steve Jurvetson [2]

微软公司的 Kinect 是目前流行消费产品运用伪随机汇编方法执行计算机视觉的绝佳例证。试想有一个红外投影仪、一个红外摄像头，一束由投影仪投射出来的红外光通过衍射光栅，从而把光线聚焦成一组红外点[19]。伪随机分布图案被设备的红外线摄像头捕捉，因为设备知悉被投影的图案，3D 场景可以以三角测距的方式构造。虽然这项技术的特定实施细

节具备了专利，但在更高的层面上，这种技术和其他结构光方法是相似的。

总的来说，伪随机汇编方法的优势是容易理解、容易实现 3D 传感。但是这项技术也有不足，例如，对噪声的忍耐度不高。此外，实现高空间分辨率也比较难，因为投影仪在 u 和 v 方向上的分辨率都是有限的。

事实证明，在结构光系统正确校准的情况下，没有必要通过建立 2D 的独立相关性来实现 3D 传感。换言之，要在结构光系统中确定 xyz 坐标，除了已校准的系统约束方程式之外，只需要一个额外的约束方程式（将在 5.3 节中讨论）。

因此，结构化图案可以在一个方向上改变，但在另一个方向上保持不变。这就消除了两个方向上的空间分辨率限制，因此得以广泛运用于计算机视觉。

5.2.2　二进制结构化汇编

图 5.2 显示了用结构光技术进行 3D 传感的示意图，图上所示的条纹可以在 u 或者 v 方向上变化，但不能在两个方向都变化。

这个系统包括三个主要单元：图像采集单元（A）、投影单元（C）和要进行测量的 3D 对象（B）。投影仪直接在对象的表面照亮垂直结构条纹，如果从另一个角度观看，对象的表面会将这些条纹从直线扭曲成曲线。然后摄像头从不同于投影角度的另一个角度下捕捉扭曲的结构图像，这样就形成了一个三角形。在这样的系统中，通过分析已知结构的变形图案获得结构化汇编信息，并以此建立对应。也就是说，系统知道应该投影哪个图案，并且通过从对象表面确定和测量图案的突变情况，对应得以建立。

二进制编码结构化图案（只有 0 和 1）在结构光系统中广泛应用，是因为：

1）简单：很容易实现，因为编码和解码算法是简单的。

2）稳健：表面特性的变化稳定，因为只有 2 个灰度水平（0 和 255）被使用和预期。

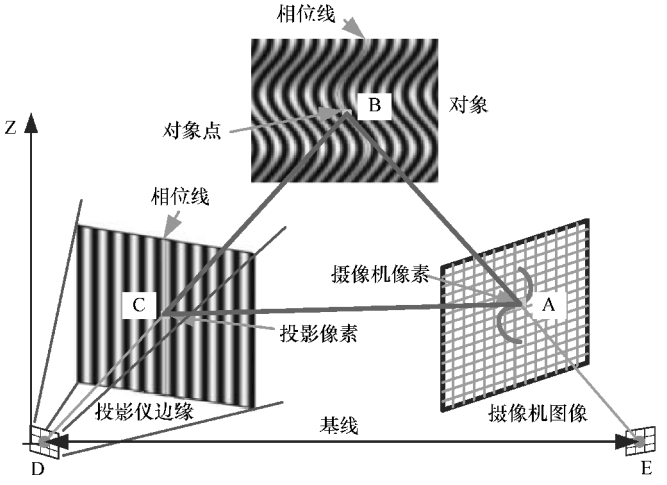


图 5.2　运用结构光技术的 3D 成像系统示意图。来源：Zhang 2010。转载获得 Elsevier 许可

在这种方法中，每个图案的每一个像素都有自己独特的 0 和 1 的码字[17]。要确定一个像素的独特的码字，一组二进制图案要依次投射和捕获，如图 5.3 所示。对每个码字的每一位的图案进行投影和捕获；找回独特的码字，再比较每个图像的每个像素，并在码字的对应位显示结果。

一旦结构化图案已经解码，并且得知每个像素的独特的码字，对应关系就可以建立。从本质上讲，这一步包括将像素的码字转化为投影坐标。如果摄像头坐标系投影坐标是已知的，假设系统经校准，三维坐标可以通过下式三角定位：

$$[u^c, v^c, 1]^{\mathrm{T}} = [P^c][x^w, y^w, z^w, 1]^{\mathrm{T}} \tag{5.4}$$

$$[u^p, v^p, 1]^{\mathrm{T}} = [P^p][x^w, y^w, z^w, 1]^{\mathrm{T}} \tag{5.5}$$

式中，P^c是摄像头矩阵；P^p是投影仪矩阵；（u^c，v^c）是摄像头的 p 坐标；（u^p，v^p）是投影仪坐标[20]。因为汇编只在一个方向，即水平，v'是未知的，这样就有三个方程式和三个未知数。这些方程式是一个线性方程式组，解出它们将得出世界坐标（x^w，y^w，z^w）。本章 5.3 节将讨论进一步探讨校准技术的细节，提供更多线性结构光系统的背景以及结构光系统如何进行校准来寻找矩阵 P^c和 P^p。

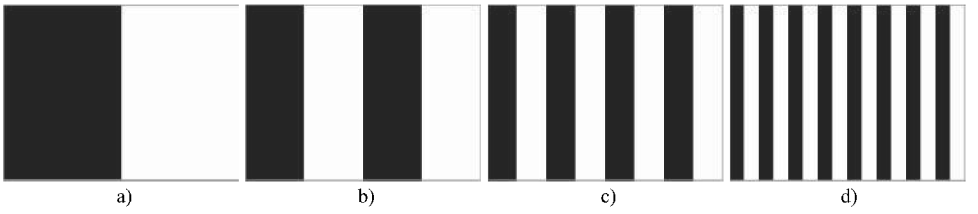


图 5.3　汇编二进制图案的示例。要恢复该码字，无论像素是黑色还是白色，都要对每个画面进行简单的比较，其结果出现在相应位的码字中

虽然二进制码因其简单性和处理系统中噪声的能力而非常方便，但它们并不是没有缺点。二进制码的两个显著的缺点是：空间分辨率；大量需要汇编的图案。

二进制码的空间分辨率被投影仪分辨率和摄像头分辨率限制。图 5.4a 展示了一个二进制图案，图 5.4b 展示了其对应的截面。在这里，黑色代表二进制 0，而白色代表二进制 1。在图 5.4b 中，以 M 和 N 中间的一个条纹来说，所有点都有相同的灰度值，因此它们不能被区分。所以，对于二进制方法来说，达到投影仪的像素级的空间分辨率是困难的，因为这个条纹的宽度一定要比投影仪的其中一个像素要大。此外，由于它不能达到像素级的配对，这项技术要达到非常高的测量精度是很难的。

第二个缺点是大量需要汇编的图案，因为一个二进制码仅使用两个灰度水平，即二进制的 0 和 1。这就限制了可以为 n 个二进制结构化图案生成最大 2^n 个独特码字。因此，要实现密集的 3D 传感，许多二进制结构化图案是必需的，从而造成了这项技术在应用于如实时 3D 传感这样的高速应用时不那么具有吸引力。

5.2.3　多进制汇编

虽然具备简单、表面特性稳定、耐噪声等优点，二进制结构化图案也有其缺点，特别是

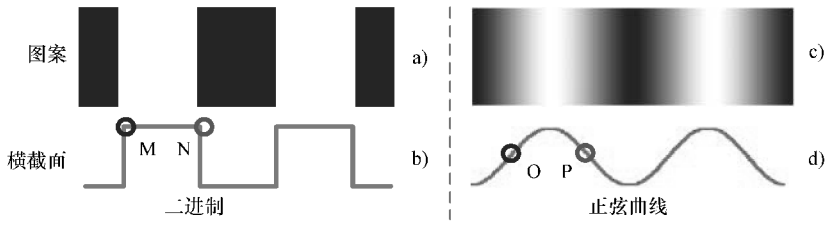


图 5.4　二进制和正弦结构化图案的比较。a）典型的二元结构化图案；b）图 a 中所示的二进制图案的截面；c）典型的正弦条纹图案；d）图 c 中所示的图案的截面

当涉及大量需要汇编的图案时。为了在不牺牲空间分辨率的前提下解决这个问题，可以利用更多的灰度值[21]。在引入附加灰度值减少结构光汇编的稳定性的同时，它提高了其传感速度；在运用于使用扫描仪的特定应用时必须对这些取舍加以权衡。

多进制方法不是只使用两个灰度值（0 和 255）来为每一个像素创建独特码字，而是利用这些值之间的一个子集。最极端的情况下会使用所有的灰度值（即全范围 0 ~ 255）。每个像素的字码可以通过灰度比计算[22,23]确定。这一计算的最基本的形式是假设一个灰度谱 0 ~ 255，包括可沿垂直列放置的线性“楔形”[17]。然后两个图案被投影到场景中：一个有上述楔形，一个没有（恒定灰度）。接下来，每个像素的比例可以通过这些值导出，用以计算出现在两个被捕捉并已经投影在场景中的图案的比例。在这种方法的替代方法中，许多楔形被按顺序投影到场景中，与此同时持续增加楔形的周期[23]。

灰度比方法在传感速度方面表现很好，因为它需要汇编的图案较少。然而，这种方法以及上述方法对噪声更为敏感（相对于二进制结构化汇编），它们受限于投影仪的分辨率，而且对于摄像头和投影仪散焦非常敏感。为了克服这种方法的局限性，可以使用相移技术和各种结构化图案，如三角形[24]和梯形[25]。这些技术可以实现摄像头像素的空间分辨率，并降低对于散焦的敏感程度；然而，它们仍然无法完全免受散焦问题的影响。

5.2.4　连续正弦相位汇编

如同 Zhang[4]标注的那样，不管是二进制、多进制，三角形还是梯形图案，只要它们是适当模糊的，最终都会成为正弦。模糊效果通常发生在一个摄像头在焦点之外捕捉到一个图像的时候和一个对象的所有的突出特征混合在一起的时候。因此，正弦图案似乎是一个自然的选择。如图 5.4c 和图 5.4d 所示，灰度随图像逐点变化。因此，达到像素级的分辨率是可行的，因为水平相邻像素之间的灰度值是可辨的。由于空间分辨率高，投影仪和摄像头之间的对应关系可以更精确地确定，这就允许了更高的精度。

正弦图案（也被称为条纹图案）有可能达到像素级的空间分辨率，因此，长期以来被从光学计量角度研究。在这个实例中使用的条纹图案是通过激光干涉产生的。数字条纹投影（DFP）技术并非使用产生可能危及测量质量散斑噪声的激光干涉，而是采用数码视频投影仪来投影计算机产生的正弦图案。原则上，DFP 技术是一种特殊的三角结构光方法，在这种

方法中的结构化图案灰度呈正弦变化。不像基于灰度的方法，DFP 技术使用相位信息来建立对应关系，通常在表面纹理的变化上相当稳定。

相位可以使用的傅里叶变换轮廓术（FTP）[26]通过傅里叶分析方法来获得。因为只有一个单一的条纹图案是必需的[27]，FTP 在简单的动态 3D 形状测量方面运用广泛。然而，这种方法对于噪声和表面纹理变化非常敏感。为了提高其稳定性，Kemao（2004）[28]提出了窗口傅里叶变换（WFT）方法。虽然成功，但由于 FTP 方法中空间傅里叶变换需要的基本限制，WFT 仍然不能实现对一般复杂程度的 3D 结构进行高质量的 3D 测量。改良 FTP 的方法是为了通过使用两个图案来获得更高质量的相位[29]。然而，没有突显的纹理和/或几何变化，表面测量仍然受限。

要测量通用表面，必须使用至少三个条纹图案。如果使用三个或更多的正弦结构化图案，并且它们的相位信息变换，可以在不知道相邻信息的情况下获得逐个像素的相位，从而使其免受表面纹理变化影响。这些方法通常被称为移相方法。

5.2.4.1　多步骤移相技术

移相方法因为下列优点在光学测量领域得以广泛运用[31]：

1）密集 3D 形状测量。相移技术允许逐个像素的 3D 形状测量，使得实现摄像头像素级的空间分辨率成为可能。

2）不受环境光的影响。移相方法不是利用灰度，而是分析相位信息的结构化图案。环境光的影响被自动取消了，但如果环境光比投影灯强太多，该信号的信噪比（SNR）可能会被牺牲掉。

3）对表面反射率变化不太敏感。通常情况下，移相方法运用反正切函数逐点计算相位，因为每一个像素点是恒定的，表面反射率信息的影响得以消除。

4）允许高速 3D 形状测量。由于整个测量区域可以一次捕获和处理，这种技术，以及其他结构光技术，可以实现高测量速度。

5）可以实现测量高精度。不像其他的结构光技术，移相方法允许投影仪和摄像头之间没有任何插值的精确的亚像素对应。因此，在理论上，如果校准正确，它可以实现高精度的 3D 形状测量。

多年来，众多的相移算法已经得以开发，包括三步法、四步法和多步算法。对于高速应用，常用三步相移算法，因为它获得逐个像素的相位所需的图像最小。具有相同的相位偏移的多步相移算法可以描述为

$$I_n(x,y) = I'(x,y) + I''(x,y)\cos(\phi + 2n\pi/N) \tag{5.6}$$

式中，$n = 1, 2, \cdots, N$；$I'(x, y)$ 是平均灰度；$I''(x, y)$ 是灰度调制；$\phi(x, y)$ 是要解出的相位。

$$\phi(x,y) = \arctan\left[\frac{\sum_{n=1}^{N} I_n(x,y)\sin(2n\pi/N)}{\sum_{n=1}^{N} I_n(x,y)\cos(2n\pi/N)}\right] \tag{5.7}$$

这个方程式只提供［$-\pi$，$+\pi$］的相位值范围。这一步也称为相位包裹，获得的相位

被称为“包裹相位”。模量为2π的包裹相位可以通过采用空间相位展开算法转换为连续相位映射$\boldsymbol{\Phi}^{\mathrm{r}}(x, y)$[32]。空间相位展开算法通过比较相邻像素的相位值定位2π不连续的相位值，通过加上或者减去2π的整数值并删除2π跳跃。

众多的相位展开算法被开发出来，其中一些在本质上是非常稳健的[33-44]。虽然在一些方面稳健，但空间相位展开算法只适用于点和点之间没有突然变化的“平滑”表面。此外，因为展开的相位总是指向映射中的包裹相位，获得的相位$\boldsymbol{\Phi}^{\mathrm{r}}(x, y)$被称为相对相位，将相位值与深度$z$的值唯一对应起来是很难的。为了唯一地确定深度$z$和相位值之间的关系，绝对相位是必要的，这将在下一节中解释[45]。

5.2.4.2　恢复绝对相位

前面提到的空间相位展开方法只为每个像素恢复相对相位，不能用于测量步高大于π或有不连续补丁的对象。为了恢复绝对相位，每个连续的补丁至少需要一个点来得到一个已知的相位值。如果传感速度是至关重要的，这些信息可以通过用标记[46]或投射一个额外的图案来编码条纹图案[47]的方式进行传输。要获得逐个像素的绝对相位，通常需要更多图像，通常采用时间相位展开算法。时间相位展开算法并非看邻近像素的相位值，而是使用来自相同摄像头像素上的其他相位值的线索。

研究人员已经开发出许多时间相位展开方法，包括两频[48]或多频[49]移相法和灰度编码加移相法[50]。用灰度编码加移相方法获得绝对相位，一系列设计的二进制编码图案独特定义每个2π相位的跳变位置来创建一个边缘序列$k(x, y)$，这样相位可以通过二进制编码图案逐个像素地恢复。简而言之，独特的码字$k(x, y)$，类似于二进制结构光方法，是为了展开相位分配给每个2π相变期的。每个码字都是由一系列二进制结构化图案建立的。一旦$k(x, y)$确定，相位就可以不用看相邻相位值而逐个像素地展开。也就是说，绝对相位可以通过以下公式获取：

$$\boldsymbol{\Phi}(x,y) = \phi(x,y) + k(x,y) \times 2\pi \tag{5.8}$$

如上所述，从一个单一频率方法得到的相位在[-π, π)的范围内。当一个条纹图案包含多个条纹，必须展开相位来得到连续相位映射。这意味着，如果另一套宽条纹图案(单条纹可以覆盖整个图像）是用于在没有2π不连续的情况下获取相位映射，第二相位映射可以在没有空间相位展开的情况下用来逐点展开另一个。要获得更广泛的条纹图案的相位，有两种方法：

1）直接使用很宽的条纹图案以使得单条纹覆盖整个测量范围。

2）使用两个高频条纹图案来生成一个等效的低频条纹图案。

前者并不常用，因为受噪声或（和）硬件限制影响，生成一个高质量的宽条纹图案是很难的。因此，后者更常被采用。本小节将简要说明该技术的原理。

这种多频相移法起源于物理光学理论，在这种方法中，理论上绝对相位$\boldsymbol{\Phi}$、光的波长λ和高$h(x, y)$之间的关系可以写为

$$\boldsymbol{\Phi} = [C \cdot h(x,y)/\lambda] \cdot 2\pi \tag{5.9}$$

式中，C是一个系统常数。所以，对于$\lambda_1 < \lambda_2$来说绝对相位是$\boldsymbol{\Phi}_1$和$\boldsymbol{\Phi}_2$，它们的区别分

别是

$$\Delta\Phi_{12} = \Phi_1 - \Phi_2 = [\,C \cdot h(x,y)/\lambda_{12}^{eq}\,] \cdot 2\pi \tag{5.10}$$

式中，

$$\pi\lambda_{12}^{eq} = \lambda_1\lambda_2/|\lambda_2 - \lambda_1| \tag{5.11}$$

是 λ_1和 λ_2之间的等效波长。如果 $\lambda_2 \in (\lambda_1, 2\lambda_1)$，我们就得出 $\lambda_{12}^{eq} > \lambda_2$。实际上，我们只有已包裹相位 Φ_1和 Φ_2。我们知道的绝对相位之间的关系是 Φ 和带有 2π 不连续性的包裹相位 $\Phi = \Phi(\mathrm{mod}\ 2\pi)$。这里模运算符用于将相位转换为 $[0, 2\pi)$ 的区间。对式（5.10）进行模数运算将导致：

$$\Delta\phi_{12} = [\,\Phi_1 - \Phi_2\,](\mathrm{mod}\ 2\pi) \tag{5.12}$$

$$= [\,\phi_1 - \phi_2\,](\mathrm{mod}\ 2\pi) \tag{5.13}$$

$\Delta\phi_{12} = \Delta\Phi_{12}$ $(\mathrm{mod}\ 2\pi)$。如果正确选择波长，则结果等效波长 λ_{12}^{eq}是大到足以覆盖图像的整个范围，即 $|C \cdot h(x,y)/\lambda_{12}^{eq}| < 1$，模运算符没有任何影响，而且不需要进行相位展开。然而，由于噪声存在，双频技术通常是不够的[48]。实际上，为了进行逐点绝对相位测量，至少需要有三个频率条纹图案。多频技术是为了等效的最宽条纹可以覆盖整个图像[51]。

假设使用另外一套含波长（λ_3）条纹图案；在 λ_1和 λ_3之间的等效波长会是 $\lambda_{13}^{eq} = \lambda_1\lambda_3/|\lambda_3 - \lambda_1|$。我们会得出

$$\Delta\phi_{13} = [\,\phi_1 - \phi_3\,](\mathrm{mod}\ 2\pi) = \{[\,C \cdot h(x,y)/\lambda_{13}^{eq}\,] \cdot 2\pi\}(\mathrm{mod}\ 2\pi) \tag{5.14}$$

$$\Delta\phi_{123} = (\Delta\phi_{13} - \Delta\phi_{12})(\mathrm{mod}\ 2\pi) = \{[\,C \cdot h(x,y)/\lambda_{123}^{eq}\,] \cdot 2\pi\}(\mathrm{mod}\ 2\pi) \tag{5.15}$$

式中，$\lambda_{123}^{eq} = \lambda_{12}^{eq}\lambda_{13}^{eq}/|\lambda_{13}^{eq} - \lambda_{12}^{eq}|$。现在我们只需要 $|C \cdot h(x,y)/\lambda_{123}^{eq}| < 1$ 来确保不展开空间相位的情况下得到绝对相位。只要得到最长的相等波长绝对相位，它就可以反过来展开其他波长的相位。最短波长的相位通常用来恢复 3D 信息，因为测量精度大约与波长成反比。

5.3　结构光系统校准

在对不同的结构光技术的论述中，我们已经讨论过，假设系统被校准，就能够将 3D 信息三角化测量。在选择一个汇编方案后，摄像头上的点可以与一个投影点（u^p，v^p）或一条线对应，而后将其与投影仪和摄像头内在和外在的矩阵用于 3D 点线性方程式解算三角化处理。结构光系统校准为摄像头和投影仪确定了这些内在和外在的矩阵。

人们已经研究出许多种结构光系统校准的方法使其达到高精度，并且结构光系统校准涉及投影仪和摄像头的校准。摄像头校准是由 Zhang（2000）[52]建立的平面棋盘法，因其简单性和标定转速，而被广泛使用。在该方法中，摄像头被视为针孔摄像头模型。摄像头校准确定其内部参数（如焦距、主点）和外部参数；坐标（x，y，z）与现实世界坐标（x^w，y^w，z^w），以及摄像头坐标系统是互相协调的。图 5.5 所示的是一个简单的针孔系统，其内在的参数可以描述为

$$A=\begin{bmatrix}\alpha & \gamma & u_0\\ 0 & \beta & v_0\\ 0 & 0 & 1\end{bmatrix} \tag{5.16}$$

式中，(u_0, v_0) 是主点坐标，或光轴和图像传感器平面的交叉点；α 和 β 是图像平面上 u、v 轴上的焦点长度；γ 是表示 u、v 坐标偏度的参数。对于现代的摄像头传感器来说，这个值通常都是零，因为 u 和 v 方向是彼此垂直的。

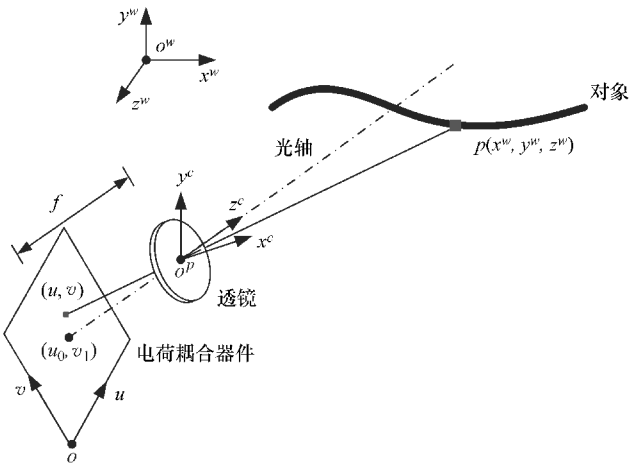


图 5.5　针孔摄像头模型。该摄像头模型描述了 3D 空间中的任意点在其坐标系下转化为了摄像头镜头坐标系，最后镜头坐标系中的 3D 坐标会被投影到 2D 成像空间。来源：Zhang & Huang PS 2006b。转载获 SPIE 许可

数学上将针孔摄像头模型的外在参数描述为

$$[R,t]=\begin{bmatrix}r_{11} & r_{12} & r_{13} & t_x\\ r_{21} & r_{22} & r_{23} & t_y\\ r_{31} & r_{32} & r_{33} & t_z\end{bmatrix} \tag{5.17}$$

式中，R 是一个 3×3 的旋转矩阵；而 t 是一个 3×1 的平移向量。

如图 5.5 所示，对于任意一个点 p，在世界坐标系（o^w；x^w，y^w，z^w）的坐标为（x^w，y^w，z^w），在摄像头坐标系（o^c；x^c，y^c，z^c）的坐标为（x^c，y^c，z^c），其在 uv 图像平面上的投影在数学上可以表示为

$$sI=A[R,t]X^w \tag{5.18}$$

式中，在图像平面上 $I=[u, v, 1]^{\mathrm{T}}$ 是图像点的齐次坐标；$X^w=[x^w, y^w, z^w, 1]^{\mathrm{T}}$ 是该点在世界坐标系中的齐次世界坐标；而 s 是一个比例因子。上述方程式描述了一个线性摄像头模型，非线性效应可以通过采用非线性模型补偿。为简单起见，本章只阐述了线性模型。

结构光系统与立体声系统不同，因为该系统中用投影仪取代了其中的一个摄像头。这种替换使得结构光系统校准非常困难，因为投影仪不能像摄像头那样捕捉图像。人们研究出了各种技术以达到完全校准结构光系统[7,53]，但这些方法大多数是非常耗时的，并且难以达

到高精度。2004 年 Legarda – S′aenz 等学者（2004）[54] 提出一种利用绝对相位的方法，通过投影一系列的条纹图案来找到投影仪校准板的标记中心。通过优化，该方法在精度方面表现良好。然而，它需要使用校准的摄像头来校准投影仪，因此，该摄像头的校准误差将被耦合到投影仪校准，这是不可取的。

从光学角度来看，投影仪和摄像头是相同的。鉴于此，Zhang 和 Huang（2006b）[47] 提出了一种新的结构光系统校准方法，同时独立地校准投影仪和摄像头。在这种方法中，水平和垂直条纹图案被用来建立一个摄像头像素和投影仪像素之间的一对一的映射。这使得摄像头的灰度图实际上为投影仪生成图像，因此，该方法允许投影仪像普通摄像头一样“捕捉”图像。

一旦投影仪“捕捉”到图像，结构光系统校准会成为一个行之有效的立体声系统校准。由于投影仪和摄像头的校准是同时独立地进行，校准的精度大大地提高了，同时校准速度也大大增加了。图 5.6 所示的是摄像头捕捉到的一个典型的棋盘图像对和映射方法转换的投影图像。它清楚地表明，投影仪的棋盘图像捕捉得很完整。继 Zhang 和 Huang[47] 的研究后，还有人研究得出了一些校准方法[55-58]。这些技术的主要目标基本上是相同的：在投影仪和摄像头之间建立一个一一映射。一旦系统被校准，可以使用绝对相位作为一个约束来计算 (x, y, z) 坐标，我们会在下面对此进行讨论。

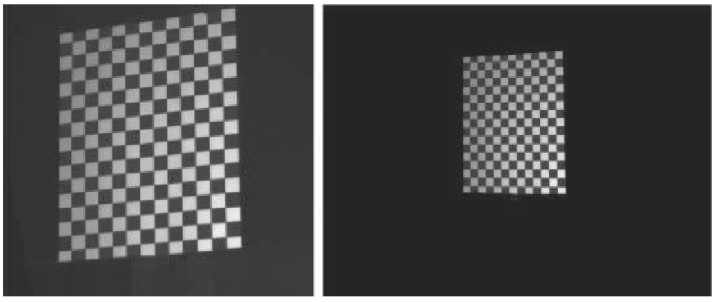


图 5.6　Zhang 和 Huang [47] 运用该技术得到的棋盘图像对。
a）摄像头拍摄到的棋盘图像；b）投影仪映射的棋盘图像，也被认为是通过投影仪捕捉的棋盘图像。来源：Zhang S and Huang PS 2006b。转载获得 SPIE 许可

绝对相位图提供了一个摄像头像素和投影线之间的一对一映射。如果摄像头和投影仪在同一个世界坐标系中校准，则这个约束就足以获得唯一的 (x, y, z) 坐标。因为对于一个结构光系统，式（5.18）可以被重新编写来代表摄像头针孔模型。

$$s^c I^c = A^c [R^c, t^c] X^w \tag{5.19}$$

式中，s^c 是摄像头的缩放因子；I^c 是齐次摄像头图像坐标；A^c 是摄像头的内在参数；$[R^c, t^c]$ 是摄像头的外在参数矩阵。这就提供了一个从坐标系到摄像头图像平面的映射。同样，从世界坐标系到投影仪图像平面的投影可以表示为

$$s^p I^p = A^p [R^p, t^p] X^w \tag{5.20}$$

式中，s^p 是投影仪的缩放因子；I^p 是齐次投影图像坐标；A^p 是投影仪的内在参数；$[R^p, t^p]$ 是投影仪的外在参数矩阵。

从式（5.19）～式（5.21）有 6 个方程式，但有 7 个未知数（x^w，y^w，z^w）、s^p、s^p、u^p、v^p。为了完全地解出世界坐标（x^w，y^w，z^w），需要有由绝对相位信息提供的另外一个方程式（或约束）：摄像头上的每个点与投影面上相同的绝对相位的一条线相对应[47]。即假设条纹是沿着 v 方向，我们可以在捕获的条纹图像和投影的条纹图像之间建立关系：

$$\phi_a(u^c, v^c) = \phi_a(u^p) \tag{5.21}$$

有了该约束方程式，（x^w，y^w，z^w）坐标可逐个像素地唯一解出[59]。

5.4　数字条纹投射（DFP）技术下的 3D 传感示例

本节展示了几则使用 DEP 技术实现高分辨率 3D 传感的实例。图 5.7 阐述的是使用三步相移测量法实现 3D 传感的实例，图 5.7a～c 所示为伴随 2π/3 相移的三相移条纹图像。图 5.7d 对这些条纹图像，应用式（5.7）后的相位图，该图清楚地显示了相位的不连续性。应用参考文献［60］中讨论的相位去包裹算法，包裹相位图可以去包裹化以得到一个连续的相位图，如图 5.7e 所示。之后，去包裹相位图可通过应用 5.3 节中所介绍的方法换成 3D 形状。3D 形状可使用 3D 图形处理库（OpenGL）得以进一步绘制，如图 5.7f、g 所示。同时，通过对这三幅条纹图像取均值的方法，得到纹理图像。而纹理图像可映射于 3D 形状上，以获得更逼真的视觉效果，如图 5.7h 所示。

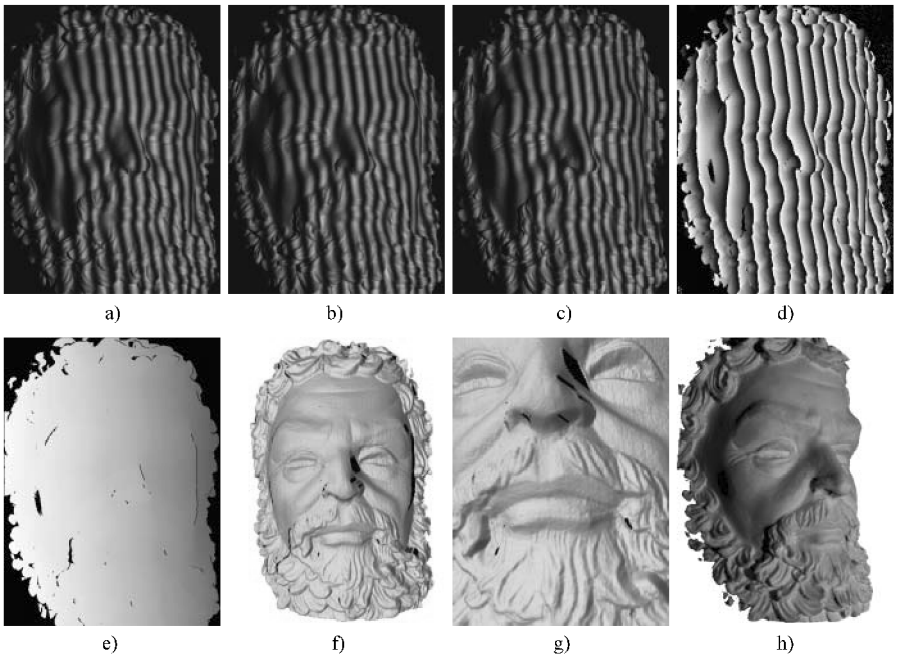


图 5.7　采用三步相移法实现 3D 传感的实例。a）$I_1(-2\pi/3)$；b）$I_2(0)$；c）$I_3(2\pi/3)$；d）包裹相位图；e）去包裹相位图；f）阴影模型绘制的 3D 形状；g）镜头拉近后的图像；h）纹理映射绘制 3D 形状。来源：Zhang S and Huang PS 2006b。转载获取 SPIE 许可

图 5.8 和图 5.9 阐述了多频相移法[61]的 3D 传感实例，我们选择了 $\lambda_1 = 60$ 像素、$\lambda_2 = 90$ 像素和 $\lambda_3 = 102$ 像素的三幅频率条纹图像，结果显示得到的等效条纹波长为 765 像素，换言之，如果我们使用投影仪产生 765 像素宽条纹图像，就不需要为了恢复绝对相位而对空间相位去包裹，于是便形成了 3D 形状。

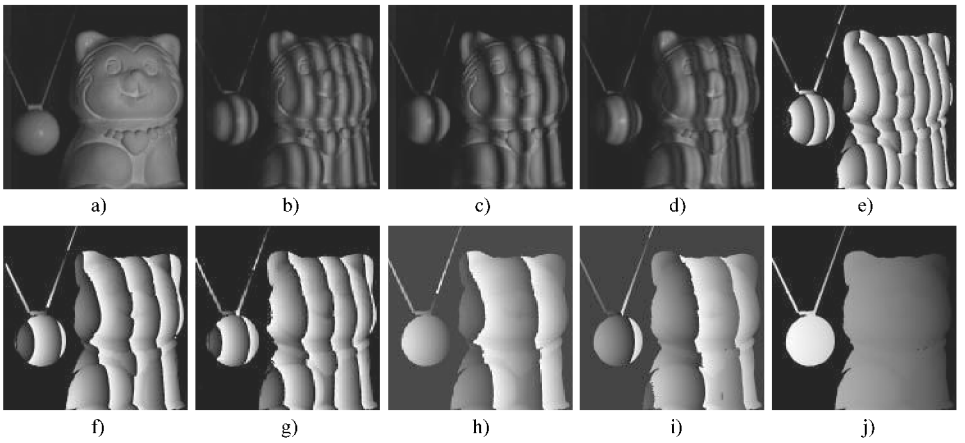


图 5.8　a）捕捉到场景的照片；b）一个条纹图案（$\lambda_1 = 60$ 像素）；c）一个条纹图案（$\lambda_2 = 90$ 像素）；d）一个条纹图案（$\lambda_3 = 102$ 像素）；e）包裹相位 ϕ_1；f）包裹相位 ϕ_2；g）包裹相位 ϕ_3；h）相应等效相位差 $\Delta\phi_{12}$；i）相应等效相位差 $\Delta\phi_{13}$；j）产生的相位 $\Delta\phi_{123}$。来源：Wang Y and Zhang S 2011。转载获得美国光学协会许可

图 5.9　使用绝对相位图得到的 3D 结果。来源：Wang Y and Zhang S 2011，转载经美国光学学会许可

如图 5.8a 所示，该实验中存在两个独立的物体，透过摄像头观看球状物的直径大约为 17mm，塑像大约高 60mm，宽 45mm。图 5.8b ~ d 显示了不同频率下的变形条纹图案。图 5.8e ~ g 分别展示了不同频率的变形条纹图案中获取的包裹相位图。图 5.8h 显示的是 λ_1 和 λ_2 对应的等效包裹相位图，而图 5.8i 显示的是 λ_1 和 λ_3 对应的等效包裹相位图。

最终，最长等效波长的相位图可通过两幅等效相位图获得，该结果如图 5.8j 所示。从图中可见，最长等效波长相位图，无 2π 不连续性，因而，不需要空间相位的去包裹化。

不同于最长等效波长的相位图，最短波长（$\lambda_1 = 60$ 像素）可以逐点去包裹，之后再用于重建 3D 信息。图 5.9 显示了重建后的 3D 结果。图中可见，3D 外形得以适当恢复，而这种效果依靠单频移相方法是无法实现的。

5.5　实时 3D 传感技术

我们可以采用随机汇编的单一结构化图案或正弦结构以满足实时的速度要求，然而，这些技术通常在表面性能需求或可达到的空间分辨率上具有很大的局限性，权衡之下，人们往往会采用快速切换的多结构化图案，以便在短时间内捕获适用于恢复 3D 形状的结构化图案。Rusinkiewicz 等人（2001）[62]研发了一种利用条纹边界编码实现实时 3D 模型采集的系统[63]。条纹边界编码是由投射二进制级结构化图案的序列决定的，如上所述，该技术的空间分辨率受投影仪分辨率大小的限制。

为避免由彩色、单色或黑白色引起的问题，通常通过使用不同的结构化图案寻找解决办法。例如，Zhang 和 Huang（2006a）[47]研发出一种以黑白正弦变化的结构化图案为基础的 3D 传感系统，Zhang 等人（2006a）通过使用三步相移算法[69]开发了一个实现实时同步数据采集、重建和显示的运行系统，该方法利用了单片机数字光处理（DLP）的独特投影机制。三种结构化图案编码进入投影仪的 RGB 通道，并顺其自然地由数字光处理（DLP）投影仪在三种图案中自动切换。

通过这些手段，我们实现了在 60Hz 的波动频率下，以每帧超过 300k 点的速度进行了 3D 表面测量[71]，我们将在本节进一步阐述该项技术的细节。

5.5.1　数字光处理（DLP）技术的原理

数字光处理（DLP）的概念最初产生于 20 世纪 80 年代的美国德州仪器公司。1996 年，德州仪器公司开始利用其数字光处理技术牟利。每一个 DLP 投影系统的核心部件都是一个光学半导体，称作数字微镜装置（DMD）。DMD 实则是一个非常精准的光开关。DMD 芯片包含一个铰链式相连的微镜阵列，每一个微镜均与投影图像上光的一个像素对应。

图 5.10 显示了微镜的工作原理。光学组件上的数据控制的静电力使得反射面在 $+\theta_L$（开）~ $-\theta_L$（关）之间移动，从而调节投射于反射面上的光线，反射面开关的速度决定了投影图像像素的亮度。图像是光从“开”反射面经由投影透镜反射到屏幕上而形成的。灰度值是通过控制在帧周期内，反射面开关时间的比例来创建的——黑即 0% 开时间，而白即

100% 开时间。

数字光处理（DLP）投影仪采用数字微镜器件（DMD）生成彩色图像。所有 DLP 投影仪都包括一个光源、一个彩色滤光系统、至少一个 DMD、数字光处理电子器件和一个光学投影镜头。对于单片数字光处理（DLP）投影仪，彩色图像是由系统中放置的色轮制作的。包含红、绿和蓝色滤光器的色轮高速旋转——因此，红、绿和蓝色通道图像会被投影到屏幕上，然而，由于刷新率很高，人眼只能捕捉到一个彩色图像而非相继出现的三个图像。

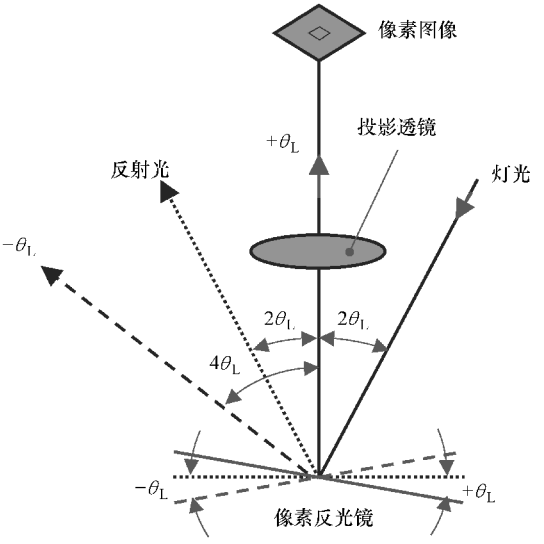


图 5.10　数字微镜装置的光学开关原理
（经允许修改自参考文献［72］）

数字光处理（DLP）投影仪由于时域积分产生灰度值[73]，我们用普乐士 U5 - 632h 单片数字光处理投影仪以投影速度为 120Hz 的单色模式进行简单测验。光敏二极管（Thorlabs FDS100）感知输出的光，光电流转化为电压信号，全过程由示波器监测。所使用的光敏二极管响应时间为 10ms，有效面积为 3.6mm × 3.6mm，带宽 35MHz。示波器 Tektronix TDS2024B 用于监测信号，所用带宽为 200MHz。

图 5.11 展现的是投影机被馈以不同灰度值的均匀图像后的典型结果。如果馈以纯绿色 RGB =（0，255，0），则信号的占空比会接近 100% “开”。当灰度值下降到 128，大约一半的通道被填充。如果输入的灰度值减小到 64，则该通道只有一小部分被填充。如果输入的灰度值是 0，那占空比将接近于 0% “开”。这些实验表明，如果馈以的灰度值介于 0 ~ 255 之间，输出信号变得不规则。因此，馈以的灰度值从 0 到 255，正弦条纹也会相应变化，整个投影周期必须得以捕获，以便获取从投影仪投影得来的图像，这就是我们使用实时 3D 传感技术[74]真正意义之所在。

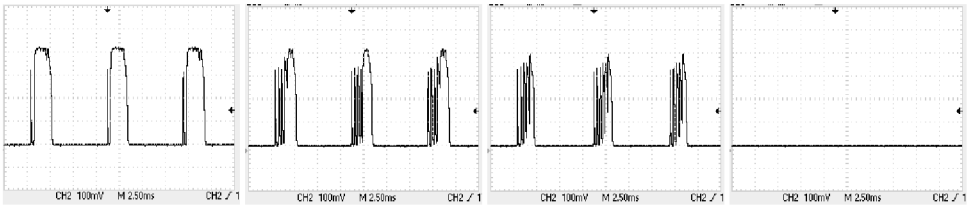


图 5.11　投影仪中绿色图像被馈以不同灰度值后所获得的投影时间信号示例。a）绿 = 255；b）绿 = 128；c）绿 = 64；d）绿 = 0。改编自 Zhang. S. , et al. 2013

5.5.2　实时 3D 数据采集

如 5.2.4 节所述，如果使用三步相移算法，三幅结构化图像可用于重建一个 3D 形状。这与数字光处理（DLP）技术完全吻合。数字光处理（DLP）技术中三种图案被编入投影仪的三基色通道中。因为彩色条纹图案在 3D 传感中不可或缺，我们基于单片数字光处理（DLP）投影仪和白光技术研发了一种实时 3D 传感系统[74]。图 5.12 展示了该系统的实际规划，计算机生成的彩色编码条纹图像发送至单片数字光处理投影仪，从而顺序且反复地将灰阶的三色通道投射到物体上，摄像头和投影仪完全同步，这样，摄像头就可以单独快速地捕捉每一个单独通道。对三幅条纹图像采用三步相移算法，3D 几何图形得以复原。取三幅条纹图像的平均值，即产生纹理图像，而纹理图像可进一步映射到复原的 3D 形状上，以加强其视觉效果。

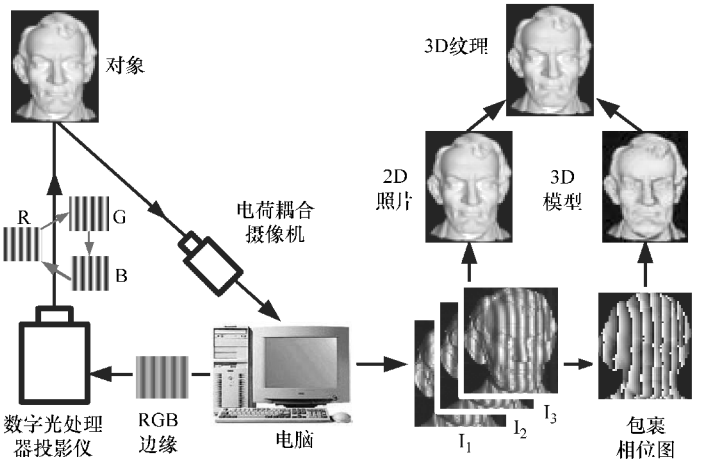


图 5.12　实时 3D 形状测量系统布局。来源：Zhang S 2010。转载获 Elsevier 许可

投影仪顺序为每一 RGB 通道投射生成单色条纹图像，颜色的产生是置于投影镜头前的色轮作用的结果。投影图像的每一个“帧”实际上都包含三个独立图像。通过移除色轮和在每一单独通道中放置单独的条纹图像，都可使投影仪以 120 帧/s 的速度生成三幅条纹图像（每个颜色通道的刷新率为 360 帧/s），因此，如果三幅条纹图像用于复原一个 3D 形状，那么 3D 测量的速度就要在 120Hz 以上，然而，由于摄像头的速度有限，所以一般摄像头都会使用两个投影周期来捕捉这三幅条纹图像，这样，测量速度就可以降至 60Hz。

图 5.13 是实时 3D 形状测量系统的时序图。由于我们使用的摄像头的速度限制（全分辨率时，最高速度为 200 帧/s），摄像头往往需要两个投影周期来捕获用于复原一个 3D 形状的三幅结构化图像，这样，我们就实现了以 60Hz 的速度来测量 3D 形状，这种方式的效果比实时（通常为 24Hz 或更高）更快。

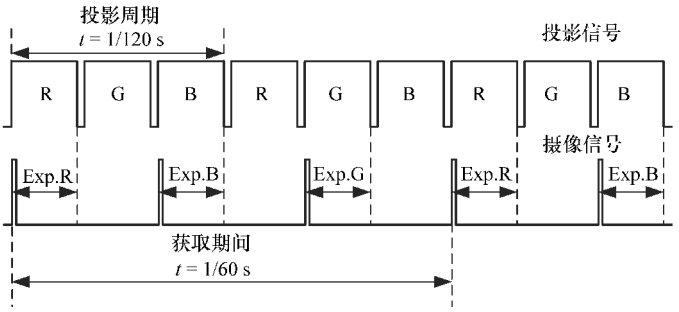


图 5.13　测量系统的时序图。来源：Zhang S and Yau ST 2007b。转载获 SPIE 许可

5.5.3　实时 3D 数据处理与可视化

根据相位计算 3D 坐标是计算密集型的，单台计算机处理器想要实现实时 3D 坐标计算非常具有挑战性。然而，坐标计算是点对点的矩阵计算，图形处理单元（GPU）则可以有效处理该计算过程。GPU 是用于个人电脑或游戏控制台的专用图形渲染器。现代图形处理单元（GPU）在处理和显示电脑图形图像方向非常有效，其高度并行结构较典型的 CPU（中央处理器）更适应并行算法。由于传输模型中没有分级存储器件或数据依赖关系，该传递途径将吞吐量最大化而无任何停滞，因此，不管图形处理单元何时被连贯地馈以输入数据，它的性能都很高，优良的可扩展结构就此形成[76]。现代图像处理单元的这种流媒体处理模型在某些通用应用程序方面超越了中央处理器（CPU），且这种优势性差异在未来可能仍会增大[77]。

图 5.14 所示为 GPU 传递途径，CPU 发送顶点数据包括顶点位置坐标和顶点法坐标至 GPU。GPU 生成各顶点的光照，创建多边形、光栅处理像素，然后向显示屏输出光栅处理后的图像。现代 GPU 允许用户执行传递途径中的顶点和像素部分指定代码，这些部分分别被称作顶点着色器和像素着色器，可编程顶点处理器上的顶点着色器适用于每一个顶点。

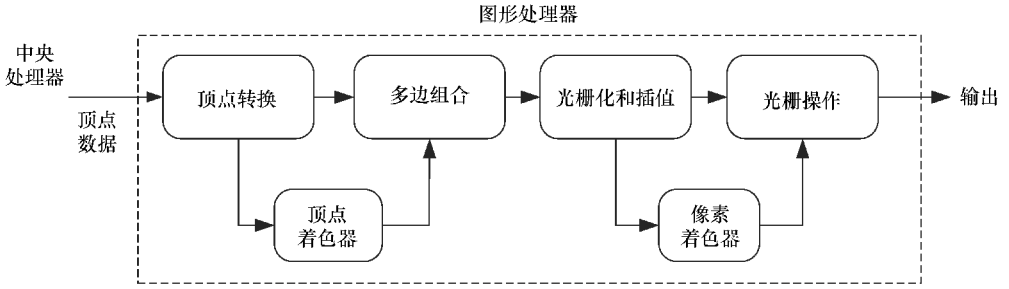


图 5.14　GPU 传递途径。顶点数据包括顶点坐标和顶点法坐标，被送往 GPU，GPU 生成各顶点的光照，创建多边形、光栅处理像素，然后向显示屏输出光栅处理后的图像。
源自：Zhang S，Royer D，and Yau ST，2006。转载获美国光学学会许可

顶点着色器承载着顶点坐标、颜色，以及从 CPU 获取的普通信息，顶点数据流入 GPU，基于数据的输入顺序在多边形的顶点被处理和组装。GPU 自动处理流数据转移程序，以便进行下一步的并行计算。虽然 GPU 的时钟速率可能明显慢于 GPU，但 GPU 的多顶点处理器可并行运行，因此，GPU 的吞吐量可超过 CPU，随着 GPU 带来了复杂性的同时，顶点处理器的数量也会增加，这样 GPU 的性能也会有更大改进。

通过利用 GPU 的处理能力，3D 坐标计算在使用配有 NVIDIA 显卡的普通个人电脑的情况下即可实时运行[79]。此外，3D 形状数据早已存入显卡，因而可以没有任何滞后地实时递交。因此，通过这些方式，实时 3D 几何可视化也可同步实时完成，同时，因为只有相位数据，而无 3D 坐标和法坐标，作为可视化数据传递到显卡，该技术显著地降低了数据传输的负荷（几乎降低为 1/6）。总之，利用 GPU 的处理能力进行 3D 坐标校准，实时 3D 图像重建和可视化，既实时也快速。

图 5.15　30 帧/s 下的同步 3D 数据采集、重建及显示。来源：Zhang S 2010。转载获 Elsevier 许可

5.5.4　实时 3D 传感实例

图 5.15 显示了一个活生生的人脸测量的实验结果，右图展示了实际对象，而左图展现了所获得的 3D 几何图形，并同时呈现在了电脑屏幕上。同步 3D 数据采集、重建和显示速度达到了 30 帧/s，且每帧可获得超过 300000 点数据。

5.6　人机交互应用的实时 3D 传感

实时 3D 传感融合了许多高新技术以使其具有在现实世界中快速精准地捕捉物体的能力。这允许软件进行准确的测量，识别预设的图案并作运行，以及通过感知的数据控制系统，等等。从该技术所在的人机交互领域中获得启示的方式是提出这样的问题：像这样的系统究竟在做什么呢？

观察诸如结构光系统的 3D 系统。通过在实时 3D 中观察，许多崭新的与电脑交互的动态响应关系就能够被发现。随着 3D 计算机视觉领域的设备和软件的发展，用户再也无需被固定在桌子后，用鼠标、键盘或其他传统输入装置操控着一个二维平面（他们的电脑工作空间）。新的机器输入方法使更具创意、更为自然的交互成为可能，而这些在数年前还是不可行的。本节列出了这些新交互方式的实例并聚焦于该类交互的细节和启示。

5.6.1　实时 3D 面部表情捕捉及其人机交互的意义

3D 传感提供的一个新交互工具就是能够捕捉并回应面部表情与其他面部动作。比如，设想在运算系统中各个面部表情都有与之对应的关联行为。目前，虽然该技术以前就一直在研究[80]，但爱荷华州立大学正在研发的结构光技术，正如前面所述，可以提供更多视觉和空间的信息。这反过来为准确的操控系统提供了可能。

传统的电脑视觉中，要让电脑识别面部特征和表情，系统会处理一个 2D 视频流。然后基于每个视频帧可识别的特征，运行的算法可以基于预设和/或习得的图案确定用户使用了哪一个表情[80]。依据同样的原则，设想一下结构光捕捉系统可能探测到的互动和细微的脸部特征，如前面所述。这种系统提供了另一种维度的数据——深度信息，它很难用传统的 2D 捕捉。结构光系统的另一个优势是能够获得高质量和细节密集的空间坐标。

实时捕捉系统再也不需要等到面部表情变得异常清晰时才能对其进行解码了。有了今天的结构光技术，无论是一个断续的假笑还是一次眼角的抽搐都能够被清晰的捕捉，如图 5.16 所示。现在就可以思考一下该技术对残障人士的启示——他们在使用传统的鼠标、键盘或麦克风作为输入设备时可能比较困难[81]。结构光 3D 捕捉系统为与电脑交互提供了另一种可能的方式——通过捕捉和处理 3D 物体与景象。

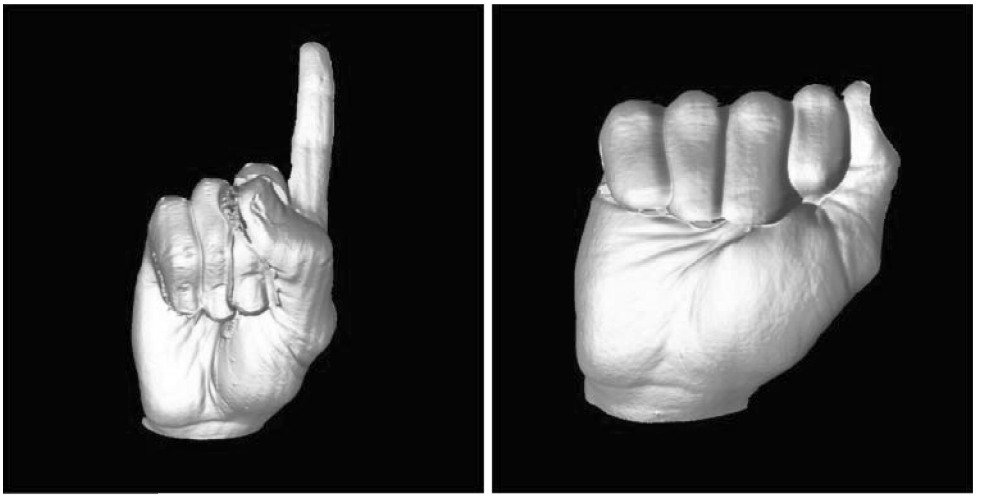
a)　　b)

图 5.16　由 3D 结构光系统捕捉的精细的面部刻画和特征的示例。a）怪笑的细节可以被捕捉；b）虽然是隐蔽的特征，但皱纹可以被轻易地探测到在额头上

5.6.2　实时 3D 身体部分姿势捕捉及其人机交互的意义

该项技术不仅可以应用于捕捉和处理非常细致的面部特征和动作，它还能应用于身体的其他各部位。以用手指指向捕捉系统为例。用传统的 2D 捕捉方法，一个直接竖起手指的动作往往因为缺乏一个清晰可辨的轮廓而很难被识别。但是如果有细节丰富的 3D 技术，指示

的手指就忽然可见了，因为它有了深度和在现实世界内的情景。为了应用于更多的姿势和交互，这些概念可以扩展到整个手、手臂和身体。图 5.17 展示了一个手指从竖起变化成握拳的例子。结构光捕捉的细节程度是很高的，因此可以破译小幅度的运动和不明显的特征。



a)　　b)

图 5.17　由 3D 结构光系统捕捉的手势位置和变化的示例；隐蔽的细节捕捉可以应用到身体的任何部位。a）竖起的手指；b）把竖起的手指向内收回

不久以前，用动作控制电脑还只是一个在科幻片中让人兴奋的概念，而且往往需要侵入式的接触型传感器才能实现；恰是诸如实时 3D 传感这样的新技术的进步才使概念变成了现实。如今，身体姿势识别已经运用到一些消费电子产品中，但我们也要考虑以下它们的某些缺点。

缺乏高质量的被感知数据和足够的可信赖空间信息是如今 2D 和 3D 捕捉系统的主要局限。虽然如此，控制这些系统的姿势仍然可以包括一只或两只手臂完全划过身体前方的动作。这个姿势可以足够明显地被识别为一个控制命令。大多数的用户能够做出这些手势吗[81]？这些姿势在久做之后会使用户疲劳吗[82]？在一些没有足够空间的环境中怎么给出这些姿势命令呢？

当更高级精准的信息由 3D 结构光系统提供时，可能的姿势数量急剧增加了，随之增加的还有姿势控制的信赖度。如前面所述，由于这些系统能够捕捉和处理十分隐蔽的面部特征和姿势，它们同样也能作用于身体各部位。轻微的手指和手控制可以在 3D 空间内执行控制电脑的动作。能够执行小幅度动作和现有的大幅度动作将使人们与电脑系统的交互变得更加自然。因为动作指令可以由身体发出，潜在用户在学习和使用这些控制时就不会遇到太多的难题[83]。

5.6.3　人机交互意义的总结

人机交互的意义是重大的，因为这些系统为用户提供了与他们接触的运算环境和服务交互的另一种方式。数据的新维度，连同更高的分辨率使基于姿势的输入设备提高了控制水平。除了前面所举的例子，还有许多其他的可以受益于该技术的人机交互领域。智慧的人机

交互践行者将驾驭这项快速发展的 3D 传感技术并创造出下一波人机交互的新理想、新控制和新技术。

5.7　最新发展

5.7.1　实时 3D 传感与自然 2D 彩色纹理捕捉

虽然实时 3D 传感已经在市场普及，但是现有实时 3D 系统仍然主要使用可见光。虽然使用成功，但它们在应用中仍有局限，诸如在人机交互、国土安全和生物计量学方面，可见光的弊端较大。为了解决这一问题，微软 Kinect 已经使用了近红外（NIR）激光来代替白光，但可能会对眼部造成伤害。随着近期 LED 技术的突破，NIR LED 光为此提供了用眼安全的解决方案。

目前关于使用条纹投影的红外高清实时 3D 传感技术的现有文献十分稀少。而且通过分析 3D 重构使用的结构模型，前述的实时 3D 传感技术能同时提供一个黑白（b/w）纹理。但是该纹理生成方法并非自然，意味着如果没有定向投射光，纹理就不能被捕捉到。这常常会因为已测量的几何图形而导致阴影投射在获得的纹理上。为了获得自然纹理的图像，投射光必须关掉，保证在纹理图像被捕捉时仅有环境光照射。这可以通过在没有投射光时捕捉额外的图像来完成。不过这会极大地降低测量速度。

想要同时获取 2D 彩色纹理更难，当然并不是不可行——用彩色相机照下 3D 几何图形和 2D 彩色纹理[60,67,84-86]。另一种做法是安装另一台专门用来捕捉彩色纹理的彩色相机[74,87,88]，建立彩色相机与黑白相机的映射。前者往往会因为内在的颜色问题（如成色显色）而影响 3D 测量的质量。后者通常要求安装复杂的硬件或苛刻的校准以实现两种相机的映射。

就我们所知，现在还没有系统可以同时实时地捕捉自然 2D 彩色纹理和高清 3D 几何图形。最近，Ou 等学者（2013）[89]已经证明了现有的 NIR 算法确实可以应用于 3D 传感。在本研究中，我们使用了近红外相机/投影仪的组合来执行 3D 传感以及另一个彩色相机来同时捕捉仅在环境光照明下的 2D 彩色图像。由于这两束光的波长并不覆盖，3D 形状和自然 2D 纹理的图像可同时获得，速度也未受影响；由于彩色相机仅捕捉可见光而不会被用来测量 3D 形状的红外光干涉，自然的 2D 彩色图像也可以获得。

图 5.18 所示为我们开发的系统的照片。它是由一个红外数码光处理（DLP）投影仪（LightCommander，Logic PD 公司）、一个高速红外 CMOS 相机（Phantom V9.1，Vision Research 公司）和一个彩色 CCD 相机（DFK 21BU04，Imaging Source 公司）组成的。应用在该投影仪的红外 LED 的波长是 850nm。在高速红外 CMOS 相机的前端装置了红外滤光器以阻挡可见光。红外 CMOS 相机可以拍摄分辨率为 576×576 的照片，彩色 CCD 相机的分辨率是 640×680。由于红外投影仪的低密度，高速投影仪设置成以 200Hz 的频率投射二进制图案，并且红外相机精准地与投影仪同步（200Hz）以拍摄 2D 条纹图案。3D 拍摄速度是 20Hz，因为它需要 10 个移相条纹图案来重构一个 3D 帧。因此彩色相机调制在 20Hz 来拍摄 2D 彩

色图像，以期精准地与 3D 拍摄速度匹配。

彩色相机和红外相机通过外部的触发定时电路来实现精准同步。图 5. 19a 展示了人脸的 3D 形状。注意到重构的 3D 几何图形并不是非常平滑（脸上有凹凸）。这可能是由于红外光射入脸部皮肤的深浅不同而导致的。从这些条纹图像中，黑白纹理可以重获并映射到 3D 几何图形上，如图 5. 19b 所示。由于投影仪的直接照明（鼻侧的阴影很明显），黑白并不自然。相反，彩色纹理在环境光下被捕捉到了，且无阴影问题显示。图 5. 19c 展示了将自然彩色纹理映射到红外 3D 几何图形上的结果。值得关注的很重要一点是，由于相机的离散效果，映射往往不会与像素实现完美对齐。该研究采用了线性内插法来投射彩色纹理。同样值得注意的是，研究使用的映射是线性的，没有考虑摄像头的非线性扭曲。但即使是线性模型也并未产生明显的伪影。

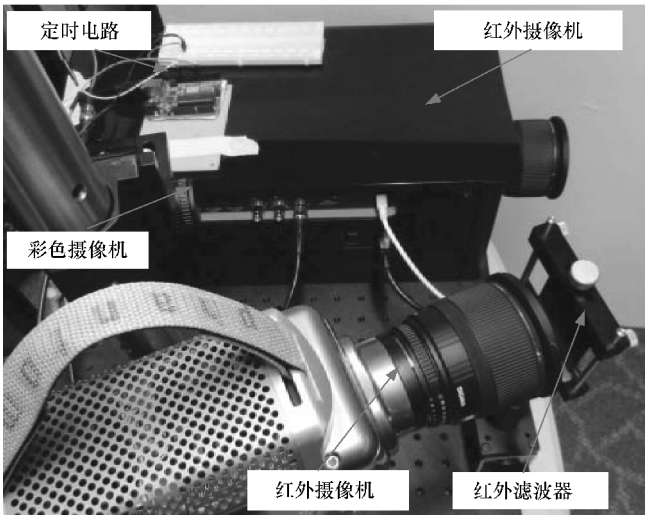


图 5. 18　同步 3D 几何图形和自然彩色纹理捕捉的系统开发。

来源：Ou P，Li B，Wang Y and Zhang S 2013。转载获得美国光学学会许可

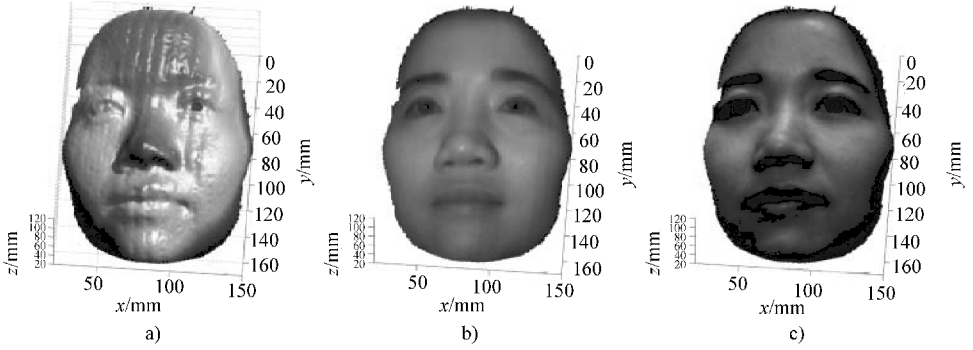


图 5. 19　实时捕捉 3D 几何图形与自然彩色纹理的实验结果。

a）使用红外条纹图案重构 3D 脸部；b）黑白纹理投影的 3D 结果；c）彩色纹理投影的 3D 结果。

来源：Ou P，Li B，Wang Y and Zhang S 2013。转载获得美国光学学会许可

5.7.2　超高速 3D 传感

实时 3D 传感技术已经应用于医学、娱乐、计算机科学和工程学的多种应用中。实时速度通常指 30Hz 或更快的速度，足够捕捉到缓慢的动作（如面部表情）。但是，要捕捉较高速的动作，如心脏跳动甚至是说话中嘴形的变化，当前的技术就有局限了。速度的突破需要能够快速抓拍变换的情景。我们最近的研究就一直在开发二进制散焦法以解决限速的问题[12]。

利用 DLP Discovery 平台，Zhang 等学者（2010）[13]已经成功研制了一个可以实现数万赫兹的 3D 传感速度。但是二进制散焦技术并非完美无瑕。相比传统的 DFP 方法，其测量能力非常有限：

1）测量的准确性由于高频谐波的影响而较低。

2）测量的深度范围相对有限，因为物体必须放置在较小的区域，使二进制图案变成质量好的正弦图案[45]。

3）需要更频繁的校准，因为大多数现有的校准技术需要投影仪对焦[90]。

4）对不同的空间频率很难同时实现高质量的条纹图案[61]。

为了改进二进制散焦法，在电子电力学领域开发的脉宽调制（PWM）技术被运用到 3D 传感领域。Ayubi 等学者（2010）[91]已经引入了正弦脉宽调制（SPWM）技术，Wang 和 Zhang（2010）[92]已经开发了最优脉宽调制（OPWM）技术以进一步改进相位质量。但如果条纹线过宽或过窄，OPWM 和 SPWM 技术仍旧面临挑战[93]。

由于条纹图案的离散本质，PWM 技术的改进局限是可以理解的。这是因为 PWM 技术毕竟本质上还是一维的。因此，如果优化可以双维度展开，未来的改进空间仍然很大。Lohry 和 Zhang（2012）[94]最近提出可以局部调制像素以模仿三角图案的技术，以期减少高频谐波的影响。

就在最近，我们注意到在数字图像处理领域的高质量打印方面，以二进制图像代表灰度图像的技术已经相对成熟。该技术称为半调或抖动，自 20 世纪 60 年代以来一直广为应用[95]。Wang 和 Zhang（2012b）[96]从简单的 Bayer 抖动技术和后来的误差扩散技术[97-101]中借用了概念并将其应用在 3D 形状测量领域[102]。我们的研究发现抖动技术在条纹线较宽时能极大改进测量质量；若条纹线较窄，则改进作用并不显著。

大多数的抖动技术不过是在灰度图像中应用了一个矩阵，通过将原始或改进的灰度与矩阵比较，使图像得以“二进制化”。这些算法都被开发用于为一般灰度图像生成高质量的视觉效果。但是它们起初不是为利用某些灰度图像的内在结构而专门设计的，如条纹图案的正弦结构。因此，如果应用独特的条纹图案的正弦结构，应该就可以有巨大的改进技术效果。近期我们已经开发了一些算法[103,104]来优化抖动技术，取得了实质性的突破。

我们测量了条纹周期在 $T=90$ 像素的 3D 雕塑，让投影仪稍微地散焦。图 5.20 展示了结果。第一行表现了捕捉的结构化图像，第二行展示了 3D 结果。很明显地可以看到，如果投影仪几乎对焦，正方的二进制图案的二进制结构是清晰的，但是抖动或优化抖动图案的本质是正弦的。该实验也显示了正方二进制法（Squared Binary Method，SBM）和 PWM 都无法生

成高清 3D 结果，虽然 PWM 稍微促进了 SBM，抖动技术和优化抖动技术在重构雕塑的相位时表现良好。优化抖动技术较抖动技术而言有相当明显的改良。

Wang 等学者（2013）[102]开发了测量活兔心跳的系统，测出心跳速度约在每分钟 200 下。该系统由数字光处理（DLP）投影仪（DLP Light Crafter，德州仪器公司）和高速 CMOS 相机（Phantom V9.1，Vision Research 公司）组成。相机由能够感应 DLP 投影仪定时信号的外部电路激活。相机用来拍摄 576 ×576 分辨率的图像，投影仪的分辨率在 648 ×648。

投影仪被设定为以 2000Hz 频率转换二进制图案，相机设定为 2000Hz 频率进行图像捕捉。一个双波长的相移技术被选用，短的波长是 OPWM 图案[92]，长的波长是 Stucki 抖动图案[101]。

图 5.21 所示为兔子心脏的典型特征，并清晰地显示了心脏的动态运动被很好地拍摄了下来。没有二进制散焦技术，跳动的兔子的心脏表面就无法正确地测量，因为我们发现至少需要 800Hz 的频率来正确测量心脏而没有明显的运动伪影。

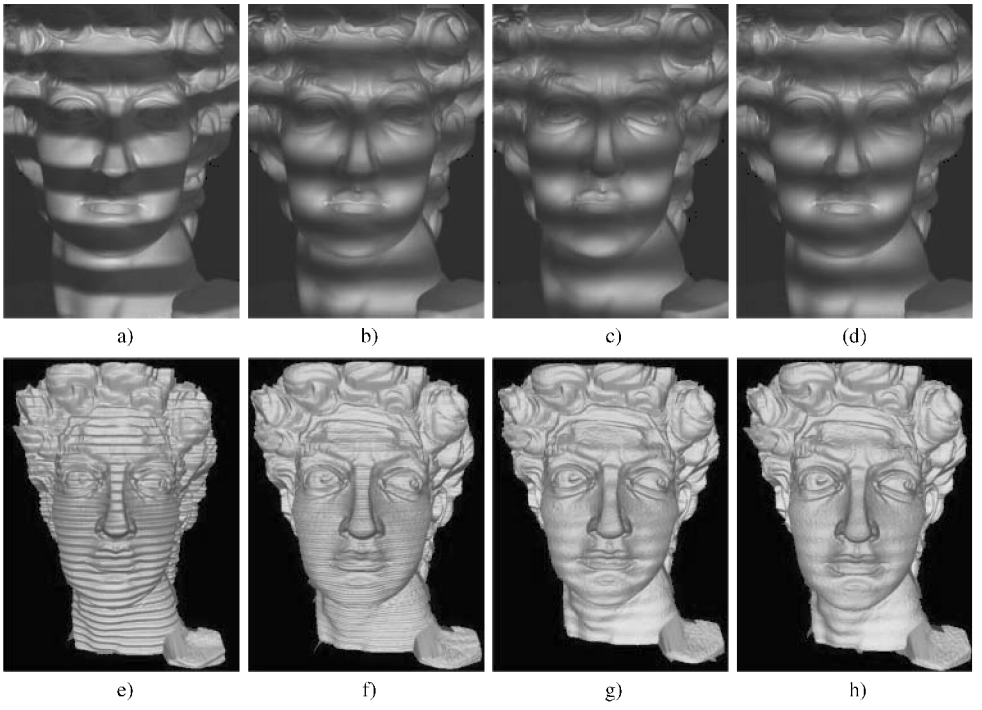


图 5.20　在投影仪稍微散焦时，捕捉图案的条纹周期 $T=90$ 像素。
a）正方二进制图案；b）PWM 图案；c）抖动图案；d）优化抖动图案；
e）～f）对应重构 a）～d）的 3D 结果。
来源：Li B，Wang Y，Dai J，Lohry W and Zhang S 2013。转载获得 Elsevier 许可

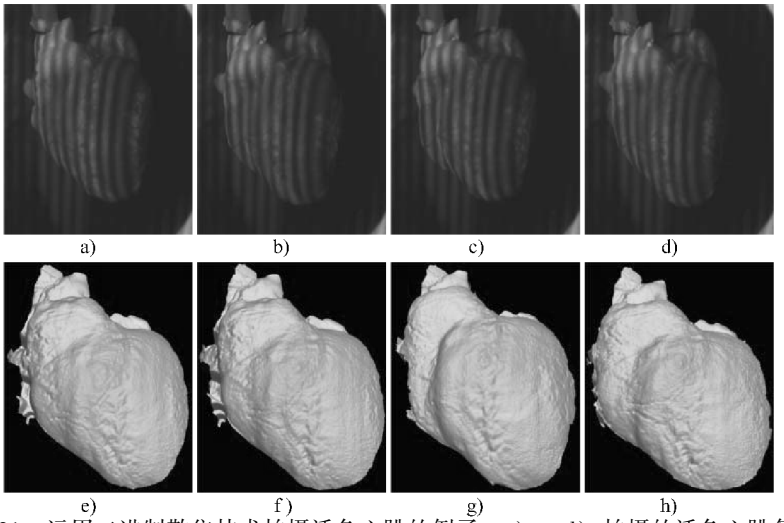


图 5.21　运用二进制散焦技术拍摄活兔心跳的例子。a）～d）拍摄的活兔心跳条纹；e）～h）相应的 a）～d）3D 结果的重构。来源：Li B，Wang Y，Dai J，Lohry W and Zhang S 2013。转载获得 Elsevier 的许可

5.8　结语

本章综述了一般 3D 传感知与结构光技术，解释了这些方法背后的原理，阐明了已经开发多年的实时 3D 传感数字条纹投影（DFP）技术，并展示了近期应用二进制散焦法开发的超高速 3D 传感技术的研究进展。在 3D 传感领域取得的突破是喜人的，而且目前已经发展到足够可以应对真实环境、日常生活的挑战的成熟阶段。在该项技术广泛地应用到我们的生活中之前，我们仍有一些问题需要解决，包括生产和购买这些设备的成本。但随着该技术向诸如人机交互等领域的迈进，其可能性和应用空间将是无穷的。

致谢

本章展示的成果是由以下人士共同合作参与的：来自美国纽约州立大学石溪分校的 Fu－Pen Chiang 和 Peisen Huang，来自哈佛大学的 Shing－Tung Yau，以及来自我们爱荷华州立大学实验室的众多学生：Laura Ekstrand，Shuanyan Lei，Beiwen Li，William Lohry 和 Yajun Wang。作者向他们致谢。该研究受到美国国家科学基金（NSF）的资助，资助编号：CMMI－1150711 和 CMMI－1300376。本章所述观点为作者观点，可能与 NSF 有出入。

参 考 文 献

1. Zhang, S. (ed.) (2013). *Handbook of 3D machine vision: Optical metrology and imaging,* 1st edition. CRC Press, New York, NY.
2. Geng, G. (2011). Structured-light 3D surface imaging: a tutorial. *Advances in Opt. and Photonics* **3**(2), 128–160.
3. Gorthi, S., Rastogi, P. (2010). Fringe projection techniques: Whither we are?. *Opt. Laser. Eng.* **48**, 133–140.
4. Zhang, S. (2010). Recent progresses on real-time 3-D shape measurement using digital fringe projection techniques. *Optics and Lasers in Engineering* **48**(2), 149–158.

5. Lei, S., Zhang, S. (2010). Digital sinusoidal fringe generation: defocusing binary patterns vs focusing sinusoidal patterns. *Optics and Lasers in Engineering* **48**(5), 561–569.
6. Guo, H., He, H., Chen, M. (2004). Gamma correction for digital fringe projection profilometry. *Appl. Opt.* **43**, 2906–2914.
7. Hu, Q., Huang, P.S., Fu, Q., Chiang, F.P. (2003). Calibration of a 3-D shape measurement system. *Opt. Eng.* **42**(2), 487–493.
8. Kakunai, S., Sakamoto, T., Iwata, K. (1999). Profile measurement taken with liquid-crystal grating. *Appl. Opt.* **38**(13), 2824–2828.
9. Pan, B., Kemao, Q., Huang, L., Asundi, A. (2009). Phase error analysis and compensation for nonsinusoidal waveforms in phase-shifting digital fringe projection profilometry. *Opt. Lett.* **34**(4), 2906–2914.
10. Zhang, S., Huang, P.S. (2007). Phase error compensation for a three-dimensional shape measurement system based on the phase shifting method. *Opt. Eng.* **46**(6), 063601.
11. Zhang, S., Yau, S.T. (2007a). Generic nonsinusoidal phase error correction for three-dimensional shape measurement using a digital video projector. *Appl. Opt.* **46**(1), 36–43.
12. Lei, S., Zhang, S. (2009). Flexible 3-D shape measurement using projector defocusing. *Opt. Lett.* **34**(20), 3080–3082.
13. Zhang, S., van der Weide, D., Oliver, J. (2010). Superfast phase-shifting method for 3-D shape measurement. *Opt. Express* **18**(9), 9684–9689.
14. Zhang, S., Wang, Y., Laughner, J.I., Efimov, I.R. (2013). Measuring dynamic 3D micro-structures using a superfast digital binary phase-shifting technique *ASME 2013 International Manufacturing Science and Engineering Conference*, Madison, Wisconsin.
15. Li B, Wang Y, Dai J, Lohry W, Zhang S. (2013). Some recent advances on superfast 3D shape measurement with digital binary defocusing techniques. *Optics and Lasers in Engineering* (doi:10.1016/j.optlaseng.2013.07.010).
16. Dhond, U., Aggarwal, J. (1989). Structure from stereo – a review. *IEEE Trans. Systems, Man, and Cybernetics* **19**, 1489–1510.
17. Salvi, J., Pages, J., Batlle, J. (2004). Pattern codification strategies in structured light systems. *Patt. Recogn.* **37**, 827–849.
18. Huang, Y., Shang, Y., Liu, Y., Bao, H. (2013). *Handbook of 3D Machine Vision: Optical Metrology and Imaging* 1st edn; CRC chapter 3D shapes from Speckle, pp. 33–56.
19. Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE Multimedia* **19**(2), 4–10
20. Hartley, R., Zisserman, A. (2000). *Multiple view geometry in computer vision* vol. 2. Cambridge Univ Press.
21. Pan, J., Huang, P., Zhang, S., Chiang, F.P. (2004). Color n-ary gray code for 3-D shape measurement *12th Intl Conf. on Exp. Mech.*
22. Carrihill, B., Hummel, R. (1985). Experiments with the intensity ratio depth sensor. *Computer Vision, Graphics and Image Processing* **32**, 337–358.
23. Chazan, G., Kiryati, N. (1995). Pyramidal intensity-ratio depth sensor. Technical report, Israel Institute of Technology, Technion, Haifa, Israel.
24. Jia, P., Kofman, J., English, C. (2007).Two-step triangular-pattern phase-shifting method for three-dimensional object-shape measurement. *Opt. Eng.* **46**(8), 083201.
25. Huang, P.S., Zhang, S., Chiang, F.P. (2005). Trapezoidal phase-shifting method for three-dimensional shape measurement. *Opt. Eng.* **44**(12), 123601.
26. Takeda, M., Mutoh, K. (1983). Fourier transform profilometry for the automatic measurement of 3-D object shape. *Appl. Opt.* **22**, 3977–3982.
27. Su, X., Zhang, Q. (2010). Dynamic 3-D shape measurement method: A review. *Opt. Laser. Eng.* **48**, 191–204.
28. Kemao, Q. (2004). Windowed Fourier transform for fringe pattern analysis. *Appl. Opt.* **43**, 2695–2702.
29. Guo, H., Huang, P. (2008). 3-D shape measurement by use of a modified fourier transform method. *Proc. SPIE* 7066:70660E.
30. Schreiber, H., Bruning, J.H. (2007). *Optical Shop Testing* 3rd edn. chapter Phase shifting interferometry, pp. 547–655. John Wiley & Sons.
31. Malacara, D. (ed.) (2007). *Optical Shop Testing* 3rd edn. John Wiley and Sons, New York.
32. Ghiglia, D.C., Pritt, M.D. (eds.) (1998). *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software*. John Wiley and Sons, New York.
33. Baldi, A. (2003). Phase unwrapping by region growing. *Appl. Opt.* **42**, 2498–2505.
34. Buchland, J.R., Huntley, J.M., Turner, S.R.E. (1995). Unwrapping noisy phase maps by use of a minimum-cost-matching algorithm. *Appl. Opt.* **34**, 5100–5108.

35. Chyou, J.J., Chen, S.J., Chen, Y.K. (2004). Two-dimensional phase unwrapping with a multichannel least-mean-square algorithm. *Appl. Opt.* **43**, 5655–5661.
36. Cusack, R., Huntley, J.M., Goldrein, H.T. (1995). Improved noise-immune phase unwrapping algorithm. *Appl. Opt.* **34**, 781–789.
37. Flynn, T.J. (1997). Two-dimensional phase unwrapping with minimum weighted discontinuity. *J. Opt. Soc. Am. A.* **14**, 2692–2701.
38. Ghiglia, D.C., Romero, L.A. (1996). Minimum *lp*-norm two-dimensional phase unwrapping. *J. Opt. Soc. Am. A.* **13**, 1–15.
39. Goldstein, R.M., Zebker, H.A., Werner, C.L. (1988). Two-dimensional phase unwrapping. *Radio Sci.* **23**, 713–720.
40. Hung, K.M., Yamada, T. (1998). Phase unwrapping by regions using least-squares approach. *Opt. Eng.* **37**, 2965–2970.
41. Huntley, J.M. (1989). Noise-immune phase unwrapping algorithm. *Appl. Opt.* **28**, 3268–3270.
42. Merraez, M.A., Boticario, J.G., Labor, M.J., Burton, D.R. (2005). Agglomerative clustering-based approach for two dimensional phase unwrapping. *Appl. Opt.* **44**, 1129–1140.
43. Salfity, M.F., Ruiz, P.D., Huntley, J.M., Graves, M.J., Cusack, R., Beauregard, D.A. (2006). Branch cut surface placement for unwrapping of undersampled three-dimensional phase data: Application to magnetic resonance imaging arterial flow mapping. *Appl. Opt.* **45**, 2711–2722.
44. Zhang, S., Li, X., Yau, S.T. (2007a). Multilevel quality-guided phase unwrapping algorithm for real-time three-dimensional shape reconstruction. *Appl. Opt.* **46**(1), 50–57.
45. Xu, Y., Ekstrand, L., Dai, J., Zhang, S. (2011). Phase error compensation for three-dimensional shape measurement with projector defocusing. *Appl. Opt.* **50**(17), 2572–2581.
46. Zhang, S., Yau, S.T. (2006). High-resolution, real-time 3-D absolute coordinate measurement based on a phase-shifting method. *Opt. Express* **14**(7), 2644–2649.
47. Zhang, S., Huang, P.S. (2006b). Novel method for structured light system calibration. *Opt. Eng.* **45**(8), 083601.
48. Creath, K. (1987). Step height measurement using two-wavelength phase-shifting interferometry. *Appl. Opt.* **26**(14), 2810–2816.
49. Cheng, Y.Y., Wyant, J.C. (1985). Multiple-wavelength phase shifting interferometry. *Appl. Opt.* **24**, 804–807.
50. Sansoni, G., Carocci, M., Rodella, R. (1999). Three-dimensional vision based on a combination of gray-code and phase-shift light projection: Analysis and compensation of the systematic errors. *Appl. Opt.* **38**, 6565–6573.
51. Towers, C.E., Towers, D.P., Jones, J.D. (2003). Optimum frequency selection in multifrequency interferometry. *Opt. Lett.* **28**(11), 887–889.
52. Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(11), 1330–1334.
53. Cuevas, F.J., Servin, M., Rodriguez-Vera, R. (1999). Depth object recovery using radial basis functions. *Opt. Commun.* **163**(4), 270–277.
54. Legarda-Sáenz, R., Bothe, T., Jüptner, W.P. (2004). Accurate procedure for the calibration of a structured light system. *Opt. Eng.* **43**(2), 464–471.
55. Gao, W., Wang, L., Hu, Z. (2008). Flexible method for structured light system calibration. *Opt. Eng.* **47**(8), 083602.
56. Huang, P., Han, X. (2006). On improving the accuracy of structured light systems *Proc. SPIE.* **6382**, 63820H.
57. Li, Z., Shi, Y., Wang, C., Wang, Y. (2008). Accurate calibration method for a structured light system. *Opt. Eng.* **47**(5), 053604.
58. Yang, R., Cheng, S., Chen, Y. (2008). Flexible and accurate implementation of a binocular structured light system. *Opt. Lasers Eng.* **46**(5), 373–379.
59. Zhang, S., Yau, S.T. (2006). High-resolution, real-time 3-D absolute coordinate measurement based on a phase-shifting method. *Opt. Express* **14**(7), 2644–2649.
60. Zhang, Z., Towers, C.E., Towers, D.P. (2007b). Phase and colour calculation in color fringe projection. *J. Opt. A: Pure Appl. Opt.* **9**, S81–S86.
61. Wang, Y., Zhang, S. (2011). Superfast multifrequency phase-shifting technique with optimal pulse width modulation. *Opt. Express* **19**(6), 5143–5148.
62. Rusinkiewicz, S., Hall-Holt, O., Levoy, M. (2002). Real-time 3D model acquisition. *ACM Trans. Graph.* **21**(3), 438–446.
63. Hall-Holt, O., Rusinkiewicz, S. (2001). Stripe boundary codes for real-time structured-light range scanning of moving objects *The 8th IEEE International Conference on Computer Vision* **II**, 359–366.
64. Geng, Z.J. (1996). Rainbow 3-D camera: New concept of high-speed three vision system. *Opt. Eng.* **35**, 376–383.

65. Harding, K.G. (1988). Color encoded morié contouring *Proc. SPIE* **1005**, 169–178.
66. Huang, P.S., Hu, Q., Jin, F., Chiang, F.P. (1999). Color-encoded digital fringe projection technique for high-speed three-dimensional surface contouring. *Opt. Eng.* **38**, 1065–1071.
67. Pan, J., Huang, P.S., Chiang, F.P. (2006). Color phase-shifting technique for three-dimensional shape measurement. *Opt. Eng.* **45**(12), 013602.
68. Zhang, S., Huang, P. (2004). High-resolution, real-time 3-D shape acquisition *IEEE Comp. Vis. and Patt Recogn Workshop* **3**, 28–37, Washington DC, MD.
69. Huang, P.S., Zhang, S. (2006). Fast three-step phase shifting algorithm. *Appl. Opt.* **45**(21), 5086–5091.
70. Zhang, S., Royer, D., Yau, S.T. (2006a). GPU-assisted high-resolution, real-time 3-d shape measurement. *Opt. Express* **14**(20), 9120–9129.
71. Zhang, S., Yau, S.T. (2007b). High-speed three-dimensional shape measurement using a modified two-plus-one phase-shifting algorithm. *Opt. Eng.* **46**(11), 113603.
72. Gong, Y., Zhang, S. (2010). *Opt. Express* **18**, 19743.
73. Hornbeck, L.J. (1997). Digital light processing for high-brightness, high-resolution applications *Proc. SPIE* **3013**, 27–40.
74. Zhang, S., Huang, P.S. (2006a). High-resolution, real-time three-dimensional shape measurement. *Opt. Eng.* **45**(12), 123601.
75. Ekstrand, L. *et al.* (2013). *Handbook of 3-D machine vision: Optical metrology and imaging* **9**, 233.
76. Ujaldon, M., Saltz, J. (2005). Exploiting parallelism on irregular applications using the gpu. *Intl. Conf. on Paral. Comp.*, **13–16**.
77. Khailany, B., Dally, W., Rixner, S., Kapasi, U., Owens, J., Towles, B. (2003). Exploring the vlsi scalability of stream processors. *Proc. 9th Symp. on High Perf. Comp. Arch.*, 153–164.
78. Zhang, S., Royer, D., Yau, S.T. (2006a). GPU-assisted high-resolution, real-time 3-d shape measurement. *Opt. Express* **14**(20), 9120–9129.
79. Zhang, S., Royer, D., Yau, S.T. (2006b). Gpu-assisted high-resolution, real-time 3-D shape measurement. *Opt. Express* **14**(20), 9120–9129.
80. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. *Conference on Computer Vision and Pattern Recognition Workshop, 2003, CVPRW '03*, 53–53.
81. Mauri, C., Granollers, T., Lorés, J., García, M. (2006). Computer vision interaction for people with severe movement restrictions. *Human Technology* **2**(1), 38–54.
82. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y. (2011). Vision-based hand-gesture applications. *Commun. ACM* **54**(2), 60–71.
83. Nielsen, M., StÃrring, M., Moeslund, T., Granum, E. (2004). A procedure for developing intuitive and ergonomic gesture interfaces for hci. In Camurri, A., Volpe, G. (eds). *Gesture-Based Communication in Human-Computer Interaction,* vol. 2915 of *Lecture Notes in Computer Science,* pp. 409–420. Springer Berlin Heidelberg.
84. Zhang, L., Curless, B., Seitz, S.M. (2002). Rapid shape acquisition using color structured light and multi-pass dynamic programming. *The 1st IEEE International Symposium on 3D Data Processing, Visualization, and Transmission*, 24–36.
85. Zhang, S., Yau, S.T. (2008). Simultaneous three-dimensional geometry and color texture acquisition using single color camera. *Opt. Eng.* **47**(12), 123604.
86. Zhang, Z., Towers, C.E., Towers, D.P. (2006c). Time efficient color fringe projection system for 3D shape and color using optimum 3-frequency selection. *Opt. Exp.* **14**, 6444–6455.
87. Liu, X., Peng, X., Chen, H., He, D., Gao, B.Z. (2012). Strategy for automatic and complete three-dimensional optical digitization. *Opt. Lett.* **37**, 3126–3128.
88. Notni, G.H., Kühmstedt, P., Heinze, M., Notni, G. (2002). Simultaneous measurement of 3-D shape and color of objects *Proc. SPIE* **4778**, 74–82.
89. Ou, P., Li, B., Wang, Y., Zhang, S. (2013). Flexible real-time natural 2d color and 3D shape measurement. *Optics Express* **21**(14), 16736–16741.
90. Merner, L., Wang, Y., Zhang, S. (2013). Accurate calibration for 3D shape measurement system using a binary defocusing technique. *Opt. Laser Eng.* **51**(5), 514–519.
91. Ayubi, G.A., Ayubi, J.A., Martino, J.M.D., Ferrari, J.A. (2010). Pulse-width modulation in defocused 3-D fringe projection. *Opt. Lett.* **35**, 3682–3684.
92. Wang, Y., Zhang, S. (2010). Optimum pulse width modulation for sinusoidal fringe generation with projector defocusing. *Opt. Lett.* **35**(24), 4121–4123.

93. Wang, Y., Zhang, S. (2012a). Comparison among square binary, sinusoidal pulse width modulation, and optimal pulse width modulation methods for three-dimensional shape measurement. *Appl. Opt.* **51**(7), 861–872.
94. Lohry, W., Zhang, S. (2012). 3D shape measurement with 2D area modulated binary patterns. *Opt. Laser Eng.* **50**(7), 917–921.
95. Schuchman, T.L. (1964). Dither signals and their effect on quantization noise. *IEEE Trans. Communication Technology* **12**(4), 162–165.
96. Wang, Y., Zhang, S. (2012b). Three-dimensional shape measurement with binary dithered patterns. *Appl. Opt.* **51**(27), 6631–6636.
97. Bayer, B. (1973). An optimum method for two-level rendition of continuous-tone pictures. *IEEE International Conference on Communications* **1**, 11–15.
98. Floyd, R., Steinberg, L. (1976). An adaptive algorithm for spatial gray scale. *Proc. Society for Information Display* **17**, 75–77.
99. Kite, T.D., Evans, B.L., Bovik, A.C. (2000). Modeling and quality assessment of halftoning by error diffusion. *IEEE International Conference on Image Processing* **9**(5), 909–922.
100. Purgathofer, W., Tobler, R., Geiler, M. (1994). Forced random dithering: improved threshold matrices for ordered dithering. *IEEE International Conference on Image Processing* **2**, 1032–1035.
101. Stucki, P. (1981). Meccaa multiple-error correcting computation algorithm for bilevel hardcopy reproduction. Technical report, IBM Res. Lab., Zurich, Switzerland.
102. Wang, Y., Laughner, J.I., Efimov, I.R., Zhang, S. (2013). 3D absolute shape measurement of live rabbit hearts with a superfast two-frequency phase-shifting technique. *Opt. Express* **21**(5), 5822–5632.
103. Dai, J., Zhang, S. (2013). Phase-optimized dithering technique for high-quality 3D shape measurement. *Opt. Laser Eng.* **51**(6), 790–795.
104. Lohry, W., Zhang, S. (2013). Genetic method to optimize binary dithering technique for high-quality fringe generation. *Opt. Lett.* **38**(4), 540–542.
105. Huang, P.S., Zhang, C., Chiang, F.P. (2002). High-speed 3-D shape measurement based on digital fringe projection. *Opt. Eng.* **42**(1), 163–168.

第6章

实时立体3D成像技术

LazarosNalpantidis
丹麦哥本哈根奥尔堡大学，机械制造工程系视觉与机器智能实验室机器人研究组

6.1 引言

立体视觉是利用两个视觉传感器同时作用重建景深的技术。其基本原理蕴含于自然之中，空间的不同两点观察同一事物形成的像差能够为感知被观察事物的景深提供足够信息。这一现象首先是由 Charles Wheatstone 爵士在大约两个世纪前发现的。他声称："……大脑是通过投射到两个视网膜上产生的两张不同图像来感知事物的三个维度的……"[1]。

计算机和机器人视觉系统的重要任务之一是将摄像头记录的场景中各点的深度及其他原始数值推算出来。从亮度图像中提取深度信息最常用的方法是安装立体摄像机（见图 6.1），然后通过该摄像机的一组同步成像照片获取所拍事物的深度信息。同一景物点在不同成像平面中像素的对应关系（也被称作立体匹配问题）就形成了所谓的视差图[2]。视差是在观察两景物点时相对应的像素坐标的差别，而景物点的实际深度值就可以根据该视差值按照一定比例换算出来。调整好的立体摄像机通常情况下垂直视差为零。这就是说，视差图一般是用来记录对应的图像像素的水平视差值的。然而，如何对视差图进行精确高效的估算却是计算机视觉领域的一个长期存在的难题[3]。

立体视觉在机器视觉[4]、计算机视觉[5]、虚拟现实、机器人导航[6]、同步定位与地图绘制[7,8]、深度测量[9]和 3D 环境再造[4]过程中的重要性显而易见。本章旨在全面描述实时立体成像算法和系统，将着重阐述立体视觉算法的主要特征，有待深入分析处辅以相关参考文献，并会介绍与实时技术相关的最新发展现状。下面将采用图 6.2 所示分类方式对实时立体 3D 成像技术进行分类。

硬质立体摄像机　　类人活动立体头部

a)　　b)

图 6.1　立体视觉传感器

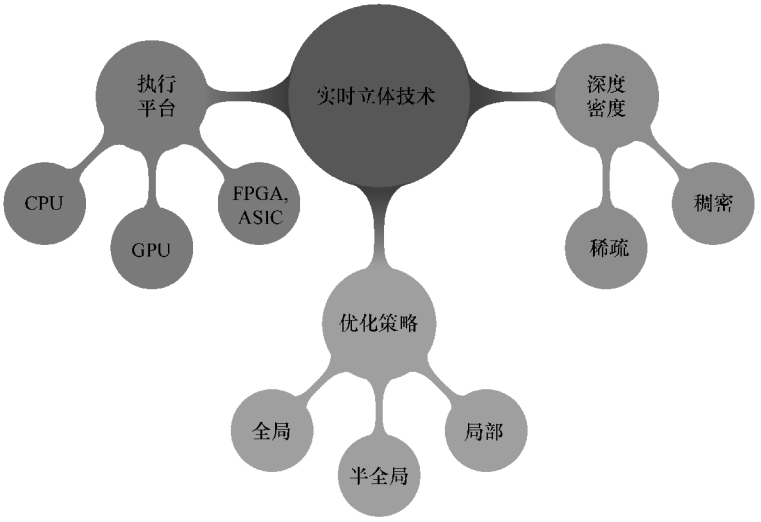


图 6.2　实时立体 3D 成像技术的分类

6.2　背景

虽然在过去的几十年里，立体视觉一直是许多研究人员关注的焦点，而今它仍然不失为一个非常活跃热门的专题。每年新增的相关文献和大量新出版的著作都表明，该项技术的最新成果仍有很大的改进空间。正因为此，每次对最新发展现状尝试进行的概述都注定很快过时。然而，研究这些概述我们仍然可以推导出，该研究重点的变化和其发展过程中的新趋势。

历史上，第一次收集、调查和比较立体视觉算法出现在 Barnard 和 Fischler[10]、Dhond 和 Aggarwal[11]，以及 Brown [12]的综述文献中。然而，对该研究问题影响最大且对研究方向有明确指引作用的著作或许是 Scharstein 和 Szeliski[13]的开创性分类方式及其综述文献。该著作中，除了详细介绍了当代算法外，还提出了算法的分类框架和一个公开可用的客观测试平台。测试平台包括一个标准的立体图像数据集，评估结果准确性的指标，以及一个承载所

有上述组件的网站，以及一个不断更新列表展示评估后的算法结果[14]。

尽管 Scharstein 和 Szeliski 的研究重点是定义能够使计算机视觉界进一步追求立体算法精度，机器人应用和实时视觉系统的快速演进表明执行速度这一影响因素和深度估计精确度同样（甚至更加）重要。这种趋势在后面的文献综述中均有所体现。例如，涉及实时硬件运行的文献[15,16]，以及重点阐述面向资源受限系统的实时算法文献[17]。

除了有的文献对完整的立体算法进行文献综述和对比演示外，还有些非常有用的著作会对立体算法的基本执行模块的各种替代解决方案进行对比。因而 6.3 节中将要对立体算法的结构进行详细定义，并明确其各组成部分。Hirschmuller 和 Scharstein 在参考文献［18，19］中，就全局、半全局以及局部立体算法的不同（不）相似测量方法（也被称作匹配成本函数）进行了对比。

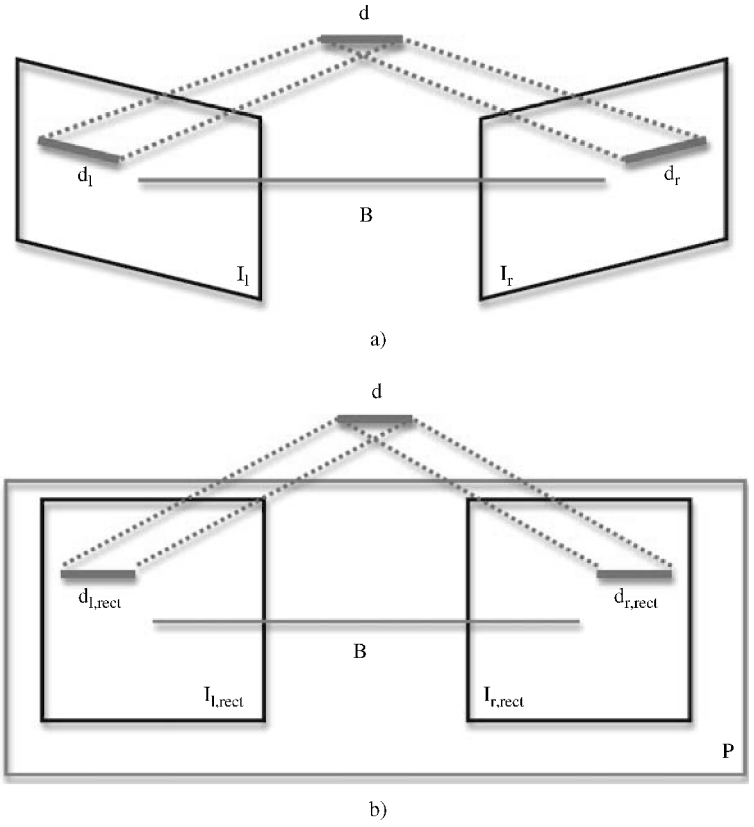


图 6.3　立体图像对的校正。同一物体 d 的两个图像 I_l、I_r 被处于公共平面 P 的校正图像 $I_{l,\,rect}$、$I_{r,\,rect}$替换

此外，在参考文献［20］中，Gong 等学者对各种基于实时 GPU 加速系统的匹配成本聚合解决方案进行了汇总。考虑到全局方法的分配差距，Szeliski 等人在参考文献［21］中提出了一些能量最小化基准，并应用这些基准对结果质量和几种常见能量最小化算法的速度进行了比较。

立体匹配问题可以通过观察这两幅立体图像的几何结构并对图像进行校正的方法得以有

效解决。通常情况下，双摄像头的两个图像平面不在同一平面内。然而立体算法可以处理这种情况，如果立体图像对已被校正，那所需的计算就会大大简化。如图 6.3 所示，校正过程包括一对原始图像 I_l、I_r 被另一对投影等效对 $I_{l,rect}$ 和 $I_{r,rect}$[5,2] 替代，原始图像被再次投射到与连接原始图像的两个光学中心的基准线 B 平行的公共平面 P 上面。

对极几何原理可以用于识别两个图像的相同特征，这为解决立体匹配问题提供了工具。如果不进行校正，那么匹配过程还将涉及在目标图像的二维区域内进行搜索。不过，如果假定精准校正后的立体图像对在水平扫描线内，且归于相同的纵向线，则这种匹配可以看作是一维搜索，如图 6.4 所示。图像平面上的点 P_1 可以是线 C_1P_1 上的任意一点，也可能出现在交替图像平面内的纵向线 E_2[4] 上的任意一点。因此，由于对应点归于同一纵向线，扫描线上的搜索过程理论上减少了。这些点的水平坐标差异就是差异值。视差图就是由图像的所有点的差异值组成的。

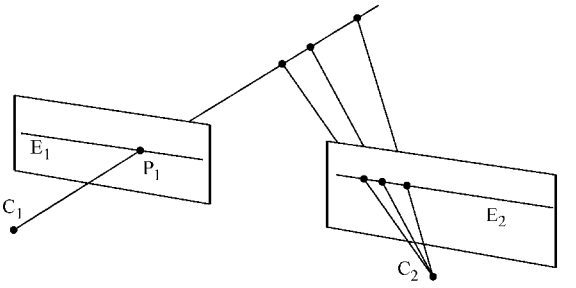


图 6.4　纵向线的几何原理，其中 C_1 和 C_2 分别是摄像机的左、右镜头中心。图像平面上的点 P_1 可以是线 C_1P_1 上任意一点，也可能出现在交替图像平面内的纵向线 E_2 上的任意一点

6.3　立体匹配算法的结构

报告中的大多数立体匹配算法都或多或少使用相同的结构集[13]。基本构建模块如下：

1）计算两个输入图像中每个像素的匹配成本函数。

2）支持区域内每个像素和每个潜在视差值的匹配成本计算的汇总。

3）图像中每一像素的最佳视差值的选取。

4）成型视差图的优化完善。

每个立体匹配算法利用匹配成本函数建立两个像素间的对应关系，该部分将在 6.3.1 节中进行讨论。匹配成本计算的结果包含视差空间图像（DSI）。视差空间图像可以被视为一个 3D 矩阵，该矩阵包括每个像素和所有潜在视差值的匹配成本计算结果[22]。DSI 的结构如图 6.5 所示。

通常情况下，匹配成本聚合要超出支持区域。这些区域可以是 DSI 立方体范围内的 2D 区域，甚至是 3D[23,24] 区域。每个像素的最佳视差值的选取将随之进行。选取过程可以是一个简单的胜者全得（WTA）的过程，也可以是更复杂的过程。而更多情况下，该过程是一个迭代过程，如图 6.6 所示。通常这一过程还会采用一个额外的视差优化步骤，旨在过滤计算过的视差值，以提供亚像素精度或为未计算的像素分配视差。大多数的立体匹配算法的一般结构如图 6.6 所示。每个模块将会在本节的剩余部分中详细讨论。

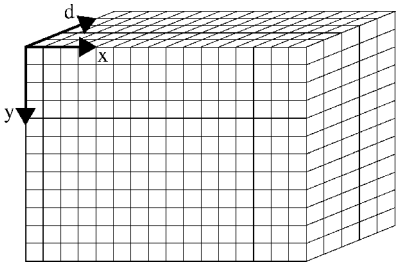


图 6.5　DSI 包含所有图像像素以及所有潜在视差值的匹配成本

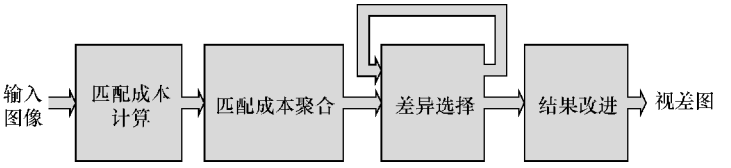


图 6.6　立体匹配算法的一般结构

6.3.1　匹配成本计算

在立体图像中检测共轭对是一个具有挑战性的研究问题，俗称匹配问题，也就是左侧图像上每个像素在右侧图像上寻找对应像素[25]。没有任何歧义实现匹配的像素应明显不同于其周围像素。为了确定两像素是否形成共轭对，采用某种匹配成本函数测量这些像素的相似性是必要的。其中最常见的匹配成本函数是绝对灰度差异（AD）、平方灰度差异（SD）和归一化互相关（NCC）。各种匹配成本的评价总结在参考文献［13，18，26］中。

绝对灰度差异（AD）是最简单的测量方法。它包括简单的减法和绝对值的计算。因此，它是文献中提及的最常用的测量方法。绝对灰度差异（AD）的数学公式是

$$\mathrm{AD}(x, y, d) = |I_{\mathrm{left}}(x, y) - I_{\mathrm{right}}((x-d), y)| \tag{6.1}$$

式中，I_{left}和I_{right}分别指左侧图像和右侧图像的灰度值；d 表示从 0 到 $D-1$ 的测试中的差值；x、y 是图像平面上像素的坐标值。

在某种程度上，平方灰度差异（SD）在表达两像素差异性时更加精准。然而，平方灰度差异的计算成本较其带来的精度增益而言较高。计算公式为

$$\mathrm{SD}(x, y, d) = (I_{\mathrm{left}}(x, y) - I_{\mathrm{right}}((x-d), y))^2 \tag{6.2}$$

归一化互相关（NCC）计算的是图像区域而非单像素的差异性。该方法能够对计算负荷成本得出可信度高的结果。其数学表达式是

$$NCC(x, y, d) = \frac{\sum_{x,y \in W} I_{\mathrm{left}}(x, y) \cdot I_{\mathrm{right}}(x, y-d)}{\sqrt{\sum_{x,\, y \in W} I_{\mathrm{left}}^2(x, y) \cdot \sum_{x,\, y \in W} I_{\mathrm{right}}^2(x, y-d)}} \tag{6.3}$$

式中，W 表示需要计算的图像区域。

亮度补偿式差异性测量法（LCDM）在参考文献［27］中进行了介绍，这种测量法主要用于在照明条件较差的环境中进行的立体成像算法。被测量的图像从最初的 RGB 转化成 HSL 色彩空间。HSL 色彩空间本质上表达了颜色的亮度和它的定性特征[28]。也就是说，无论环境的光照条件如何，一个对象产生的 H 值和 S 值不变。基于上述假设，亮度补偿式差异性测量法（LCDM）忽视亮度值来计算两种颜色差异性。

色彩空间表达法略去纵轴 L 则形成一个仅由 H 和 S 定义的 2D 圆形。这样，任何颜色都可以描绘以圆心为起始点的平面向量。因此，任何颜色 $\boldsymbol{P}_k$ 可以描述为一个极向量，或者模为 Sk，幅角为 Hk 的复数。因此，两个颜色 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 的差别，即亮度补偿式差异性测量法（LCDM）可以通过计算这两个复数之差来计算。

$$\begin{aligned}\mathrm{LCDM}_{P_1,P_2} &= |\boldsymbol{P}_1 - \boldsymbol{P}_2| \\ &= |S_1 e^{iH_1} - S_2 e^{iH_2}| \\ &= \sqrt{S_1^2 + S_2^2 - 2S_1S_2\cos(H_1 - H_2)}\end{aligned} \tag{6.4}$$

另一方面，对于一个特定区域的所有像素，秩变换替代了同秩间的像素灰度，然后计算可能匹配间的绝对差异[29]。另一个匹配成本算法是交互信息（Mutual Information），主要用于像素相关性。该算法通过计算和评估联合及个体熵和概率分布可得出一个图像的两部分的相似性。可以通过相应图像部分的直方图得出概率分布。最后，还有基于相位的方法，可以把图像看作是信号，并执行相位相关函数的匹配。

参考文献［19］对确定像素相关性的各种匹配成本进行综述和评价。

6.3.2　匹配成本聚合

通常，匹配成本聚合在支持区域进行。支持区域一般指支持窗口或聚合窗口。它可以是正方形，也可以是矩形；大小可以固定，也可以根据需求变化。上述成本函数的聚合就是多数立体视觉法的核心。以绝对差之和（SAD）为例，数学表达式如下：

$$\mathrm{SAD}(x,y,d) = \sum_{x,y\in W} |I_{\mathrm{left}}(x,y) - I_{\mathrm{right}}((x-d),y)| \tag{6.5}$$

以平方差之和为例（SSD）：

$$\mathrm{SSD}(x,y,d) = \sum_{x,y\in W} (I_{\mathrm{left}}(x,y) - I_{\mathrm{right}}(x-d,y))^2 \tag{6.6}$$

式中，I_{left}、I_{right}表示左、右图像的灰度值；x、y 表示像素的坐标；d 表示需要考虑的视差值；W 表示聚合后的支持区域。

一般来讲，越复杂的聚合方法越耗时。基于可扩展局部方法的聚合方式为了获得更精确的结果[30]舍弃了计算的简洁性。基于自适应支持权重（ASW）的方法[31,32]通过使用固定大小的支持窗口获取精确的结果。在这种方式中，像素在聚合阶段的贡献各不相同，而这决定了各像素和窗口中心像素的关联程度。尽管这些方式被广泛接受，但使用何种关联函数仍然是一个悬而未决的专题。

基于自适应支持权重的匹配搜索方法在参考文献［31］中进行了介绍。给定的支持窗口中，像素的支持权重要根据颜色相似性和几何接近度进行调整，以减少图像的模糊程度。参考文献［27］中检查非理想光照条件下的立体视觉和参考文献［33］中心理物理学启发下的立体算法均采用了上述的类似方法。

另一方面，一些匹配成本函数中含有不可分割的成本计算过程，因而不再需要额外的聚合步骤。例如，传统归一化互相关（NCC）算法和秩变换要求对需要计算的区域有一个先验定义。

在参考文献［20］中对适用于可编程 GPU 运行平台并面向实时系统的不同匹配成本聚合方法进行深入综述和评论。

6.4　特征分类

正如上面所讲，每年有大量与立体视觉算法和系统相关的论文出版。这份不断增长的论文清单需要依据某种有意义的标准进行分类，以易于管理搜索。传统意义上来说，最常见的是根据计算视差图的密度和最终视差值的分配策略对立体算法文献进行分类。本节余下部分将对分类的门类进行剖析，并对一些指示性算法进行讨论。

6.4.1　深度估计密度

立体匹配算法可以分为两大类别：一种生成密集结果，另一种生成稀疏结果。密集立体算法并不一定要获得每一像素的视差估算。事实上由于立体对遮蔽现象（即使有些技术可以让信息传递到遮蔽区域），通常也不能获得 100% 的密集结果。因此，密集和稀疏算法的界限划分，要取决于该算法真正在所有像素匹配中使用，还是仅仅用于某特定像素子集中。

6.4.1.1　密集算法

随着越来越多的计算能力应用于视觉系统中，密集立体算法（dense stereo algorithms）也变得越来越流行。如今，有关密集算法的出版物占据了相关文献的主要份额。除了所需计算资源的可用性，另一个促进密集算法使用的因素是用于评估的标准测试平台和计算结果准确性的客观比较的存在。因此，根据 Maimone 和 Shafer[34] 的观点，当前立体匹配算法的研究是由 Scharstein 和 Szeliski[13, 14] 持有的在线工具所支配。该网站提供了一个共享数据集，并且支持所产生的视差图的上传，自动评估，并于其他结果同时列出以便于比较。

正如已经讨论过的，并不是图像上的每一个像素都有一个立体对图像与之对应。这主要是由于遮蔽造成的。然而应用某种全局最优化或“填充”机制的算法可达到 100% 的覆盖率。参考文献［35］提出了一种基于互信息匹配成本的分层计算法。这种算法的目的是尽量通过从各个方向聚合各像素的匹配成本来最小化适当的全局能量函数，而非通过迭代优化。最终的视差图达到亚像素精准级，并能够检测出遮蔽区域。该算法基于主频为 2.8GHz 的英特尔 Xeon 处理器平台，处理 Teddy 图像集的速度是 0.77 帧/s。结果发现未遮蔽区域的误差比例小于所有标准图像集的 3%。

前述方法的增强版也是由参考文献［36］的同一个作者提出的。互信息又一次被用作成本函数，扩展概念在其中的应用会导致无纹理区域内灰度一致视差选择和视差图中填充漏洞时的不连续保护性插补。它可以成功处理复杂的形状，并在无纹理区域使用平面模型。双向一致性检查和亚像素评估，连同无效视差插值都参与在此过程中。实验结果表明在非遮蔽区域，Tsukuba、Venus、Teddy 和 Cones 图像集中不匹配像素的比例分别为 2.61、0.25、5.14 和 2.77，且每次搜索的差异水平为 64，而且 2.8GHz 计算机记录的运行速度是小于 1 帧/s 的。

6.4.1.2　稀疏算法

稀疏立体算法（sparse stereo algorithms）只为所有图像像素中的某有限子集提供景深估

算。稀疏视差图通过提取并匹配有区别度的图像特征，或根据某种可靠性测量法排除掉密集算法的某些结果来获得。很多情况下，这些算法源于对人类视觉的研究，并且基于例如对两张图像的部分或边缘进行匹配。

由于大多数现代应用程序需要密集的视差信息，因此，使用稀疏或半密集算法产生的视差图往往不太有吸引力。然而这些算法在如下情况却非常有用：

- 需要非常快速的执行时间。
- 不需要整个图像的细节。
- 所获的景深估算的可信度要求比密度要求高。

这些类型的算法通常只关注图像的明显特征，遮蔽的和纹理不佳的区域内像素却无法使用这些算法找到匹配对。Veksler 在参考文献［37］中提出一种算法可以检测并匹配立体对中左右图像的密集特征，进而形成半密集视差图。密集特征是在左图像中一组关联像素集和其对应右图像中的关联像素集。这两个像素集的边界灰度边缘比其匹配误差更加明显。所有这些计算过程均在立体匹配过程中进行。这些算法计算出的视差图对非遮蔽区域的 Tsukuba 图像对分为 14 个视差级别，得出 66% 的密集度和 0.06% 的平均误差。

另一种算法也是由 Veksler[38] 基于与之前方法相同的基本概念提出的。主要区别在于这种方法针对密度特征提取使用了图形切割算法。结果显示，这种算法生成半密集结果，且特征检测区域的半密集结果相当精确，从密集和误差比率来看，计算结果更加精确，但计算需要更长的时间。而对于 Tsukuba 图像对，密集度可高达 75%，非遮蔽区域内的总误差比率为 0.36%，运算速度为 0.17 帧/s。对于 Sawtooth 图像对，相应值分别为 87%、0.54% 和 0.08 帧/s。所有上述结果均是从主频为 600MHz 的奔腾 III 电脑上获取的。因而我们可以断言，使用更强大的运算系统可以大幅提速。

参考文献［39］中，通过使用 Harris 角点检测器提取特征点启动稀疏立体算法。这些特征点构成图谱，再利用图谱切割法在损耗最小全局能量的前提下，解决标记问题。此外，稳定的光照变化结构张量描述器也被用于相似测量中，以便获取更加精确的结果。

最近，Schauwecker 等人[40] 将改进的高效快速特征检测器（FAST feature detector）与稀疏立体匹配算法结合。额外的一致性检测可以滤除可能的错误匹配，最终在简单的双核 CPU 电脑平台上达到 200 帧/s 的运行速度。

6.4.2　优化策略

根据像素视差分配方式的不同，立体匹配算法可以分为三大类。首先，有些算法会根据局部相邻像素提供的信息，决定每一像素的视差。这种算法称为局部或基于区域的方法。这些方法也被称作基于窗口的方法，这是因为给定点的视差计算只取决于有限支持窗口内的灰度值。其次，有些算法对每一像素的视差分配是基于整个图像提取的信息。因此这种算法被叫作全局算法。有时这些算法也被称作基于能量的算法，因为这些算法意在最小化全局能量函数，包括数据项和平滑项，并将整个图像考虑在内。最后一类被称为半全局算法，这种算法沿扫描线选择视差值，以使能量函数最小化[41]。当然还有很多其他的算法[42] 不能严格

地包含在这三大类中。立体匹配问题也引入了多种多样的计算工具，先进的智能计算技术并不少见，其中包含着很多有趣而又混杂的研究方向[43,44]。

6.4.2.1　局部算法

在匹配成本的计算和聚合（如有必要）之后，应进行最佳候选匹配像素（和视差值）的实际选择。多数局部算法只选择呈现最小匹配成本的像素作为匹配候选对象。在计算方面这种简单的“胜者全得”的方法很有效，但也往往不够准确。它会导致错误和不连贯视差值的产生。

在参考文献［45 – 47］中，Ogale 和 Aloimonos 提出了一种结合很多早期可视模块（例如，分段、形状和深度评估、遮蔽探测和局部信号处理）的组合方法。采用的相异性测量是不同频率通道的相位差。因而这种方法可以通过对比处理图像，并区别于其他错误匹配。

6.4.2.2　全局算法

相对于局部算法，全局算法计算结果准确，但耗时也多。这类方法将视差分配步骤视为标记问题，其目的是通过结合数据项和平滑项，来寻找最佳差异函数 $d = d(x, y)$，从而减少全局成本函数。

$$E(d) = E_{\text{data}}(d) + \lambda E_{\text{smooth}}(d) \tag{6.7}$$

式中，E_{data}综合考虑整个图像中 x、y 的像素值；E_{smooth}提供了算法的平滑假设；λ 是权重因子。

使用合适的迭代算法能够使能量函数最小化。常用的算法还包括图形切割算法[48, 49]和环路置信传播（loopy belief propagation）[50]。然而，全局算法的主要缺点是耗时多，计算量大。

参考文献［21］中对全局立体匹配的各种能量最小化方法进行了全面综述和对比。

6.4.2.3　半全局算法

最受欢迎的半全局立体匹配算法是基于动态规划（DP）的。DP 是一种广泛应用的优化方法，该算法可沿图像扫描线评估视差值 d[41,51]，这也是它被称为半全局算法的原因。DP 立体算法的基本思想是将匹配问题作为能量最小化问题重新处理，但仅限于沿扫描线范围内。因此，可以建立能量函数式（6.7），式中引入平滑项的概念来处理每个扫描线的景深不连贯和遮蔽情况。

DP 算法介于局部算法和全局算法之间，它以可接受的帧速率提供了准确性高的结果。而且，近年来基于动态规划的立体算法似乎在这两方面有显著改善。参考文献［52，53］阐述了 DP 的硬件平台，该平台可提供很高的执行速度。此外，数据显示，自适应支持权重聚合方案的引入进一步提高了所生成的景深图的准确度和精细度[54]。

6.5　实施平台的分类

许多立体视觉算法可以实现实时或者近实时操作。这样的执行速度可以通过优化局部立体算法，或采用定制的加速计算硬件实现。基于中央处理单元（CPU）、图形处理单元

（GPU）以及现场可编程门阵列（FPGA），或专用集成电路（ASIC）的实现均可用于实时系统中，这些运行平台会在本节中进行讨论。

6.5.1　仅用 CPU 的方法

在参考文献［55］中，Gong 和 Yang 提出了一种基于可信度的动态规划（RDP）算法。这种算法应用了一种不同的策略方法来评价匹配的可信度。根据这一点，所提算法的可信度是包括全局最佳差异分配和不包括全局最佳差异分配间的成本差异。DP 算法中，扫描线间的连贯性问题是通过可信度阈值的处理过程减少的。其运算结果形成了一个半密集明确的视差图。处理 Tsukuba 图像对时，密集度为 76%，错误率为 0.32%，运行速度为 16 帧/s。处理 Sawtooth 图像对时，密集度为 72%，错误率为 0.23%，运行速度为 7 帧/s。相应的，对于 Venus 和 Map 图像对，相应结果为 73%、0.18%、6.4 帧/s 和 86%、0.7%、12.8 帧/s。所以，报告结果显示，如果半密集度视差图可以接受的话，运行在 2GHz 奔腾 4 计算机上的实际操作结果很令人振奋。

参考文献［56］中阐述了 Tombari 等人提出的一种通过有效分割型 AD 成本聚合策略实现速度 - 精度权衡最大化的局部立体算法。报告中 Tsukuba 图像对的处理速率为 5 帧/s，Teddy 和 Art 立体图像对的处理速率为 1.7 帧/s。

最后，参考文献［57］阐述了一种实时计算密集视差图的局部立体匹配算法。所使用支持窗口的两种型号提高了基础绝对差之和（SAD）的准确性，同时也保证了计算过程的低成本。

6.5.2　GPU 提速的方法

计算机系统中 GPU 的有效使用可以明显地通过利用 GPU 并行计算的能力提高执行速度。

参考文献［58］中报告的是基于可编程 3D 图形处理单元（GPU）的分层视差评估算法。这种方法既可以处理校准的图像对，也可以处理未校准的图像对。双向匹配连同绝对灰度差的局部聚合总和一起使用。在 ATI Radeon 9700 Pro 的运行平台上，对 256 × 256 像素输入图像，运行速率可达到 50 帧/s。

参考文献［59］中阐述的是基于 GPU 的立体算法。这种立体算法可以达到每秒 4839 百万次视差评估（MDE/s）的实时处理性能。通过在匹配决策规则中使用改进版的平方差之和，以及根据可信度标准过滤计算结果，来获取高精度的运算结果。

在参考文献［60］中，Kowalczuk 和他的同事阐述了一种实时立体匹配方法，该方法通过使用双通道逼近法处理自适应支持权重聚合以及低复杂度迭代视差细化技术。基于可编程 GPU 运行平台使用 CUDA 能够实现 152.5MDE/s，对于 320 × 240 且视差水平为 32 的图像，运算速度可以达到 62 帧/s。

Richard 等人的工作[61]对 Yoon 和 Kweon[31]的算法进行了改进，并在 GPU 上运行。运行速度可以达到 14 帧/s 以上，且仍能保留原始算法的高质量。

参考文献［62］中报告的是一种可实时产生高质量结果的算法。这种算法是基于全局

能量函数最小化而建立的。分层置信传播（hierarchical belief propagation）算法可迭代优化平滑项，但过程中包括删除冗余计算，因而聚合速度很快。为了实现实时操作，作者利用了GPU 的并行优势。实验结果显示，该算法对于 320 ×240 像素自记录图像，在 16 视差级的条件下，处理速度为 16 帧/s。非遮蔽区域内，Tsukuba、Venus、Teddy 和 Cones 图像集的不匹配像素的比例分别为 1.49、0.77、8.72 和 4.61。所使用的计算机为 3GHz 电脑，GPU 为 NVIDIA GeForce 7900 GTX，且配有 512MB 显存的显卡。

此外，参考文献［63］还阐述了另一个在 GPU 运行平台上基于分层置信传播的全局立体匹配算法。该种算法使用近似视差图计算或在更高层面使用运动预估处理，来达到限制搜索空间，同时不影响所得结果精度的效果。

参考文献［54］中，Wang 等人提出了一种将高质量的结果与实时性能相结合的立体算法。该算法中，DP 与自适应聚合步骤结合使用。只有在垂直方向进行逐个像素的匹配成本聚合，从而提高了扫描线间的一致性和明显的目标边界。这项工作，如参考文献［31］所述，对于固定支持窗口内的像素，利用基于颜色和距离接近的权重分配。实时性能是由于计算机 CPU 和 GPU 并行使用实现的。这种算法可以在 16 视差级上，以 43.5 帧/s 的速度处理 320 × 240 像素图像，或是在 16 视差级上，以 9.9 帧/s 的速度处理 640 ×480 像素图像。测试系统为 3.0GHz 的电脑，配以 ATI Radeon XL1800 的 GPU。

最后，参考文献［64］阐述了基于 RDP 算法的近实时立体匹配技术。该算法可以产生半密集视差图。算法中使用两条正交 RDP 路径，沿水平和垂直两条扫描线寻找可靠视差值。因此，扫描线间的一致性得以明显加强。通过利用可编程图形硬件的计算能力优势，该算法的计算速度得到进一步提高。在英特尔奔腾 4 电脑，配以可编程 ATI Radeon 9800 XT 的 GPU 和 256MB 显存，以 3GHz 的频率测试该算法。结果显示，对于 Tsukuba 图像对，密集度为 85%，运算误差为 0.3%，运行速度为 23.8 帧/s。对于 Sawtooth 图像对，密集度为 93%，错误率为 0.24%，运行速度为 12.3 帧/s。对于 Venus 图像对，各值为 86%、0.21%、9.2 帧/s。对于 Map 图像对，各值为 88%、0.05%、20.8 帧/s。如果需要，该算法也可用于产生密集度更高的视差图，但运行速度会有所降低。

6.5.3　硬件执行（FPGA，ASIC）

使用精心设计的硬件可以真正提高立体算法的运算性能。然而，并不是所有的立体算法在硬件上都易于有效运行[16]。此外，这样的立体算法的运行所需的资源和时间也多，而且进一步的改善都很难实现。

参考文献［65］中开发的基于 FPGA 的系统可以在固定窗口上利用 SAD 方法实时计算密集视差图。整个算法包括径向畸变校正、高斯拉普拉斯（LoG）过滤、匹配搜索和视差图计算，都是在一个简单的 FPGA 上实现的。该系统可以在 64 视差级和 8 位深度精度的前提条件下，以 30 帧/s 的速度处理 640 ×480 像素图像，或以 50 帧/s 的速度处理 320 ×240 像素图像。

另一方面，参考文献［66］提出了一个比之前算法略微复杂的算法，该算法使用自适应大小窗口，基于 SAD 算法实现。该方法通过分层降低窗口大小迭代优化匹配结果。由该

方法获得的结果优于固定窗口算法 10%。该算法系统架构是完全并行的，可以同时处理所有的像素和窗口。在 8 位灰度图像精度和 64 视差级的条件下，处理 64 ×64 像素图像的速度为 30 帧/s。资源消耗为 4. 25 万个逻辑单元，相当于 82% 的 FPGA 器件。

参考文献［67］提出了算法的核心是使用自适应窗口的 SAD 聚合。在单个 FPGA 器件上，实现了基于硬件的 CA 并行流水线。对于 640 ×480 像素图像，在视差范围为 80 像素的条件下，实现了接近 275 帧/s 的运行速度。基于硬件的算法可达到较高的运行速度，同时降低了准确性。设备利用率是 83%，共使用 14. 9 万个门器件（gates）。

参考文献［52］中在 FPGA 板上实现了 SAD 算法，FPGA 板的特点为外部存储器和 Nios Ⅱ嵌入式处理器，运行速度为 100MHz。该算法在 32 视差级条件下，以 14 帧/s 的速度处理 320 ×240 像素，产生密集 8 位视差图。关键资源是大约 1. 6 万个逻辑单元，通过迁移到更复杂的设备，该算法可以升级得到更好的效果。

同样的作者，在参考文献［53］中提出了一种改进的基于平方差之和的算法，这种算法需在固定的 3 × 3 聚合窗口和硬件媒体加强滤波器的配合下使用。该系统可以在 64 视差级条件下，以 162 帧/s 的速度处理 640 ×480 像素图像。这一算法需要 3. 2 万个逻辑单元，相当于大约 6. 3 万个门器件。

参考文献［68］中的 Ambrosch 和 Kubinger 提出了局部立体算法在 FPGA 中实现，这种算法将基于递归的统计变换和基于自适应支持窗口的 SAD 相结合，该算法可以以 60 帧/s 的速度处理 750 ×400 像素图像。

Zicari 等学者在参考文献［69］中对 FPGA 实施了配以额外一致性检查的 SAD 算法，这种算法可以在 30 视差级条件下，以 97 帧/s 的速度处理 1280 ×780 的灰度图像。

此外，Kostavelis 等人[70]的工作主要包括在 FPGA 上实施基于 SAD 的密集立体算法，该算法可用作行星自主机器人的视觉系统。这种在 Xilinx Virtex 6 FPGA 器件实现的立体算法能够在 200 视差级别和 1/4 像素精度条件下，以 0. 59 帧/s 的速度处理 1120 ×1120 像素图像，研究发现这种算法远远超过了空间探测车上的需求精度。

参考文献［71］探讨了 DP 的使用，并在网格解空间上实现了使用 DP 搜索方法。它可以处理双摄像头，即光轴相交的相机。从一对摄像头得到的图像可通过使用线性内插法矫正后计算出视差值。该体系架构为线性脉动阵列式，且使用的是简单的处理单元，该设计规范简单，且易于实现并行计算。算法运行需要 208 个处理单元，产生的系统在 208 视差级条件下，以 15 帧/s 的运行速度可以处理 1280 ×1000 像素图像。

上述方法的扩展版在参考文献［72］中进行了阐述，扩展版与之前版本最主要的不同在于考虑结合从上一行得到的数据，从而更好地保证扫描线间的不一致。该算法的速度为 30 帧/s，可在 128 视差级条件下，处理 320 ×240 像素图像。使用的处理单元数量为 128。对于 Tsukuba、Map、Venus 和 Sawtooth 图像集，在遮蔽区域内视差误差大于 1% 的像素百分比分别为 2. 63%、0. 91%、3. 445% 和 1. 88%。

最后，参考文献［53］中提出了一个自定义并行 DP 算法，而且运行过程中也会使用固定 3 ×3 聚合窗口和硬件媒体增强过滤器。此外，还利用了扫描线间的支持。该系统在 65 视

差级条件下，可以以 81 帧/s 的速度处理 640 × 480 像素图像。运算平台需要 27 万个逻辑器件，相当于大约 160 万个门器件。

基于 ASIC 的立体算法可以产生非常快速的系统，甚至比使用 FPGA 时速度更快。然而，选择 ASIC 的成本更高，除非在大规模生产的情况下。原型机制造时间相当长，而且运算结果也是有高度的过程依赖性。任何的改进都很困难，而且也会费时费力。因而在多数情况下，ASIC 平台的性能优势并不能证明它比其他硬件更有利。这也是基于 ASIC 实现的立体算法文献与基于 FPGA 的文献相比少之又少的主要原因。已发表的文献认为基于 ASIC 硬件的立体匹配算法[73,74]仅限于使用平方差之和（SAD），报告中的架构广泛使用了并行运算，似乎很有前景。

6.6　结语

立体视觉仍然是解决 3D 成像问题的一个很有吸引力的解决方案。本章讨论了立体视觉算法背后的基础理论，根据各算法的主要特点和运行平台进行了分类，并对立体算法的简要现状予以介绍。

由此得出的结论是，立体视觉关注的重心似乎发生了转移，人们不再像以前那样追求精度，如今更多是对实时性能的需求。这个问题的解决方案大体分三类：一是在强大的先进 CPU 中运行简单算法；二是充分利用可编程 GPU 协同处理器；三是开发与 FPGA 相兼容的硬件运行平台。后两种的选择似乎越来越普及，因为它们可以将实时执行速度和非常精确的深度评估有效结合。另一个有趣的现象是局部算法不再是实现实时计算的唯一算法。研究表明，半全局甚至是纯粹的全局立体算法也可以达到可接受的帧速率。这种趋势似乎在不断增长，这是由于更强大的平台正在逐渐变得可用，并且更多高效优化的视觉算法也在被人提出。

立体视觉技术的成熟以及其适应室内室外环境的能力使得立体视觉技术在与其他 3D 传感技术的激烈竞争中保持地位稳固。因而，立体视觉算法的实时实现在类似现代人机交互系统、家庭娱乐系统和自主机器人等系统中占有一席之地。

参 考 文 献

1. Wheatstone, C. (1838). Contributions to the physiology of vision – part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, 371–394.
2. Faugeras, O. (1993). *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge.
3. Marr, D., Poggio, T. (1976). Cooperative computation of stereo disparity. *Science* **194**(4262), 283–287.
4. Jain, R., Kasturi, R., Schunck, B.G. (1995). *Machine vision*. McGraw-Hill, New York, USA.
5. Forsyth, D.A., Ponce, J. (2002). *Computer Vision: A modern Approach*. Prentice Hall, Upper Saddle River, NJ, USA.
6. Metta. G, Gasteratos, A., Sandini, G. (2004). Learning to track colored objects with log-polar vision. *Mechatronics* **14**(9), 989–1006.
7. Murray, D., Little, J.J. (2000). Using real-time stereo vision for mobile robot navigation. *Autonomous Robots* **8**(2), 161–171.
8. Murray, D., Jennings, C. (1997). *Stereo vision based mapping and navigation for mobile robots*. IEEE International Conference on Robotics and Automation, **2**, 1694–1699.

9. Manzotti, R., Gasteratos, A., Metta, G., Sandini, G. (2001). Disparity estimation on log-polar images and vergence control. *Computer Vision and Image Understanding* **83**(2), 97–117.
10. Barnard, S.T., Fischler, M.A. (1982). Computational stereo. *ACM Computing Surveys* **14**(4), 553–572.
11. Dhond, U.R., Aggarwal, J.K. (1989). Structure from stereo – a review. *IEEE Transactions on Systems, Man, and Cybernetics* **19**(6), 1489–1510.
12. Brown, L.G. (1992). A survey of image registration techniques. *Computing Surveys* **24**(4), 325–376.
13. Scharstein, D., Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo cor- respondence algorithms. *International Journal of Computer Vision* **47**(1–3), 7–42.
14. http://vision.middlebury.edu/stereo/ (2010).
15. Sunyoto, H., van der Mark, W., Gavrila, D.M. (2004). *A comparative study of fast dense stereo vision algorithms.* IEEE Intelligent Vehicles Symposium, 319–324.
16. Nalpantidis, L., Sirakoulis, G.C., Gasteratos, A. (2008). Review of stereo vision algorithms: from software to hardware. *International Journal of Optomechatronics* **2**(4), 435–462.
17. Tippetts, B., Lee, D.J., Lillywhite, K., Archibald, J. (2013). Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing.*
18. Hirschmuller, H., Scharstein, D. (2007). *Evaluation of cost functions for stereo matching.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota, USA.
19. Hirschmuller, H., Scharstein, D. (2009). Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(9), 1582–1599.
20. Gong, M., Yang, R., Wang, L., Gong, M. (2007). A performance study on different cost aggregation approaches used in real-time stereo matching. *International Journal of Computer Vision* **75**(2), 283–296.
21. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C. (2008). A comparative study of energy minimization methods for markov random fields with smoothness-based priors. IEEE transact. pattern anal. mach. intell. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(6), 1068–1080.
22. Muhlmann, K., Maier, D., Hesser, J., Manner, R. (2002). Calculating dense disparity maps from color stereo images, an efficient implementation. *International Journal of Computer Vision* **47**(1–3), 79–88.
23. Zitnick, C.L., Kanade, T. (2000). A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(7), 675–684.
24. Brockers, R., Hund, M., Mertsching, B. (2005). Stereo vision using cost-relaxation with 3D support regions. *Image and Vision Computing New Zealand*, 96–101.
25. Barnard, S.T., Thompson, W.B. (1980). Disparity analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**(4), 333–340.
26. Mayoral, R., Lera, G., Perez-Ilzarbe, M.J. (2006). Evaluation of correspondence errors for stereo. *Image and Vision Computing* **24**(12), 1288–1300.
27. Nalpantidis, L., Gasteratos, A. (2010). Stereo vision for robotic applications in the presence of non-ideal lighting conditions. *Image and Vision Computing* **28**, 940–951.
28. Gonzalez, R.C., Woods, R.E. (1992). *Digital Image Processing.* Addison-Wesley Longman Publishing Co, Inc, Boston, MA, USA.
29. Zabih, R., Woofill, J. (1994). *Non-parametric local transforms for computing visual correspondence.* European Conference of Computer Vision, 151–158.
30. Mordohai, P., Medioni, G.G. (2006). Stereo using monocular cues within the tensor voting framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(6), 968–982.
31. Yoon, K.-J., Kweon, I.S. (2006). Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(4), 650–656.
32. Gu, Z., Su, X., Liu, Y., Zhang, Q. (2008). Local stereo matching with adaptive support-weight, rank transform and disparity calibration. *Pattern Recognition Letters* **29**(9), 1230–1235.
33. Nalpantidis, L., Gasteratos, A. (2010). Biologically and psychophysically inspired adaptive support weights algorithm for stereo correspondence. *Robotics and Autonomous Systems* **58**, 457–464.
34. Maimone, M.W., Shafer, S.A. (1996). *A taxonomy for stereo computer vision experiments.* ECCV Workshop on Performance Characteristics of Vision Algorithms, 59–79.
35. Hirschmuller, H. (2005). *Accurate and efficient stereo processing by semi-global matching and mutual information.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 807–814.
36. Hirschmuller, H. (2006). *Stereo vision in structured environments by consistent semi-global matching.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 2386–2393.

37. Veksler, O. (2002). Dense features for semi-dense stereo correspondence. *International Journal of Computer Vision* **47**(1–3), 247–260.
38. Veksler, O. (2003). Extracting dense features for visual correspondence with graph cuts. *IEEE Computer Vision and Pattern Recognition* **1**, 689–694.
39. Mu, Y., Zhang, H., Li, J. (2009). *A global sparse stereo matching method under structure tensor constraint.* International Conference on Information Technology and Computer Science, **1**, 609–612.
40. Schauwecker, K., Klette, R., Zell, A. (2012). *A new feature detector and stereo matching method for accurate high-performance sparse stereo matching.* IEEE/RSJ International Conference on Intelligent Robots and Systems, 5171–5176.
41. Cox, I.J., Hingorani, S.L., Rao, S.B., Maggs, B.M. (1996). A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding* **63**, 542–567.
42. Liu, C., Pei, W., Niyokindi, S., Song, J., Wang, L. (2006). Micro stereo matching based on wavelet transform and projective invariance. *Measurement Science and Technology* **17**(3), 565–571.
43. Binaghi, E., Gallo, I., Marino, G., Raspanti, M. (2004). Neural adaptive stereo matching. *Pattern Recognition Letters* **25**(15), 1743–1758.
44. Kotoulas, L., Gasteratos, A., Sirakoulis, G.C., Georgoulas, C., Andreadis, I. (2005). *enhancement of fast acquired disparity maps using a 1-D cellular automation filter.* IASTED International Conference on Visualization, Imaging and Image Processing, Benidorm, Spain, 355–359.
45. Ogale, A.S., Aloimonos, Y. (2005). *Robust contrast invariant stereo correspondence.* IEEE International Conference on Robotics and Automation, 819–824.
46. Ogale, A.S., Aloimonos, Y. (2005). Shape and the stereo correspondence problem. *International Journal of Computer Vision* **65**(3), 147–162.
47. Ogale, A.S., Aloimonos, Y. (2007). A roadmap to the integration of early visual modules. *International Journal of Computer Vision* **72**(1), 9–25.
48. Boykov, Y., Veksler, O., Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(11), 1222–1239.
49. Boykov, Y., Kolmogorov, V. (2001). An experimental comparison of min-cut/max-ow algo- rithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 359–374.
50. Yedidia, J.S., Freeman, W., Weiss, Y. (2000). *Generalized belief propagation.* Conference on Neural Information Processing Systems, 689–695.
51. Bobick, A.F., Intille, S.S. (1999). Large occlusion stereo. *International Journal of Computer Vision* **33**, 181–200.
52. Kalomiros, J.A., Lygouras, J. (2008). Hardware implementation of a stereo co-processor in a medium-scale field programmable gate array. *IET Computers and Digital Techniques* **2**(5), 336–346.
53. Kalomiros, J., Lygouras, J. (2009). Comparative study of local SAD and dynamic programming for stereo processing using dedicated hardware. *EURASIP Journal on Advances in Signal Processing* 1–189.
54. Wang, L., Liao, M., Gong, M., Yang, R., Nister, D. (2006). *High-quality real-time stereo using adaptive cost aggregation and dynamic programming.* Third International Symposium on 3D Data Processing, Visualization, and Transmission, 798–805.
55. Gong, M., Yang, Y.-H. (2005). Fast unambiguous stereo matching using reliability-based dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(6), 998–1003.
56. Tombari, F., Mattoccia, S., Di Stefano, L., Addimanda, E. (2008). *Near real-time stereo based on effective cost aggregation.* International Conference on Pattern Recognition, 1–4.
57. Gupta, R.K., Cho, S.-Y. (2010). *A correlation-based approach for real-time stereo matching.* International Symposium on Visual Computing, **6454**, 129–138.
58. Zach, C., Karner, K., Bischof, H. (2004). *Hierarchical disparity estimation with programmable 3D hardware.* International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, 275–282.
59. Drazic, V., Sabater, N. (2012). *A precise real-time stereo algorithm.* 27th Conference on Image and Vision Computing New Zealand, 138–142.
60. Kowalczuk, J., Psota, E.T., Perez, L.C. (2013). Real-time stereo matching on cuda using an iterative refinement method for adaptive support-weight correspondences. *IEEE Transactions on Circuits and Systems for Video Technology* **23**(1), 94–104.
61. Richardt, C., Orr, D., Davies, I., Criminisi, A., Dodgson, N. (2010). A. *Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid.* European Conference on Computer Vision (ECCV), ser. Lecture Notes in Computer Science, **6313**, 510–523.
62. Yang, Q., Wang, L., Yang, R. (2006). *Real-time global stereo matching using hierarchical belief propagation.* British Machine Vision Conference, **3**, 989–998.

63. Grauer-Gray, S., Kambhamettu, C. (2009). *Hierarchical belief propagation to reduce search space using cuda for stereo and motion estimation.* Workshop on Applications of Computer Vision, 1–8.
64. Gong, M., Yang, Y.-H. (2005). *Near real-time reliable stereo matching using programmable graphics hardware.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **1**, 924–931.
65. Jia, Y., Xu, Y., Liu, W., Yang, C., Zhu, Y., Zhang, X., An, L. (2003). A miniature stereo vision machine for real-time dense depth mapping. International Conference on Computer Vision Systems, ser. Lecture Notes in Computer Science, **2626**, 268–277.
66. Hariyama, M., Kobayashi, Y., Sasaki, H., Kameyama, M. (2005). FPGA implementation of a stereo matching processor based on window-parallel-and-pixel-parallel architecture. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science* **88**(12), 3516–3522.
67. Georgoulas, C., Kotoulas, L., Sirakoulis, G.C., Andreadis, I., Gasteratos, A. (2008). Real-time disparity map computation module *Journal of Microprocessors and Microsystems* **32**(3), 159–170.
68. Ambrosch, K., Kubinger, W. (2010). Accurate hardware-based stereo vision. *Computer Vision and Image Understanding* **114**(11), 1303–1316.
69. Zicari, P., Perri, S., Corsonello, P., Cocorullo, G. (2012). Low-cost FPGA stereo vision system for real time disparity maps calculation. *Microprocessors and Microsystems* **36**(4), 281–288.
70. Kostavelis, I., Nalpantidis, L., Boukas, E., Rodrigalvarez, M., Stamoulias, I., Lentaris, G., Diamantopoulos, D., Siozios, K., Soudris, D., Gasteratos, A. (in press). SPARTAN: Developing a vision system for future autonomous space exploration robots. *Journal of Field Robotics.*
71. Jeong, H., Park, S. (2004). Generalized trellis stereo matching with systolic array. International Symposium on Parallel and Distributed Processing and Applications, **3358**, 263–267. Springer Verlag.
72. Park, S., Jeong, H. (2007). *Real-time stereo vision FPGA chip with low error rate.* International Conference on Multimedia and Ubiquitous Engineering, 751–756.
73. Hariyama, M., Takeuchi, T., Kameyama, M. (2000). *Reliable stereo matching for highly-safe intelligent vehicles and its VLSI implementation.* IEEE Intelligent Vehicles Symposium, 128–133.
74. Hariyama, M., Sasaki, H., Kameyama, M. (2005). Architecture of a stereo matching VLSI processor based on hierarchically parallel memory access. *IEICE Transactions on Information and Systems* **E88-D**(7), 1486–1491.

第7章

飞行时间法3D成像技术

Daniel Van Nieuwenhove
比利时 SoftKinetic 传感器公司

7.1 引言

在过去十年，实时 3D 用户交互技术推动了新应用程序的不断开发，人们越来越多地意识到 3D 独特的成像优势。人们熟知的大多数 3D 传感技术包括立体视觉、飞行时间法（TOF）以及结构光方法。本章将论述飞行时间法技术。

本章的前面部分我们将会详细介绍这种技术，以此来区分不同类型的 3D 成像技术，例如，脉冲飞行时间法和持续飞行时间法。然后，我们将会介绍操作原则和主要方程式，并论述这些原理的精准性。最后，我们会探讨存在的挑战和有待改进之处，一些摄像系统的典型性能价值，以及对当前全球在分辨率方面的尖端研究。

7.2 飞行时间法 3D 传感

就在最近，各种飞行时间法（TOF）3D 成像技术已经证实了其在更广泛的 3D 应用中的可靠性[1-3]。大体来说，在所有的 TOF 3D 成像方案中，均有一束调制光波投射在背景上，其反射可以被检测到并用来确定光波的往返时间和距离[4]。通过把反射光线聚焦在像素矩阵上，完整的深度图像会立即呈现。该方法存在的挑战是，感光范围和动态范围需要在比光波高达几个数量级的环境光的存在下，依然能测量微弱的反射信号。在这方面的一些研究已经开展[2,5]。

在飞行时间法的计算中，距离是通过测量光线在光源和目的地之间的往返时间得出的。该往返时间通过与光速的乘积被转换成距离：

$$c = 3 \times 10^8 \text{m/s} = 2 \times 150 \text{m/}\mu\text{s} = 2 \times 0.15 \text{m/ns} = 2 \times 0.15 \text{mm/ps}$$

该光学雷达技术最早曾在光探测和测距（LIDAR）设备中实施过，这个装备应用了带有

单点探测器的激光器来获取 TOF 或距离。扫描背景可以形成一个完整的 3D 图像。这个曾经乃至现在仍被应用于多个领域，并且数年来得到了很大程度的完善，但扫描过程导致速度非常缓慢。而且该设备昂贵易损，在扫描景象时需要机械移动各部件。

近年来，由于集成电路技术的改进，基于飞行时间法原理，建立微型飞行时间法传感器的矩阵已经可行[6,7]。这样就可以制作成完整的 3D 摄像系统。在这些系统中，通常使用发光二极管（LED），整体的景象立刻会被照明。而反射光线将会聚集在 TOF 探测器的阵列中（见图 7.1）。每个探测器同时检测到一点的距离，从而立刻就会获得一个完整成像的范围信息。为了把错误率降到最低，活跃光源和接收器会被布置在相距很近的地方。这样能使设备紧凑排放并避免了遮蔽效应。

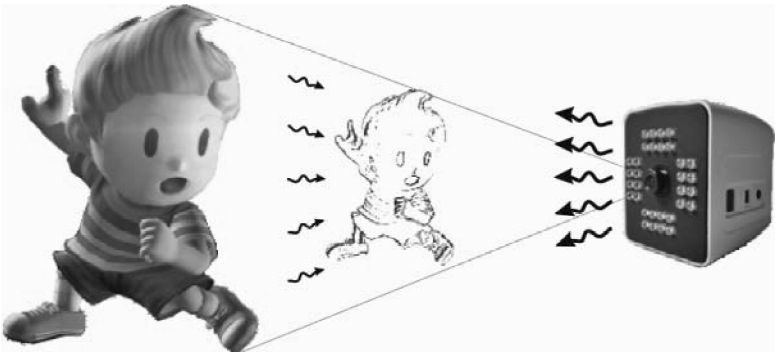


图 7.1　飞行时间法范围成像技术。该主动光照反射了在整个场景前形成的光波，包括飞行时间法所需信息以推断出场景中所有点的距离

这种被称为间接或持续飞行时间法的技术因为速度快和稳定性高而具有优越性。与其他飞行时间法技术相比，其因为无需移动部件而受到欢迎，具有良好的发展前景。根据参考文献［8］论述，间接飞行时间法 3D 摄像将会自然取代现存相应的 2D 视觉技术。基本操作原则在图 7.2 说明。该技术可进一步分成脉冲式和持续式两种。

本章接下来将简短地综述脉冲飞行时间法技术，随后重点讨论持续飞行时间法技术。因此我们将此概括地称为“飞行时间法”（TOF）。如果提到脉冲飞行时间法，我们会明确说明。

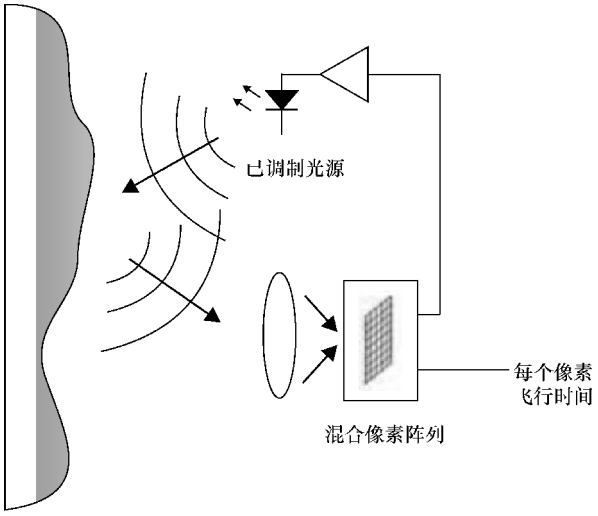


图 7.2　飞行时间法原理图。每个像素中，距离是通过测量每个像素的飞行时间获得的，然后除以 2 再乘以光速 c

7.3　脉冲飞行时间法

在飞行时间法测距法中，光脉冲会投射在背景上，与此同时高精度的秒表开始测量每个像素所用的飞行时间。光脉冲需到达目的地然后返回原点。一旦发现光脉冲返回，像素里的一个机制将会终止秒表记时，这样就会显示出光脉冲飞行时间。

由于光脉冲行经了两次路线（往返），测量的每 6.67ns 的时间对应 1m 的距离。一个精准的时间要比 1mm 所需的 7ps 测试的效果要好得多。可以通过重复测量所需的次数以及求所得结果的平均值来提高精准度。

飞行时间法的最大的缺陷是，在接收一端必须同时有高动态范围和大带宽。在运用这种技术时，让接收路径精准地探测反向散射的光脉冲是很困难的，原因如下：

1）光学阈值不是一个固定的值，而是会因为物体的背景和距离以及目标反射率而改变。

2）大气衰弱导致光脉冲散布，并且使接收到的脉冲斜坡变平。因此，一个大功率的脉冲光源是非常必要的。

另一方面，除了需要发射大功率，光源还需要能够生成快上升超短脉冲光，这对于确保入射的脉冲光的测量精度是很有必要的。最近市场推出的激光器或激光二极管是唯一能提供短脉冲宽度的具有高功率的光学元件。它通常能在 10Hz 不断重复脉冲。其较低的重复率极大地限制了脉冲 TOF 系统的帧频[9]。

7.4　持续飞行时间法

我们将开始讨论持续飞行时间法，相较于前面所述的运用单一脉冲测距，该方法发射的是连续调制光。这种持续性测量方案能生成更高的信噪比，且使用较少的峰值功率，从而对光源要求不高。更典型的方法是使用重复的脉冲波或正弦波调制。

这个技术的优点是对于带宽和功率的光源要求非常低以及较高的信噪比和可配置性。对长距离的测量将会导致更高的信噪比并且反之亦然。这是在距离精度和图像刷新率之间一个良好的权衡。这个系统是稳固的，因为它不包含任何可移动的部件。它在视觉方面很安全，因为它依靠散布的而非校准的光线。它本身的光源可以是 LED 或激光器，其中激光器可以产生更快的调制频率。因为这个技术会立刻抓取完整的图像，它可以实时进行操作，而且能很容易地达到 200Hz 以上的帧速率。除此以外，这个系统不包含特别昂贵的各种组件，一个低成本的 3D 相机就可以完成这个任务。另外，图像传感器芯片的输出可以通过一些简单的公式换算为深度图，因此无需繁琐的后续过程来完成这一目标。

这个技术的不足之处是距离的计算是模糊不清的，因为测量的目标要比离相机可见目标返回的距离范围远很多，这使得测量目标看起来比实际要更近一些。在大多数情况下，可见距离是由调制频率或脉冲率决定的。比如，一个典型的 20MHz 调制波可以产生 7.5m 的可见

距离范围。在测量距离中存在的模糊性通常被称为混叠现象。

7.5　计算方法

在本章中，我们将会推导一些常用的 TOF 公式。本章假定正弦调制，类似的公式也可以通过其他调制波获得。飞行时间法 t_d 可以表示成介于发送和接收调制信号的相位差 α：

$$\alpha = t_d \omega \tag{7.1}$$

式中，ω 是调制的角频率。飞行时间法的目标就是要求出这个相位差。时间延迟和距离可以用下面这个公式获得：

$$\text{distance} = t_d c = \frac{\alpha}{\omega} c = \frac{\alpha}{2\pi f_{ML}} c \tag{7.2}$$

式中，f_{ML}是调制频率（如 20MHz）；c 是光速度，为 $3 \times 10^8 \text{m/s}$。为了简便，我们假定空气折射率应该为 1。如图 7.3 所示，通过测量同相（I）和正交（Q）所要求的相位差参数，我们就可以在这个公式中找到唯一不可知数，就是相位差 α。相位差可以通过以下公式获得：

$$\alpha = \arctan\left(\frac{Q}{I}\right) \tag{7.3}$$

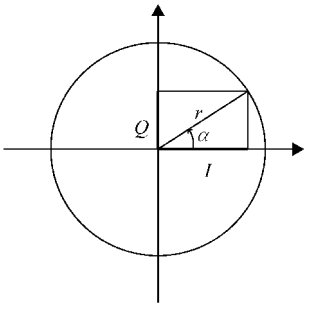


图 7.3　I、Q、α 之间关系的测角圆圈，用来测算基于正弦调制信号的时差距离

图 7.4 所示为一个连续时差测距的典型信号路径。首先，调制光发射入场景中，随后其反射被聚焦在探测器节点，转化成电流信号。电流信号是由在 0°和 180°的原始调制信号的各相移混合产生。随后产生的信号逐渐积分并且彼此缩减。在本章的后面我们会详细地介绍，其整个过程需要进行两次，分别使用 0°和 90°的相移调制光来获得电压信号，即为 I。

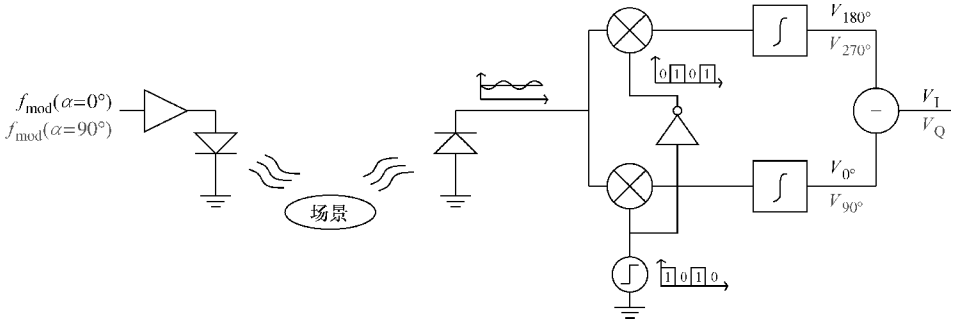


图 7.4　使用持续飞行时间法原理获得距离测算的过程包含了诸多不同组件。LED 从左边发射光到场景中，在右端反射后被检测到。可知，它与原始调制信号的 0°和 180°的各种相移发生了混合，并随着时间的推移进行了积分，缩减成对应的电压 V_I 或 V_Q（$= V_{0°} - V_{180°}$）。随后，90°的相移调制照明被用来生成一个电压信号，即为 Q（$= V_{180°} - V_{270°}$）

我们现在详尽地研究一下该方法的数学计算。在图 7.5 中，上图代表探测信号电流幅值和时间。该信号有一个背景光组件，为了简化，假定该组件随着时间变化保持恒定。一个调制光组件如下所示：

$$I_{\text{det}} = I_{\text{BL}} + I_{\text{ML}}\sin(\omega_{\text{ML}}t + \alpha) \quad (7.4)$$

式中，ω_{ML}为 $2\pi f_{\text{ML}}$；I_{BL}对应背景光产生的电流；I_{ML}对应调制光产生的电流。

为了求出 α，需要用信号乘以方波（见图 7.5 的中图）。一次乘以与发送的基本信号同相的方波，一次乘以与基本信号异相 180°的方波。通过这个方法，我们把周期电流分成了两个部分，如图 7.5 下图所示。各个部分随着时间进行积分，从而产生相应的电压：

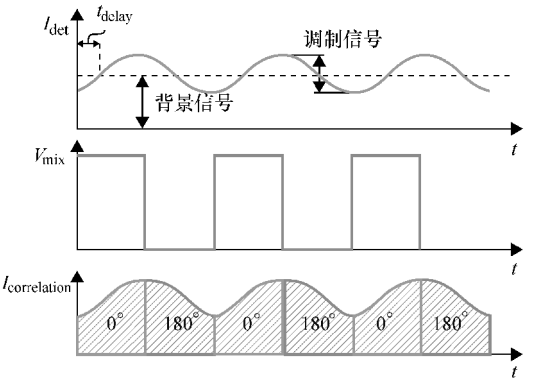


图 7.5　持续飞行时间法的原理。上图显示出探测器电流与时间的关系，正弦调制信号上面有一个 DC 背景信号组件。中图显示应用于混合器的电压与时间的关系。下图代表到达每个混合器输出的电荷，有助于 0°和 180°相移的相关测量

$$V_{0^\circ} = \frac{1}{C_{\text{int}}}\int_0^{t_{\text{int}}} I_{\text{det}}K(t)\,\mathrm{d}t \quad (7.5)$$

式中，$K(t)$ 描述的是由混合器执行的方波乘法：

$$\begin{aligned} K(t) &= 1, \text{ 对于 } (n-1)T < t < n\frac{T}{2} \\ &= 0, \text{对于 } n\frac{T}{2} < t < nT, \text{ 其中 } n \in N \end{aligned} \quad (7.6)$$

假设 $t_{\text{int}} = zT,\ z \in N$：

$$\begin{aligned} \Rightarrow V_{0^\circ} &= \frac{1}{2C_{\text{int}}}\left(t_{\text{int}}I_{\text{BL}} + \frac{t_{\text{int}}}{T}I_{\text{ML}}\int_0^{\frac{T}{2}}\sin(\omega_{\text{ML}}t + \alpha)\,\mathrm{d}t\right) \\ &= \frac{t_{\text{int}}}{2C_{\text{int}}}\left(I_{\text{BL}} + \frac{I_{\text{ML}}}{\omega_{\text{ML}}T}(-\cos(\pi + \alpha) + \cos(0 + \alpha))\right) \\ &= \frac{t_{\text{int}}}{2C_{\text{int}}}\left(I_{\text{BL}} + I_{\text{ML}}\frac{\cos\alpha}{\pi}\right) \\ &= \frac{V_{\text{BL}}}{2} + \frac{V_{\text{ML}}}{2\pi}\cos\alpha \end{aligned} \quad (7.7)$$

以同样的方式，我们可以获得 V_{180°的表达式：

$$V_{180^\circ} = \frac{V_{\text{BL}}}{2} - \frac{V_{\text{ML}}}{2\pi}\cos\alpha \quad (7.8)$$

用式（7.7）减去式（7.8），我们可以得到背景水平的独立测量 I 值：

$$V_{0^\circ} - V_{180^\circ} = \frac{V_{\text{ML}}}{\pi}\cos\alpha \propto I \quad (7.9)$$

继续测量循环，与 90°和 270°相移信号混合，我们可以获得一个与 Q 有一定比例的值：

$$V_{90^\circ} - V_{270^\circ} = \frac{V_{\mathrm{ML}}}{\pi}\sin\alpha \propto Q \tag{7.10}$$

于是可以从下面的公式求得想要的相位延迟：

$$\alpha = \arctan\left(\frac{V_{90^\circ} - V_{270^\circ}}{V_{0^\circ} - V_{180^\circ}}\right) \tag{7.11}$$

连同式（7.2），我们可以使用持续飞行时间法找到一个通用的表达式求得距离：

$$\mathrm{distance} = \frac{c}{2\pi f_{\mathrm{ML}}}\arctan\left(\frac{V_{90^\circ} - V_{270^\circ}}{V_{0^\circ} - V_{180^\circ}}\right) \tag{7.12}$$

上述我们可以推断出为了获得距离估算，需要用到四个测量数据，即为 V_{0°、V_{90°、V_{180°、V_{270°。需要注意的是，当 V_{0° 和 V_{180° 近乎相等的时候，式（7.12）会得出较差的结果。因为在这种情况下，分母会很小。解决这个难题的办法就是要运用反余切法而不是反正切法，于是我们可以从一个大分母中获得一个小指数。

7.6　精度

对最终深度精度产生影响的不同噪声成分进行综述非常重要。这个是通过研究随机噪声对测量 V_{0°、V_{90°、V_{180°、V_{270°关于相位错误 $\delta\alpha$ 产生的影响而完成的。通过式（7.12），并且运用错误传播的规律，我们可以得到一个关于 $\delta\alpha$ 的通用表达式：

$$\sqrt{\left(\frac{\delta\alpha}{\delta V_{0^\circ}}\right)^2 \Delta^2 V_{0^\circ} + \left(\frac{\delta\alpha}{\delta V_{90^\circ}}\right)^2 \Delta^2 V_{90^\circ} + \left(\frac{\delta\alpha}{\delta V_{180^\circ}}\right)^2 \Delta^2 V_{180^\circ} + \left(\frac{\delta\alpha}{\delta V_{270^\circ}}\right)^2 \Delta^2 V_{270^\circ}} \tag{7.13}$$

我们可以用这个公式解决这些特殊的相位值，例如，0°、45°、90°、135°和 180°，这样就可以得到[6]：

$$\delta D_{\mathrm{f}} = \frac{D_{\mathrm{u}}}{2\pi}\delta\alpha = \frac{D_{\mathrm{u}}}{\sqrt{8}}\frac{\sqrt{B}}{2A} = \frac{D_{\mathrm{u}}}{\sqrt{8}}\frac{1}{2\mathrm{SNR}'} \tag{7.14}$$

式中，D_{u} 是一个由调制频率决定的模糊距离，A 是光电子数量的调制信号产生的幅值，$\sqrt{B}$ 对应的是由背景光散射噪声产生的光电子数量。值得注意的是，这个情况下信噪比（SNR）没有以 dB 表示。代入公式 $D_{\mathrm{u}} = \frac{c}{2f_{\mathrm{ML}}}$，我们可以得到：

$$\delta D_{\mathrm{f}} = \frac{c}{2f_{\mathrm{ML}}}\frac{1}{\sqrt{8}}\frac{1}{2\mathrm{SNR}} \tag{7.15}$$

我们看到，两个参数影响相机系统的精度——信噪比和调制频率。最大化这些将导致最佳的相机精度。在大多数情况下，噪声性能受到不可避免的散粒噪声的限制，因此我们从中得到结论：提高系统性能的一个关键方法是优化信号幅度并使用高调制频率。光源和像素都需要能够处理这些更高的频率。通常 LED 可以支持高达几十 MHz（例如 20MHz），其中激光器倾向于支持高达几百 MHz。

7.7　局限性与改进

7.7.1　时差测距的挑战

时差测距技术具有很多优势，并且满足了很多市场需求，但是这个系统还是需要克服很多挑战。出于实际考虑，我们在图 7.6 中对这些参数进行了简短的综述，图中展现出 TOF“蜘蛛网”。

为了实现高精度，需要在时差测距系统中检测到极小的时移。由于光波传播速度为 $3 \times 10^8 m/s$，则每隔 15cm 的距离将对应 1ns 的往返时间。因此，为了实现毫米深度分辨率的测量，我们需要能够区分调制光波所用 ps 的飞行时间。

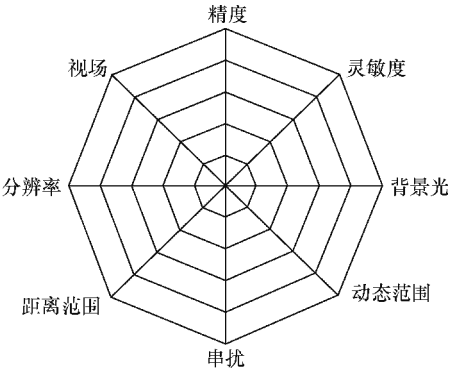


图 7.6　飞行时间法参数“蜘蛛网”

因为相机的功率预算有限，为了能用现有光观察到尽可能远，相机传感器的敏感性同样也是非常重要的。此外，现场的背景光光强度需要达到若干数量级，并且旁边有来自相机光源的调制光。在这种情况下，必须避免额外的噪声和/或饱和度以防造成信息的丢失。除了面对这些挑战，摄像系统的动态范围需要进行优化，这样所有的物体，无论距离是近或远都可以进行测量。

同时，非常重要的一点是避免双方间的串扰，必须考虑到像素间和相机间的串扰。前者是由检测到的迁移到相邻像素红外光子引发的，后者是受到了光信号在使用一个以上的相机时产生的干扰，从而照亮了同一个场景。

另一个挑战是优化距离范围。正如前面简单讨论过的，因为发射调制信号而发生混叠现象会带来限制。我们一定要确保循环的相位作为 2π 的倍数。

最后同样重要的一点是，在解决上述提到的所有问题中，我们必须确保仍然能够获得小像素来构建高分辨率的 3D 像素阵列，从而可能获得细致的 3D 图像。

光学设计同样也对时差测距系统有影响，其中有很多重要参数，比如视场（FOV）和镜头特性（F #，失真，等等）。

在系统中，这些问题需要在一个或者多个水平层次中解决，比如像素的设计、成像器的设计、控制逻辑和应用程序（见图 7.7）。

7.7.2　理论局限

时差测距系统的局限可归因于多样化的部件（例如，在 LED 中可用的速度和强度的局限）。但是在本节中，我们要找出系统可实现的理论最大精度。在硅成像的各个噪声源中，光子散射噪声是不可避免的。这种噪声是由离散电荷载流子的统计学波动造成的，并且被准确地

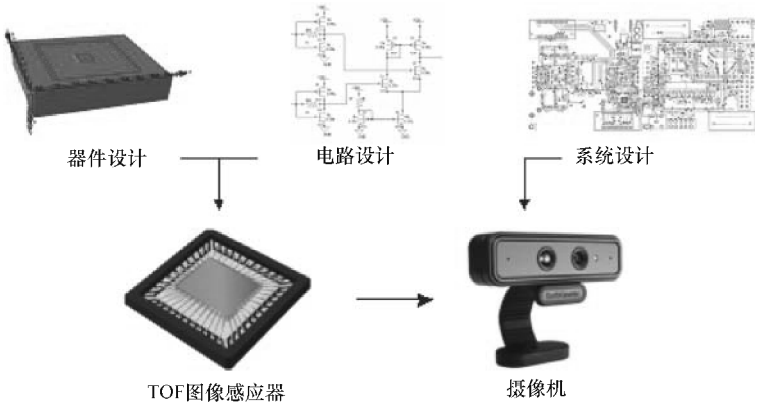


图 7.7　时差测距系统的示意图

建模成泊松过程（Poisson process）。因此，噪声振幅被定义为光生载流子数量 Q 的平方根：

$$\delta Q_{\text{xhot}} = \sqrt{Q} \tag{7.16}$$

该噪声源是设计时差测距系统的关键，我们的目标一直是使图像电路和相机产生的噪声降低或等于散射噪声，从而使系统的整体噪声性能至少达到 70% 的理论最大值。

7.7.3　距离混叠

在持续飞行时间法范围内，飞行时间被转化成一个相位差。因此，一个物体在相对略多于一个周期相位延迟的距离将被测量，结果发现定位会更近。这个范围由波长的一半定义，称为“不模糊范围”或者“混叠限制”（aliasing limit）。我们可以用很多方法来解决和改进这种与模糊问题有关的持续飞行时间法。下面将在本节讨论这些方法。

能够改进不模糊范围的第一个方法是使用稍有差异的调制频率进行两次测量。结合两次测量，这个不模糊距离成为每个频率所定义的最大距离时间间隔的最小公倍数。该结果能够运用数学方法获得：

$$\begin{aligned} D_1 &= \left(\frac{\alpha_1}{360} + n_1\right)\frac{c}{f_1 2} \\ D_2 &= \left(\frac{\alpha_2}{360} + n_2\right)\frac{c}{f_2 2} \Rightarrow D = \left(\frac{\alpha_1}{360} + n_1\right)\frac{c}{f_1 2} \\ D_1 &= D_2 \qquad\qquad \left(\frac{\alpha_1}{360} + n_1\right)f_2 = \left(\frac{\alpha_2}{360} + n_2\right)f_1 \end{aligned} \tag{7.17}$$

另一个可以规避混叠限制的方法是持续性伪噪声调制法。在这个方法中，一个有限长度的字作为调制基础。如果选中的字有自相关，如图 7.8 所示，第一个移位为非零，其他的为零，不模糊距离即明确的间距被扩展然后乘以字的位长度。

这个方法的缺点就是高频带宽度是必需的要素。伪噪声代码可能存储在只读存储器中，甚至通过使用线性移位反馈寄存器生成。这些都是从标准的数字单元中建立的，像逆变器和触发器，使得代码非常适合用芯片或可编程逻辑电路生成。

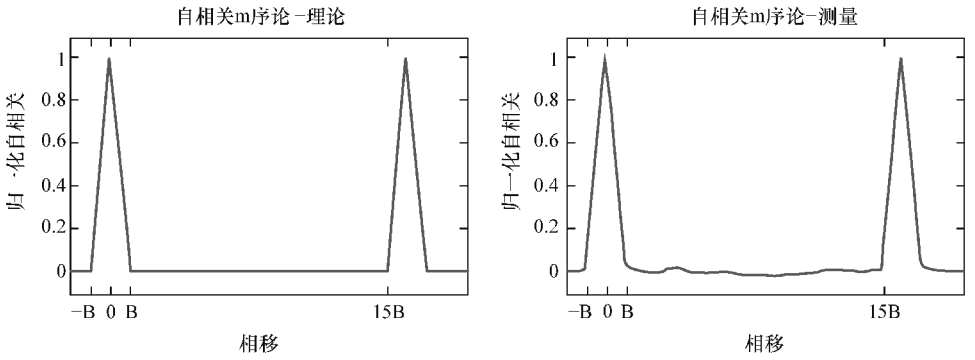


图 7.8　选定的伪噪声序列的自动校正图。模拟（左），测量（右）

7.7.4　多径与散射

时差测距成像方法的另一个副作用是多径和散射效应。

多径效应是由多元的、直接和间接返回的多条路径引发的，其中一束光能够在它进入测距相机的光学系统之前跟随这些路径。传感器的输出是不同路径距离的加权平均及其强度。

散射具有类似的效应，不过在镜头内规模较小。由于传感器表面的吸收性并非 100%，光线将会进行表面弹跳，然后部分反射在镜头上，并且可能重新进入一个不同像素位置上的传感器阵列。

这两种效果都将导致测距出现误差，特别是对于接收到来自现场的较弱的返回信号的像素而言。在本章写作的过程中，各研究中心正着力研究解决方案来补偿或消除这些效应。

7.7.5　功率分配与优化

3D 成像应用程序中的主要部分是嵌入式或电池供电的各个装置，这常常要求电源能够实现最优分配，实现各部件耗电最小化。在时差测距系统中，功率主要是由负责在一定条件下的深度测量质量的照明单元所消耗的。对所需照明功率有作用的主要参数是调制频率和调制对比度。因此，我们定义了一个特定的度量单位，称之为调制效率（ME）。如式（7.18）所示，它被定义为这个频率下的调制频率和解调对比度的乘积。

$$\text{Modulation Efficiency(ME)} = f_{\text{mod}} \times C_{\text{demod}} \tag{7.18}$$

该度量数值受系统设计的各个方面影响，比如照明光源的明暗和调制信号的质量，作为时差测距传感器的本机属性。图 7.9 所示为不同 ME 值对应的一个长范围时差测距系统的功耗图。像素间距被假定为 10μm，且其他参数保持恒定。

图 7.9 中的曲线分别代表 2cm、1cm 和 5mm 的噪声目标。噪声越低，能够检测到的移动就越精确。从图中可以看出时差测距系统的功耗在很大程度上取决于调制效率。在今天的系统中，大约 70～100 的值都是可以实现的（取决于供应商），要求 4～20W 功率量，与检测的手或手指的水平移动所需要达到的精度有关。这个参数的不断改进使得总功耗低于 1W。因此，在未来，时差测距系统将具有足够低的功耗，成为日常生活所用的各种电池供

电设备的一部分，比如笔记本电脑、平板电脑或者移动设备等。

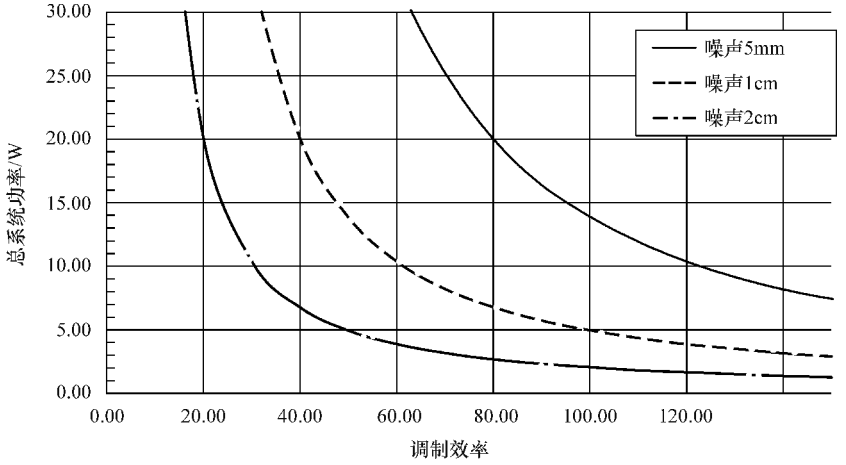


图 7.9　功耗与三种不同的噪声目标（2cm、1cm、5mm）的调制效率。系统配置参数：距离为 4m，时差测距传感器的分辨率为 VGA，像素间距为 10μm，FOV = 70° × 50°（H × V），F#镜头为 1，反射率为 50%

7.8　飞行时间法摄像组件

在飞行时间法摄像系统有许多不同的组件需要非常协调的运作，比如数字、模拟和光学。

光学元件决定了对于正常成像拥有的相同属性，比如捕获光的分配和视野。由于飞行时间法成像器通常比传统的成像传感器的像素数低得多（如 10k 像素），光学要求也在一定程度上较低。但是为了降低光照所需的光学功率，“快速”低值的 F#柔性焦距透镜组是理想的。同时必须要优化照明以发射高达几百兆赫兹的高频率波。此外，光照角度需要与成像器的视野相匹配。

摄像的核心就是时差测距成像的传感器芯片，它能够使景象中调制光的反射聚焦并转化为深度信息。成像器输出被数字化并通过使用通信协议（例如 USB）而最终发送到外部世界。

此外，还需要一些数字逻辑把所需的混合器和调制信号发送到飞行时间法成像器芯片和照明板。

7.9　标准值

在本节中，我们将呈现一些持续飞行时间法 3D 成像系统内关键参数的标准值，比如光功率、检测器电流和背景光水平。

7.9.1　光的功率范围

本节我们将计算一些影响每个像素的光功率分配的标准值。如图 7.10 所示，如果我们

假定场景中物体有朗伯反射，反射强度与 $\cos\theta$ 成正比。通过计算直径为 D 的球冠的余弦加权积分（即进入镜头时光学功率的份额）与整个半球的余弦加权积分的比值，我们就能得出接触镜头的光学功率[9]：

$$P_{\text{lens}} = P_{\text{sent}} \frac{\int_0^{2\pi}\int_0^{\theta_c} \sin\theta\cos\theta \mathrm{d}\theta\mathrm{d}\alpha}{\int_0^{2\pi}\int_0^{\tau} \sin\theta\cos\theta \mathrm{d}\theta\mathrm{d}\alpha} = P_{\text{sent}}(\sin\theta_c)^2 = P_{\text{sent}}\left(\frac{D}{2R}\right)^2 \tag{7.19}$$

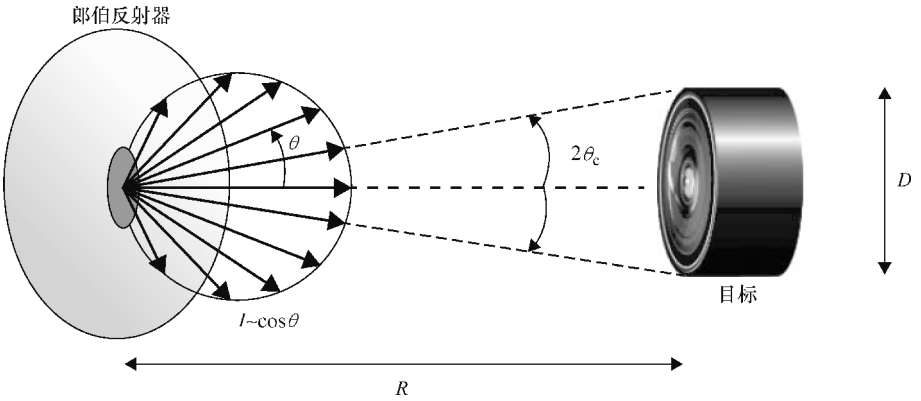


图 7.10　朗伯反射原理，其中反射强度与反射角度之间呈余弦关系。大多数日常生活物体都有一个由郎伯反射支配的表面[9]

然后我们可以计算入射到每一个像素上的光功率：

$$P_{\text{pixel}} = \frac{P_{\text{lens}}}{\#\text{pixels}} S_{\text{R}} \cdot v_{\text{total}} \tag{7.20}$$

式中，S_{R} 是表面反射率；v_{total} 是填充因数（通常为 0.7）与镜头和滤光器效率（通常为 0.9）的乘积。入射到像素的光功率因此取决于到物体的距离 R 和镜头光圈 D。

作为一个典型的例子，我们可以考虑一下发射光的光功率为 300mW 和镜头光圈为 2mm 的情形。如果我们进一步考虑一个标准的 0.3A/W 的传感器响应率、一个 70% 的填充因数、一个 50% 的标准反射率和 10k 的像素数，我们就可以像表 7.1 所示那样计算出不同距离的由检测器导致的电流值。当使用这个配置的时候，对于 100% 反射在 0.5m 的最大可能电流值为 23pA。这个值定义了这个系统所需的动态范围上限。

表 7.1　光学功率预算与检测器电流的标准值

距离/m	P_{lens}/nW	P_{pixel}/pW	I_{det}/pA
0.5	1200	38	11
1	300	9.5	2.8
2.5	48	1.5	0.45
5	12	0.38	0.11
10	3	0.095	0.028
20	0.75	0.024	0.0071

7.9.2　背景光

使用锁定原理，我们可以区别在场景中呈现的背景光与调制光。然而，因为所有光都与检测器接触，所以诱发的散射噪声水平是由调制光和背景光共同引起的。因此，相机在强的环境光下总是要么需要更多功率，要么表现出更差的性能。

能够降低入射到检测器上的光功率的一个简单方法是通过使用一个光学过滤器，以便来选择所使用的波长（在某一个光谱带宽范围内）并且减弱其他所有波长的光。使用这样的过滤器，背景光通常可以衰减 20 倍。

图 7.11 展示了太阳的光谱。我们可以看到在 930nm 周围，由于大气吸收效应使光谱显示出了一个局部最小值。因此，这将是一个在飞行时间法 3D 摄像机中所使用光源的一个不错的波长。

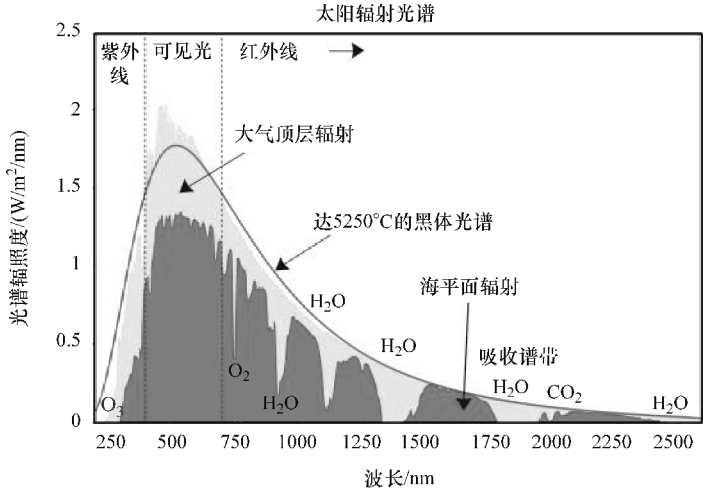


图 7.11　大气内部和大气外部太阳能光谱[10]。来源：来自 Nick84［CC－BY－SA－3.0（http：//creativecommons.org/licenses/by－sa/3.0）］，维基共享资源

通过美国材料试验学会（ASTM E 490，AM 0）公布的结果[10]，我们发现对于 250～1100nm 区域的太阳光功率密度的最大值是 1006.9W/m^2。表 7.2 已经显示了在不同条件下的光功率密度。通过使用一项纽波特 OPM840 光学功率测量单元可以获得室内和多云室外值。同样的，我们注意到一个标准的硅检测器中的检测器电流，参数包括 0.3A/W 响应率、70% 填充因数和 30μm×30μm 的面积，同时与标准值数量为 2f 的镜头连接。并且，我们假定场景反射达到最大值（100%）。为了获得这些值，我们重新改进式（7.19）和式（7.20）如下：

$$P_{\text{pixel}} = \frac{\text{PD}_{\text{BL}}}{2M_{\text{f}}} A_{\text{pixel}} \left(\frac{1}{2 \cdot f/\#} \right)^2 S_{\text{R}} \cdot V_{\text{total}} \tag{7.21}$$

式中，PD_{BL}是背景光的功率密度，并且被 2 整除，因为 DC 光以标准的 TOF 像素传播到不同

的微分检测器节点。$f/\#$是镜头使用的数量，A_{pixel}是像素面积，M_f 是光学过滤器的衰减系数，使用标准值 20。还要注意一点，这个公式是距离无关的。表 7.2 给出了背景光诱导的标准电流值。如表 7.1 所示，通过把它们和由调制光反射获得的值进行对比，我们可以看到诱导的背景光电流值几乎高达调制光的 5 个数量级。

表 7.2　不同背景光场景中的光学功率密度（有与没有光学过滤器），**以及标准硅检测器和镜头设置的相应检测器电流**

	$PD_{BL}/(W/m^2)$	PD_{BL}（过滤）$/(W/m^2)$	I_{det}（过滤）/pA
室内	3	0.15	0.8
室外（阴天）	30	1.5	8
室外（晴天）	1006.9	50.3	270

7.10　技术发展最新水平

飞行时间法技术受到学术界和行业研究实验室的广泛研究，因此在该领域涌现了大量的科学著作和许多可用的科研产品。本书我们将阐述现有的像素计数的方案和对未来的展望。

从研究和发展的角度来看，视频图形阵列（VGA）分辨率在今天已不稀奇，一个图像传感器已经能有高达 300k TOF 像素了。如图 7.12 所示，研究报告显示，TOF 传感器分辨率

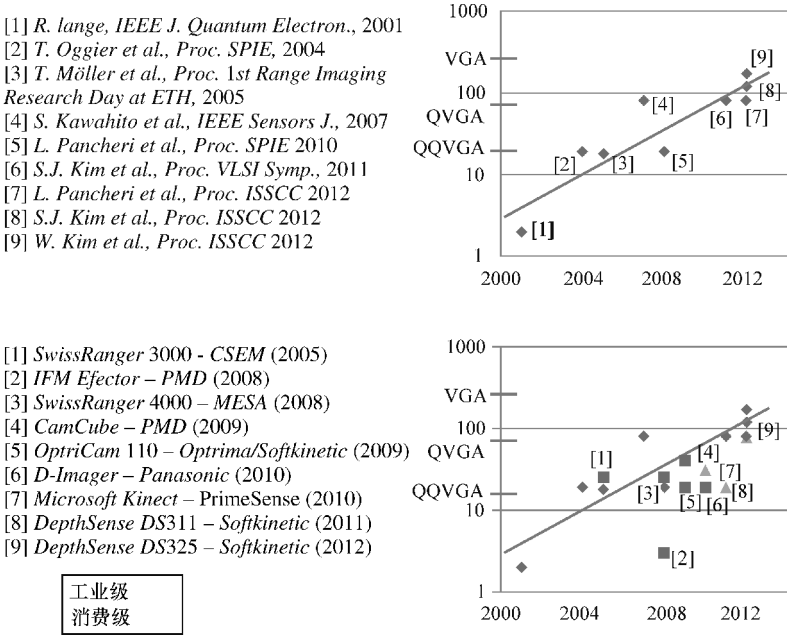


图 7.12　时差测距成像传感器分辨率的综述和趋势（来源：参考文献［11］）。上图显示了科学文献中的结论，而下图显示出可利用的各类产品的分辨率。从测量最小的可检测到的特征尺寸推断出的 Kinect 分辨率，注意到这比特定的视频图形阵列的分辨率还要低得多

一直保持着每 4 年翻 4 倍的增长速度。顺应这一趋势，2016 年之前 720p 的分辨率将在科研人员的努力下实现。与该发展势头相似，新技术在市场中的应用将在研究结果更新后的 4 年内涌现。现有的产品已经在提供 QVGA 分辨率，而未来将会被改进得更加完美。

7.11　结语

本章对飞行时间法 3D 成像技术的基本原理进行了综述。对于持续飞行时间法的设备，我们已经得出了很多重要的公式和标准的系统参数。我们证明了飞行时间法 3D 成像系统中最重要的参数是能够使系统节省功率并达到高可重复性和高精度的调制效率。

随后，我们综述了这一技术所需要面临的挑战和问题，并且讨论了它的解决办法。最终，我们展现了最新发展成果及相关产品的科学界和产业界的发展趋势。

总的来说，时差测距 3D 成像技术在实时操作和分辨率方面有许多的优势。我们相信，时差测距系统将会渗透需要 3D 成像性能的许多部分，并且在广泛的应用范围中促进与用户的实时交互。

参 考 文 献

1. Gulden, P., Vossiek, M., Heide, P., Schwarte, R. (2002). Novel opportunities for optical level gauging, 3-D-imaging with the photoelectronic mixing device. *IEEE Trans Instrum Meas* **51**, 679–684.
2. Viarani, L., Stoppa, D., Gonzo, L., Gottardi, M., Simoni, A. (2004). A CMOS smart pixel for active 3-D vision applications. *IEEE Sensors J* **4**, 145–152.
3. Van der Tempel, W., Van Nieuwenhove, D., Grootjans, R., Kuijk, M. (2007). Lock-in pixel using a currentassisted photonic demodulator implemented in 0.6m standard CMOS. *Japanese Journal of Applied Physics* **46**, 2377–2380.
4. De Nisi, F., Stoppa, D., Scandiuzzo, M., Gonzo, L., Pancheri, L., Betta, G. (2005). Design of electro-optical demodulating pixel in CMOS technology. *Proc. IEEE International Symposium on Circuits, Systems* **1**, 572–575.
5. Oggier, T., Kaufmann, R., Lehmann, M., Buttgen, B., Neukom, S., Richter, M., Schweizer, M., Metzler, P., Lustenberger, F., Blanc, N. (2005). Novel pixel architecture with inherent background suppression for 3D time-of-flight imaging. *Proc of SPIE Electronic Imaging* 1–8.
6. Lange, R., Seitz, P. (2001). Solid-state time-of-flight range camera. *IEEE J Quantum Electron* **37**(3), 390–397.
7. Oggier, T., Lehmann, M., Kaufmann, R., Schweizer, M., Richter, M., Metzler, P., Lang, G., Lustenberger, F., Blanc, N. (2004). An all-solidstate optical range camera for 3D real-time imaging with subcentimeter depth resolution (SwissRanger). *SPIE Optical Design, Engineering* **5249**, 534–545.
8. Hosticka, B., Seitz, P., Simoni, A. (2006). Optical Time-of-Flight Sensors for Solid-State 3D-Vision. *Encyclopedia of Sensors* **7**, 259–289.
9. Lange, R. (2000). *3D Time-of-Flight Distance Measurement with Custom Solid-State Image Sensors in CMOS/CCD-Technology*. PhD thesis, University of Siegen; September 2000.
10. Newport Corporation (2008). *Introduction to Solar Radiation*. http://www.newport.com/Introduction-to-Solar-Radiation/411919/1033/catalog.aspx#.
11. Stoppa, D., Pancheri, L., Perenzoni, M. (2012). *Sensors Architectures for 3D Time-of-Flight Imaging*. 6th Annual Global Conference on Image Sensors.

第8章

凝视跟踪

Heiko Drewes
德国慕尼黑路德维希 – 马克西米利安大学 LFE 媒体科学

8.1 引言和研究动机

我们与计算机设备的交互稳步增加。因此，在 HCI（人机交互）领域的研究一直在寻求更有效、更直观、更简易的交互方法。“更有效”意味着我们可以尽可能快速地进行交互。我们也尽量避免电脑教程讲座或操作培训。这意味着我们喜欢更直观的界面。最后，我们不喜欢做体力或脑力费力的事情，而是希望与设备的交互轻松容易。

传统的交互设备，如鼠标和键盘，如果使用过多可能会导致身体损伤，如腕管综合征。键盘和鼠标在移动的环境中使用也不太实际。因此，我们总是寻找它们的替代品，甚至比它们更好的东西。使用我们的凝视来完成与计算机交互似乎是一个很有前途的想法。视线移动是快速的，我们可以很轻松直观地移动，因此凝视跟踪满足上述所有标准。

此外，如果大批量生产，眼动仪的成本很低。一个小型的眼动仪由一个摄像头、一个 LED、一个处理器和软件组成。在智能设备，如智能手机、平板电脑、笔记本电脑，甚至一些新的电视机上，所有这些组件都已经存在，但即使单独生产，这些组件的成本也比制造光电鼠标的成本要低。先进的眼动仪为了视图更立体而使用两个摄像头，有时配有多个 LED。尽管如此，其成本依旧，尤其大批量生产的话是可以负担的。

认为凝视跟踪技术将成为未来交互科技的更进一步的原因是，人与人之间的交互中，凝视是很重要的。

与动物眼睛相比（见图 8.1），人类眼睛的眼白非常明显。动物的眼睛，特别是与人类眼睛运作方式相似的哺乳动物的眼睛里，却并未见到眼白，由于这个原因，确定动物凝视的方向比人类要难。认识到凝视的方向在我们物种进化过程中起了什么作用还不确定。可以确定的是我们用眼睛来交流。通常情况下，我们用凝视来定位人物或者物体。如果有人问别人：“我可以拿这个吗?”别人能看到这个人在看什么，因此知道“这个”指的是什么。如

果我们想要人机交互达到近似人与人之间的交互，计算机需要具备凝视的自觉性。

图 8.1　黑猩猩的眼睛和人类的眼睛

眼球凝视交互带来了进一步的优势。它没有物理接触，因此是一个非常卫生的交互方式。没有碰触，设备也不需要清洗。眼动仪上没有活动件，这意味着它不需要维护。如果组装上一个变焦镜头，那么眼动仪的工作距离就会比我们的手臂要长，可以作为遥控器使用。

此外，眼动仪能让我们的电脑交互更安全，因为它们明确要求我们集中注意力。需要视线接触才能拨出的移动电话不会因为装在口袋里意外按键就拨出电话。最后，眼动仪可以通过检测我们的活动，有潜力让我们的交互更方便。例如，当我们阅读时，系统可以将非紧急通知推迟，例如软件更新的通知。

本章将继续介绍人类眼睛的基本知识和概述凝视跟踪技术。下一节将阐述凝视交互遇到的反对和障碍，其次是对在过去 30 年凝视跟踪技术研究的一个简短的总结。接下来的三节介绍研究凝视交互的三种方式。

第一种也是最明显的方式是关注眼睛的指向。这类似于鼠标的指向，但精度不高。眼睛指向部分包括鼠标和眼睛指向的比较，并将讨论手和眼睛的协作。

第二种是使用凝视姿势。凝视姿势不是很直观，但姿势属于标准的交互方式库。除了姿势识别和姿势字母表外，这一节也讲解凝视姿势和自然眼球运动的区别。

第三种是把眼睛的凝视作为情景信息。在这里，眼睛的动作不会触发有目的的指令，但是系统会观察和分析眼睛动作，从而以聪明的方法帮助和支持用户。这一节大致讲解活动识别，特别是阅读和注意力检测。本章最后展望凝视交互技术进一步发展的前景。

8.2　眼睛

在医学、生物学、神经科学和心理学等领域，关于眼睛的知识无穷无尽。这里展现的眼睛的知识是经过简化，仅阐述服务理解本章所需的必要的事实。

从技术的角度来看，眼睛可以看作是一对同步运动的动态稳定相机。每个眼球有三对拮抗肌（见图 8.2），它们可以对头部三个自由度起到补充水平向、垂直向和围绕视线旋转方向的作用。

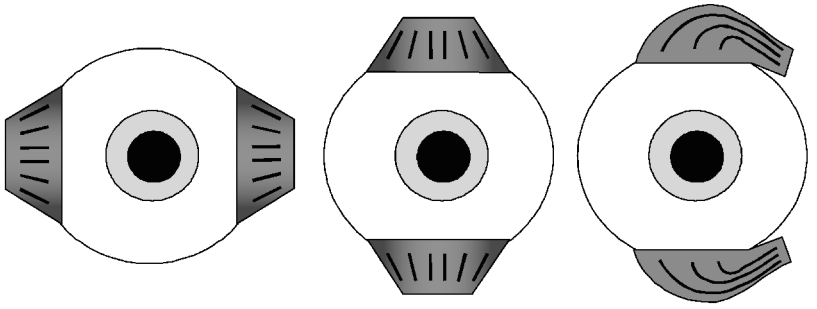


图 8.2　三对拮抗肌可以补偿头部的所有动作

图 8.3 是眼睛的简化示意图。眼睛和相机类似，虹膜就像是光圈，视网膜像是光敏面，晶状体就像是镜头。与相机相比，眼睛通过改变晶状体的形状来聚焦，而不是改变它的位置。相机和眼睛的光敏面有很大的不同。相机的光敏面是二维的且具有均匀分布的光接收器，通常接收红色、绿色和蓝色的光。眼睛的光敏面是圆形的，光接收器不均匀分布。除了

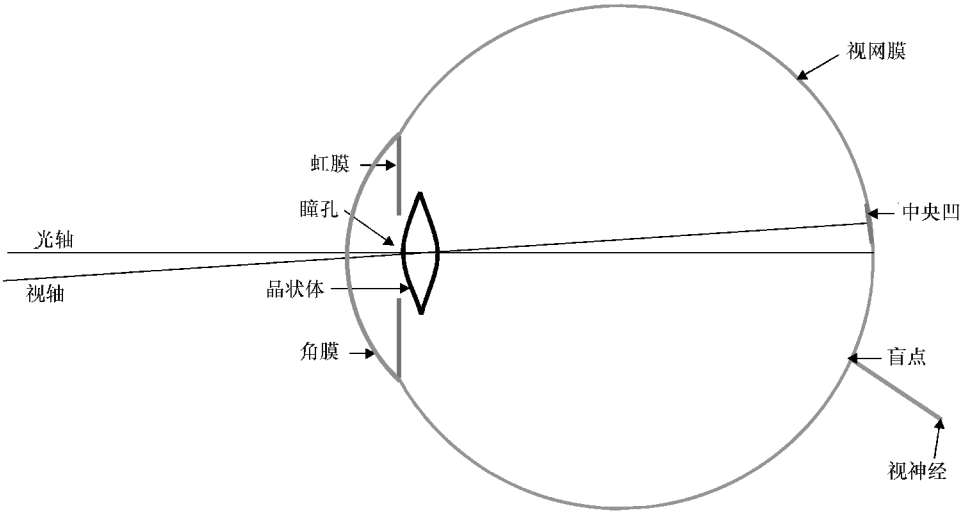


图 8.3　眼球示意图

接收这三种光的接收器（视锥细胞），眼睛还具有另一种接收器（视杆细胞）不对光色区分，但光敏度更高。视杆细胞赋予我们夜视能力。视网膜上的视锥细胞密度低，除了在与瞳孔相对的小点上密度高，我们称这个小点为中央凹。因此，我们只能在狭窄的 1° ~2°的范围内看得清楚。这个距离相对于手臂长度的距离而言是非常小的。我们所感觉的高清晰度画面是大脑产生的幻觉。

在小范围内我们能看见高清晰度的画面，但也有所代价，那就是我们要移动眼球。我们总是把眼睛转到能直接用中央凹看到物体的位置。有两种类型的眼球运动来实现这点：

- 一种眼球运动是补偿运动。当我们视线固定于某物，而转动头的方向时，这种运动就会发生。图像传输的稳定性是必要的，因为我们需要一个稳定的图像投影到中央凹。我们观看一个移动的物体时，也会有这样的动态平衡。保持图像稳定的运动是平滑的。
- 另一种类型的眼球运动是突然快速移动，这被称为扫视。通常情况下，眼睛极快地运动，视线落到感兴趣的点并停留一段时间，这段停留被称为定睛。在此之后，眼睛做另一个扫视运动，依此类推。大多数时候，我们的眼睛做扫视运动。

当眼睛运动的时候，其位置不会变，而是围绕其中心旋转。因此一个扫视的长度由扫视开始和结束时瞳孔正常的角度定义。图 8.4 所示为扫视的时间与视角的关系。很清楚地看见，扫视存在最小时间，由视角决定。然而，对于大视角，时间增加幅度很小。视角大于 5°的扫视持续时间约为 100 ~150ms。

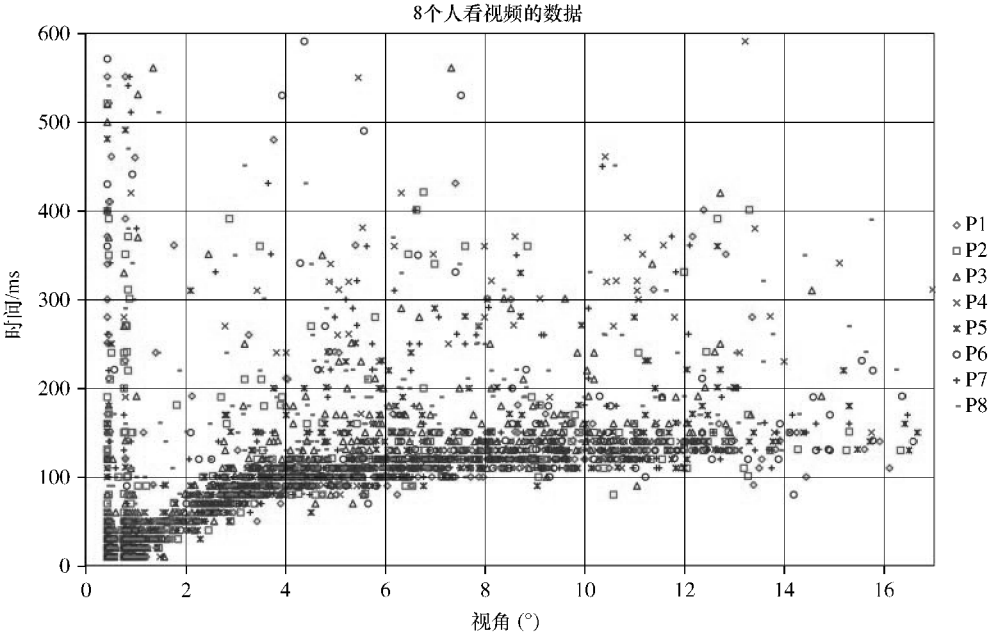


图 8.4　扫视时间和视角的关系

扫视运动速度可高达每秒 700°，以至于视网膜上的光接收器没有足够的时间来检测图像，因此在扫视期间我们处于失明的状态。因此，没有控制反馈回路来引导眼睛看向目标。

心理学把眼睛的扫视运动叫作弹道运动。这意味着，扫视运动不服从费茨定律[1]，即使一些在 HCI（人机交互）界的出版物持相反态度[2-5]。与服从费茨定律的运动不同，弹道运动的时间不取决于目标的大小。

Carpenter 在 1977 年[6]测量了扫视的旋转幅度和持续时间。他用线性近似来表达扫视的时间 T 和其幅度 A 的关系：

$$T = 2.2\text{ms}/^\circ \cdot A + 21\text{ms}$$

1989 年，Abrams、Meyer 和 Kornblum[7]提出了一个模型，即肌力随时间增加而持续增强。由于眼睛的质量和形状不会改变，加速度 $a(t)$ 和肌力成正比例，也不断随时间增加而增加：

$$a(t) = k \cdot t, \text{ 其中 } k \text{ 为常数}$$

做两次时间的积分并求解幅度的方程式，结果显示时间和幅度存在一个立方根的关系。

$$T = c \cdot A^{1/3}$$

常数 c 取决于常数 k 和眼睛的惯性矩。

见图 8.4，它显示正如 Carpenter 做的那样，假设数据都在一条直线上面，线性近似在某个范围内是合理的。但是，参考文献［7］中的模型，与实验数据更吻合。

眼动仪测量到的定睛时间范围通常在 0～1000ms。定睛时间一般不会超过 1000ms，而短的定睛也需要仔细分辨。因为眼睛和大脑需要一些时间来进行图像处理，因此定睛应该持续一段时间。很短的定睛是毫无意义的或者是扫视检测算法中的假象。

8.3　眼动仪

眼动追踪这个术语没有精确的定义。在某些情况下，眼动追踪表示追踪眼球的位置，而在某些其他情况下，它表示检测凝视的方向。也有人把眼睛作为一个整体来追踪，包括眉毛，并尝试由此检测情绪状态，例如，参考文献［8，9］。这种眼动追踪是分析面部表情的一部分。在本书中，术语“眼动仪”是对凝视方向的检测，因此有时也被称为“凝视追踪器”或“视线追踪器”。对眼睛位置的追踪是视线追踪系统的一个子任务，它允许头部自由地在显示屏前运动。

8.3.1　眼动仪的种类

有三种不同的方法来追踪眼球的运动。

最直接的方法就是将传感器固定在眼球上。把小杠杆固定在眼球上就属于这一类方法，但是我们并不推荐，因为其造成伤害的风险高。使用隐形眼镜一种是把传感器放入眼睛更安全的方式。隐形眼镜中的集成镜面可以测量反射光[10]。此外，隐形眼镜中的集成线圈能够检测出磁场中线圈的方向[11]。连接线圈与测量设备的细线对实验对象而言很不舒服。使用这种方法很大的一个好处是精准度高，并且即时获得近乎无限的高分辨率。出于这个原因，医学行业和心理研究都使用此方法。

另一种方法是眼电图（EOG），传感器连接到眼睛周围的皮肤测量电场。最初，传感器被认为是测量眼睛的肌肉电位，后来发现眼睛的电场是一个电偶极。该方法对电磁干扰很敏感，但因技术先进成熟，效果不错。同时该方法相关知识资料充足，工业标准齐全[12]。这种方法的最大优点是即便在闭着眼睛睡觉的时候，它都能检测眼球的运动。现代硬件技术允许把传感器集成到眼镜上打造可穿戴的 EOG 眼动仪[13]。

目前所述的两种方法有点突兀并且不适合用于凝视交互。第三种方法，也是对于凝视交互我们所推崇的方法，是基于视频的一种方法。此方法的核心部分是用一个视频摄像机连接到计算机进行实时图像处理。图像处理接收从摄像机传送的图像，并检测瞳孔来计算视线的方向。视频眼动跟踪的方法有一大优势，就是它不突兀。因此，它是构建人机交互视线接口的方法。基于视频的角膜反射方法将在下一节详细描述。

一般有两种类型的基于视频的眼动仪：固定眼动仪和移动眼动仪。

固定眼动仪，如图 8.5 所示，显示凝视方向是相对于用户的空间，通常显示为屏幕坐标。简单的固定眼动仪需要眼睛保持稳定，因此，使用者需要头部固定。除了眼睛其他身体部位都不能移动的残疾人可以使用这样的系统。非残疾人更喜欢能在显示屏前自由移动的系统。这样的系统通常有一对提供立体视图的摄像机，它不仅追踪视线方向，还能追踪头部的位置和方向。固定眼动仪是一个独立的设备，可以追踪落在物体上的视线。然而，因为许多眼动追踪应用程序都是在显示屏前运行，有些眼动仪直接集成到显示器上，甚至有可能不会被用户注意到。

图 8.5　拓比公司（Tobii）的固定眼动仪，一个独立的系统和一个集成到显示器的系统。来源：拓比公司（Tobii）转载许可

移动眼动仪连接到用户的头部。这种类型的眼动追踪器根据头部的朝向来确定视线的方向。通常，移动眼动仪配有一个头戴式摄像头来捕捉用户所看到的画面。从眼动追踪器的数据可以计算出用户正在看的位置，并且能在头戴式摄像头记录的图像上标记出来。随着最近

小型摄像机设计的发展，眼动仪可以集成到眼镜上。图 8.6 展示了一副眼动追踪眼镜。

Vertegaal 等人还介绍了另外一种视频眼动仪，我们把它称作 ECS 视线接触传感器[14]。ECS 视线接触传感器并不为视线方向提供坐标，而仅仅是为视线接触提供信号，在 10m 之内有效。xuuk 公司的 Eyebox2 视线接触传感器如图 8.7 所示。

图 8.6　SMI 公司的眼动追踪眼镜。来源：SMI Eye Tracking Glasses。转载已获 SensoMotoric Instruments 公司的许可

图 8.7　xuuk 公司的视线接触传感器 Eyebox2。来源：xuuk 公司。转载已获许可

8.3.2　角膜反射法

视频眼动仪的一般任务是分析摄像头记录的图像来估计视线的方向。

检测虹膜的一个可行办法是利用眼白和暗虹膜的高对比度。此方法的结果在水平方向上精准，但垂直方向不精准，因为虹膜的上部和下部被眼睑遮盖了。由于这个原因，大多数视频眼动仪取而代之检测瞳孔。摄像机图像中检测瞳孔是图像识别的一种任务，即边缘检测，来估计瞳孔的椭圆轮廓[15]。另一种检测瞳孔的算法是 Starburst 算法，将在参考文献［16］中解释。

有两种方法来检测瞳孔，分别检测暗瞳孔和亮瞳孔的方法。暗瞳法，图像处理时在摄像机拍摄的图像中定位黑色瞳孔的位置。但这种方法不适用于深棕色眼睛，因为棕色虹膜和黑色瞳孔之间的对比度是非常低的。亮瞳法使用附加的照明，使用与摄像机同一方向的红外线的光照射。因此，红外 LED 必须安装在摄像机内部或靠近摄像机的地方，这就对设备有要求。视网膜反射红外光，这使得在摄像机图像中的瞳孔呈白色。用闪光拍摄人脸的时候，这种效果就被称为“红眼”。对于个体之间的红外亮瞳响应的差异性，见参考文献［17］。

大部分眼动仪使用从角膜反射的图像，也叫第一 Purkinje 图像，来估计的视线方向。由

于角膜是一个完美的球形，闪光点停留在同一位置，视线方向就可以计算出来了（见图 8.9）。

眼动仪的图像处理软件检测闪光点的位置和瞳孔的中心。闪光点到瞳孔中心的矢量是计算视线方向的基础（见图 8.8），最后在画面上确定视线的位置。直接计算不仅需要眼动仪的空间几何形状、红外 LED、显示器和眼睛，还需要知道眼球的半径，每个使用眼动仪的用户眼球半径都不同。出于这个原因，校准过程估计闪光点和瞳孔的矢量在屏幕上位置的映射的参数。校准程序要求用户在校准过程中看多个校准点。四个点的校准程序使用靠近显示屏四角的校准点。

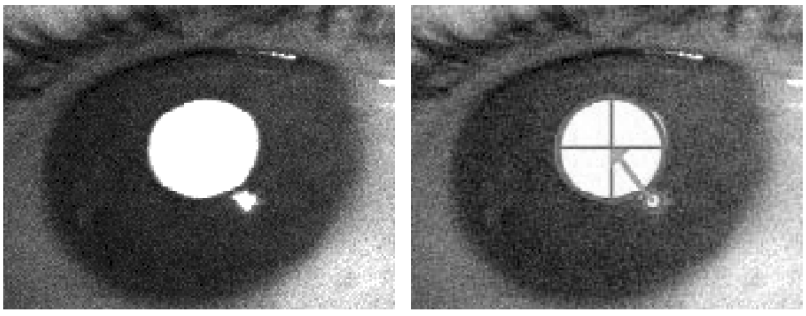


图 8.8　闪光点到瞳孔中心的矢量是计算视线方向的基础

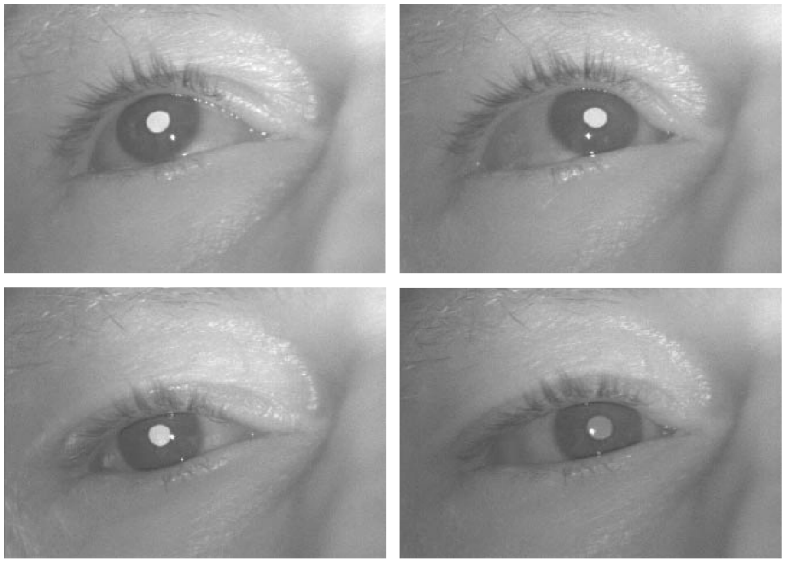


图 8.9　看向显示屏的四个角落——反射都保持在同一位置

角膜反射法不适合眼球变形的人。当人们戴隐形眼镜的时候也会有问题。带镜片眼镜问题就小得多。虽然眼镜可能改变闪光点的位置，但是反射保持在相同的位置。校准能弥补眼镜的光学畸变。

目前为止，我们所阐述的角膜反射法要求眼睛保持在一个稳定的位置，这意味着它要求使用者头部固定。对于人机交互，允许使用者在设备前自由移动更加理想。这种眼动仪使用立体视图，双摄像头，也可以只使用单个摄像头配置多个闪光源。对于这样的系统如何工作，请参见参考文献［18，20］。然而，通常商业眼动仪的运行方式是一个商业秘密。

上面提到的视线接触传感器也使用了角膜反射法。对于照明，一组红外 LED 被安装在红外摄像头的轴周围。当摄像机提供一幅闪光点在瞳孔内的图像时，这意味着测试者是直接看摄像机的。这种方法的一个很大的优点是它不需要校准。

8.4　反对和障碍

在引言中，凝视交互看上去似乎很有前途。然而，并不是所有的承诺都很容易实现，实现凝视交互的过程中有一些障碍。

8.4.1　人为方面

大家普遍关注的是，凝视界面将与眼睛的首要任务视觉相冲突。眼睛可能会有一个输入－输出之间的冲突，视觉和交互需要不同的眼睛运动。Zhai 等人在 1999 年中写道：“第二点，也许也是更重要的一点，我们的眼睛，作为我们的主要感知设备之一，还没有进化成为一个控制器官。有时它的动作是主动控制的，而在其他时候它是由外部刺激所驱动的。”[21]

视野中的变化可能引发眼球运动。如果一个凝视感知界面显示了一个闪烁的物体，眼球则很有可能转向闪烁的物体。如果眼球运动触发了一个新的命令，用户将会无意地调用这个命令。在一般情况下，在凝视界面创建这样的输入输出冲突是可能的。然而，凝视界面的开发人员需要构建这样的冲突。通常情况下，这样的冲突不会发生，也没有科学家报告这种类型冲突的严重问题。

也有反对意见称我们不能控制我们的眼球运动。尽管我们的眼球运动是由视觉任务所驱动的，我们能意识到自己的凝视，并且可以控制它。否则，我们就违反了社交协议。

反对凝视交互的另一个观点称眼睛可能会疲劳，如重复性压力损伤等问题就可能发生。然而，我们的眼睛是不停运动的，即使在我们睡觉的时候也是如此。如果一个人的眼睛一分钟不移动，我们就开始担心他是否失去了知觉。因此，疲劳对眼睛而言似乎并不是一个问题。

最后，另一个反对眼动追踪的观点是接受度问题。摄像头和所有电子设备的网络连接将为 Orwell“老大哥正在看着你呢”的设想提供基础设施。眼动仪带有摄像头我们可能会习惯，因为我们的周边早已经充满了其他的摄像头。然而，眼动仪的用户可能感觉到被监视了，并且他们也许不会接受在私人空间，如浴室中使用眼动仪。对眼睛运动的分析似乎可以判断我们的阅读能力，因此透露我们的智商或者至少智商的形成。通过凝视数据得出的结论可以吓唬人，尤其是受雇者。

8.4.2　室外应用

固定眼动仪通常运行良好且可靠，因为它们通常位于室内，光照条件相对稳定。在室外环境中，光的变化范围很广，这意味着可能有极端的光线反差。此外，光线还可能快速改变，例如移动的阴影，坐在汽车内尤其如此。这样的情况对摄像头而言仍然是一个挑战。大部分商用系统都是以红外光为基础，因为太阳是明亮的红外辐射源，因此可能会出问题。这使得它很难可靠地在摄像机图像中检测出瞳孔和闪光点。具有与摄像机的帧速率或者偏振光的使用同步的差分图像和红外照明的方法是很有前途的[22]。因此，对于大多数光条件的眼动追踪，除了极端情况，应该都是可以实现的。

8.4.3　校准

眼动仪需要进行校准以达到良好的精准度。使用红外 LED 闪烁的方法取决于眼球的半径，因此，需按用户校准。尽管眼动仪能够在无闪光点的时候确定瞳孔在空间的方向，不依赖眼球的半径，但是仍然需要校准。其原因是，每个人中央凹的位置不同。光轴（位于瞳孔中心的法线）并不完全是视轴（从中心凹到瞳孔中心的一条线）。

校准过程的优点是它仅需要做一次。配备了眼动仪的个人系统仅需要校准一次。而对于精准度要求较高的公共系统的凝视检测，例如自动柜员机，校准程序是一个真正的难题。

避免校准问题的一种方法是不使用绝对凝视位置，而仅仅采用相对凝视运动。只检测相对运动意味着使用手势。这是一种选择，但它也是一个严重的限制。

8.4.4　精度

精度包括两个方面：一是眼动仪的精度；一是眼球运动的精度。

可用的眼动仪的精度还远远没达到物理极限。眼动仪具有空间和时间分辨率。时间分辨率取决于摄像头和处理器的速度和算法。对于固定眼动仪，空间分辨率主要是摄像头分辨率的问题。由于摄像头的分辨率和处理器速度会不断增加，我们可以预期，在不久的将来我们的眼动仪会有更高的精度。对于移动眼动仪系统，空间精度还依赖于头戴式系统的机械稳定性。

时下的眼动仪一般声称为 ±0.5°的精度。臂长的距离是从显示屏到眼睛的典型距离，该距离下眼动仪的精度约为指甲大小。这样的精度是不够用眼睛凝视来代替鼠标的。典型的图形用户界面使用的交互元素是比指甲小的。

我们眼睛的精度是一个微妙的问题。问题不仅是我们如何准确地定位我们的视线，而且我们应该如何准确地做。即使我们的视线能相当准确地停留，问题是我们需要集中多少注意力。识别物体的时候，物体在中心凹上投影就足够了，因此，我们发现在这个精度范围内眼睛能够定位自己的位置。这种情况似乎与在晚上用火把发现一只昆虫是相当的，如果昆虫是在光圈内，那么这个精度就够了。但是不值得这么努力去把它作为中心。

Ware 和 Mikaelian 是这么说的：

“关于眼球正常运动的研究文献告诉我们，准确到10′角可视角度（相当于0.16°）的定睛是可以达成的，但不受控制的自发运动会造成眼睛视线间歇性离开目标。但是，当观察者连续地定睛多个目标，那么眼睛的精度会大大减少，大幅度的错位也可能变得普遍。”[2]

8.4.5　点石成金（Midas Touch）问题

虽然凝视指向似乎和手指指向非常相似，但是它们有一个重要的区别：我们不能像举起手指一样举起我们的目光。在触摸屏上，我们可以指向一个交互元素，并通过触摸表面触发该指令。对于凝视，我们也可以指向一个交互元素，但我们并不能触摸屏幕。如果我们仅仅是看着就能触发指令，那么我们就会遇到一个大问题，即就算我们仅仅想看看屏幕上有什么，我们都将触发指令。Jacob称这为点石成金问题，他解释道：

“起初，它非常简单，看你想要什么，它就会发生。但是不久之后，它就变得像点石成金那样。你看的每一个地方，都有一个指令被激活；你看的所有地方都会触发指令。”[23]

当Jacob发现点石成金的问题时，凝视指向就出现在他的脑海里。站在更广泛的层面上来讲，点石成金的问题是决定眼睛的活动是为了发出一个指令，还是只是视觉任务中的一部分的问题。即使改变了交互方法，问题依然存在。凝视手势也存在这样的危险，即它们在自然眼睛运动中也可能发生，并触发意外的指令。手势并不意味着触摸，这里的术语点石成金可能会产生误导作用。因此，最好是把手势从自然动作分离出来阐述。

8.5　凝视交互研究

使用凝视来交互的想法已经有30年了，自那时起就已经有了大量的研究。因此，下面的概述是这一领域发展研究历史中的一小段。

凝视交互之所以成为可能，是因为有电子视频摄像头和足够强大能够进行实时图像处理的计算机。第一批系统是在20世纪70年代建立用来帮助残疾人的。给残疾人用的典型的应用程序就是眼睛打字[24]。眼睛打字时，显示屏上显示标准键盘。如果凝视在这个虚拟键盘上指向了一个键，则该键被高亮显示。如果凝视停留在这个键的时间比一个预定的停留时间更长，通常在500ms左右，这意味着该键被按下了。

1981年Bolt[25]对健全的人使用的多模态凝视交互进行了预言。他描述了配有巨大显示器的媒体室，其中有15~50个窗口同时显示动态的内容，他把它命名为“视窗世界”。他的想法是将一些用户正在观看的窗口进行放大。基于停留时间，他描述了一种界面方法，还讨论了多模态界面技术。一年后，Bolt发表了论文《Eyes at the Interface》[26]，总结凝视在沟通时的重要性，并得出结论：界面技术需要的凝视意识。他的构思目前尚未完全实现。

1987，Ware和Mikaelian对凝视指向[2]进行了一系统研究。在他们的论文《An evaluation of an eye tracker as a device for computer input》中，他们介绍了三种不同的选择方法，他们称之为“停留时间按钮”“屏幕按钮”“硬件按钮”，并测量了眼睛做选择时需要的时间。对于停留时间按钮，视线要在按钮上停留一定的时间（停留时间）来触发与按钮相关的指

令。屏幕按钮是一个双目标的任务。视线移动到所选的按钮，然后，移动到屏幕键来触发指令。硬件按钮的方法是当视线在选定的按钮时，使用手指来按一个键。前两种都仅仅是凝视，而硬件按钮则需要使用一个额外的操作模式。Ware 和 Mikaelian 将他们的数据代入“改进后的”费茨定律，然后把得到的结果与 Card 等人[27]的鼠标装置的实验结果做对比。然而，费茨定律并不适用于眼睛。

1990 年，Jacob[23]系统地研究了用眼睛操作 GUI（图形用户界面）所需要的交互——选择对象、移动对象、滚动文本、触发菜单命令和设置键盘焦点落到同一个窗口。这篇论文的一大贡献是（8.4 节有阐述）发现了点石成金问题。Jacob 的论文的普遍性导致了所有进一步的研究将重点放在眼睛凝视交互的单个或更加专业的问题。

1999 年，Zhai 等人提出了建议来处理凝视指向固有的低精度的问题（8.4 节有阐述），并把它命名为 MAGIC（鼠标和凝视输入级联）指向[21]。MAGIC 指向使用凝视进行粗糙定位和用传统鼠标进精准定位。

2005 年，Vertegaal 等人引入了 EyePliances 媒体，其中远程遥控可以与多种媒体设备交互[14]。只需要用眼睛看着就能选择想要远程控制的设备。为此，他们给设备增加了一个简单的眼动仪，称为视线接触传感器 ECS。在同年的另一篇论文中[28]，Vertegaal 等人在与移动设备组合的过程中使用了 ECS。移动设备一直不受充分关注，因为使用移动设备的用户必须注意她或他的周围环境。他们用两个应用程序 seeTXT 和 seeTV 展示了如何使用眼睛凝视来检测注意力和如何使用这方面的文本信息来控制设备。seeTV 是一个视频播放器，当用户不看它的时候，它会自动暂停。seeTXT 是一个阅读应用，只有当用户看着屏幕时，它才会翻页。

近年来，眼动交互的研究已成为流行，出版物数量也增加了不少。自 2000 年以来，就形成了一个围绕该主题的专题会议，被称为 ETRA（眼动跟踪研究和应用）。自 2005 年以来，COGAIN（视线交互通信）倡议，在欧盟的支持下还组织了会议并在互联网上提供了研究论文的目录。

8.6 凝视指向

把凝视作为电脑输入最显而易见的方法是凝视指向。看东西是我们的直觉，眼睛能够快速且轻松地执行此任务。指向也是与图形用户界面交互的基本操作，用眼睛来完成指向将加速我们的交互。因此，大多数视线交互的研究都涉及凝视指向。然而，正如在反对和障碍的那节（8.4 节）中已经提到的，凝视指向存在一些固有的问题，如点石成金的问题和低精度问题。

8.6.1 解决点石成金问题

视线感知界面中有这样的问题，当视线凝视交互对象时就会触发相应指令，即使我们只想看看有什么。这个问题被称为点石成金问题。有几种方法可以解决它。凝视系统，通常用于残疾人，引进了停留时间的概念。这意味着用户要想在交互中触发一个指令，视线就必须

在屏幕上停留一个特定的时间，即停留时间。停留时间通常在 500 ~ 1000ms 的范围内，并且会耗掉快速眼动所节省的时间。

解决点石成金问题的另一种方法是使用另一种形式，如一个凝视键。按凝视键激活眼睛在看的命令。使用凝视键允许快速交互，但也消除了凝视界面的部分好处。附加一个按键也意味着界面不再卫生，因为有了需要接触的东西。此外，因为够得着的距离内要有一个按键，所以它不适用于较长的距离。最后，残疾人使用凝视界面时不能使用这个按键，因为他们无法按键。更深入地思考凝视键这个问题就会发现只有目的是输入二维坐标才是有意义的。通过看“保存交互对象”输入一个命令（例如，一个保存操作），然后按下凝视键，这会引出问题：为什么不干脆按 CTRL – S 键，而要看特殊的交互对象呢？

我们经常提到的眨眼睛以触发一个命令的建议似乎并不是一个选择。眼睛一眨一眨来保持眼睛的湿润，因此，一个眨眼的指令要比自然地眨眼用的时间要长，而且任何速度上的利益都会受损。在凝视位置所要触发的命令正好是眼睛闭合的时候，这是矛盾的。然而，不使用眨眼的主要的原因是，那会让人感到不舒适。眨眼有可能会替代鼠标点击。当我们操作图形用户界面时执行鼠标点击的次数很多，一般都会超过每小时点击 1000 次。眨眼 1000 次会使眼睛神经紧张。

8.6.2　精度问题的对策

当前眼动仪的精度，以及我们的眼球运动的精度，都不允许我们处理微小的对象，甚至是一个像素。在解决精度问题方面做了大量的研究。最简单的解决方案是扩大交互对象。假设在 72dpi 显示屏上显示 0.5in 的精度意味着交互对象不应小于 36 × 36 像素。现有的图形用户界面使用按钮的大小为 16 × 16 像素，菜单项或文本线的高度为 8 ~ 12 像素。这意味着图形用户界面必须在每一个维度要增大 3 倍，或者我们的显示器需要增大大约 10 倍。对于大多数情况下，这样浪费显示区域是不能被接受的。

研究提出了几个如何解决的精度问题的建议，即原始和精细定位，增加智能，扩大目标，以及使用进一步的输入方式。以下会对这些进行讨论。

Zhai 等人提出了一个处理低精度问题的建议，被称为 MAGIC（鼠标和凝视输入级联）指向[21]。MAGIC 指向在原始定位中运用凝视，同时在精细定位中使用传统的鼠标设备。MAGIC 指向的基本思想是在没有活动进行的一段时间以后将鼠标光标定位在第一个鼠标移动时的凝视位置上。运用 MAGIC 指向时，凝视将鼠标指针放置在靠近目标的位置，并将鼠标用于精细定位。在 Zhai 等人的研究中他们发现有超出目标的问题，因为在定位时手和鼠标已经在移动中。他们建议了一种从距离和初始运动矢量计算得出的补偿方法。

Drewes 等人提出了该原则的改善方法，并称之为 MAGIC 触摸[29]。他们制作了一个有触摸感测器的鼠标，在触摸鼠标时将鼠标指针放置在凝视位置。改进之处在于补偿方法的缺失，因为将手指放在鼠标键的时候，鼠标没有移动。另一个优点是用户可以选择凝视定位的时间，并且不需要一段鼠标闲置的时间。

当看到人们在大屏幕或双显示器设备前工作时，会很容易发现他们有时候很难找到鼠标

指针。此外，很多时候鼠标指针离我们想要指向的目标很远。因此，很有必要将鼠标从屏幕很远的地方拖过来。用凝视进行原始定位和用鼠标进行精细定位的原则不仅可以解决精确问题，还能够避免找鼠标指针的麻烦。此外，它还节省了鼠标覆盖的长距离，因此有助于防止重复性压力损伤。

2000 年 Salvucci 和 Anderson 提出了他们的智能凝视界面[30]。启动这个界面的系统，凝视追踪器向标准的 GUI 提供 X – Y 位置，用户所观看的交互对象就会变亮。凝视键，类似于鼠标键，为用户提供触发动作的可能性。为解决精度问题，系统智能地理解凝视输入：它将凝视点映射到用户可能留意的条目。为了找出这些条目，该系统使用概率算法，通过凝视位置（即接近所报告的凝视点的条目）和任务的上下文（例如，一个命令后会有另一个命令的可能性）来确定。

解决精度问题的另一种方法是使用目标扩大。Balakrishnan [31] 和 Zhai[32] 研究了手动指向中扩大目标的运用，并指出该技术对指向性任务有所帮助。Miniotas、Špakov 和 MacKenzie 把这项技术应用于眼凝视指向中[33]。在他们的实验中，扩大的目标并没有视觉地呈现给用户，但界面响应一个扩展的目标区域。他们称这种技术为静态扩展。在第 2 篇文章中，Miniotas 和 Špakov 对扩大目标进行了动态研究 [34]，即目标的扩大对用户可见。这项研究是针对菜单目标的，结果表明，增加选择的时间会显著降低选择菜单项的错误率。

Ashmore 和 Duchowski 在同一年发表了利用鱼眼晶状体来支持眼睛指向的观点[35]。

2007 年 Kumar 等人提出了一种眼凝视界面，并称之为视点 [36]。此界面使用了交互目标的扩大技术，并且也使用了一个键作为额外所需的输入方式。当按下这个键时，所凝视的屏幕区域就会扩大。在这个放大的屏幕区域内，用户用凝视选择目标，用户一松开键就会触发动作。

凝视指向的不准确性意味着如果多个目标离得很近，指向性行为对于目标就会有模糊性。Minotas 等人从中得到了启发，运用一个额外的语音指令来确定目标[37]。他们用不同颜色的目标，并要求用户大声说出目标的颜色。他们发现这种方法可以处理大小在 0.85°内，相互距离在 0.3°的目标。

在速度方面该方法并没有带来好处。操作标准的图形用户界面的情况下，这个方法带来的更精确的指向性是否值得额外加语音的麻烦还不清楚。然而，这个概念还是很有趣，因为它与人和人交互十分相似。通常情况下，我们都能知道其他人朝什么方向看，但精度远低于眼动仪。当我们说，“请给我绿色的书”，并朝桌子看，别人会给我们从桌子上拿绿色的书，而不是从书架上拿。我们假设其他人知道我们在看哪里，只需要在那个范围内指定对象。

8.6.3　鼠标指向和凝视指向对比

更深入地了解凝视指向的一个好办法就是将之与其他指向方法相比较。除了精度和速度，指向设备还在以下方面存在不同：空间要求、反馈提供、是否支持多指针、指向模式。表 8.1 给出了这些属性的概述。

费茨定律给出指向设备的速度和精度之间的关系。指向操作要达到更高的精度要求更多

的时间。在触摸屏幕或凝视指向的情况下，人体组织尺寸限制了精度。指向目标的大小是由手指或中央凹的大小来决定的，并不是由速度－精度来决定。

鼠标在桌子上运动需要一定的空间。当你坐火车或飞机的时候，空间往往是不足够的，因此移动设备通常使用轨迹球、轨迹杆，或触摸板。触摸屏不需要额外的空间，但手指的精度低，因为指尖会遮挡住一些视觉信息。为了实现在触摸屏上的高精度，人们使用尖细的触控笔。凝视指向所需要的空间和精度类似于手指点击触摸屏。但是凝视不会遮挡视觉信息，但使用触控笔来增加精度不是不可取的。

对于间接工作的指向装备，鼠标指针的反馈是必要的。对于直接在触摸板上指向，反馈是没有必要的。凝视指向也是一个直接工作的方法，不需要反馈。之所以凝视指向也可以要反馈，是来确保眼动仪所报告的坐标就是凝视的位置。8.6.5 节将讨论凝视指向的反馈和引入凝视指针却适得其反的原因。

多指针的使用是当前研究的一个主题。有很多关于双手的交互和使用所有手指指向的讨论。很明显对于眼睛，双眼是同步移动的，我们不能独立地使用两只眼睛。多凝视指针只对多个人有意义。

许多图形用户界面操作使用指针配合在鼠标键上的点击，这意味着需要一个额外的模式。触摸屏不是这种情况，触摸能提供指向以及点击。触摸板不可能做到与触摸屏相同，因为间接方法不允许直接触摸目标。触摸发生在反馈指针指向目标之前，因此，触摸对触发目标相应的指令是没用的。触摸板可以增加压力代替点击，但安装了触摸板的商业设备通常会提供额外的鼠标键。与传统的指向装置相比，触摸屏是与凝视指向最相似的。最大的不同在于，手指可以被提起来移动到另一个位置，而凝视并不能。因此，凝视不能像手指一样进行点击。

表 8.1　指向设备的属性

	鼠标	轨迹球	轨迹杆	触摸板	触摸屏	凝视
速度	快	快	一般	快	快	非常快
精度	时间	时间	时间	时间	手指的大小	中央凹的大小
空间要求	多	少	少	少	无	无
反馈	是	是	是	是	否	否
方法	间接	间接	间接	间接	直接	直接
多指针	双手	双手	双手	10 只手指	10 只手指	一双眼睛
内在点击	否	否	否	是（否）	是	否

8.6.4　鼠标和凝视协调

图 8.10 和图 8.11 展示了典型的鼠标和凝视指向目标的移动。有趣的是，视线直接移动到目标，不看鼠标指针的位置。在周边视野区域，运动检测效果良好，而且通过凝视指向无需点击鼠标指针。

凝视指向不会给眼部肌肉造成额外的负担，不会比使用传统鼠标给眼睛造成更大的压力。原因很简单，因为我们不看目标就无法选中它。在特殊条件下，如我们能用周边视觉看到大的目标的条件下，那么跟随鼠标光标的运动并将其引导到仅有运动图像的目标，但凝视

不会点击目标，这种情况将成为可能。

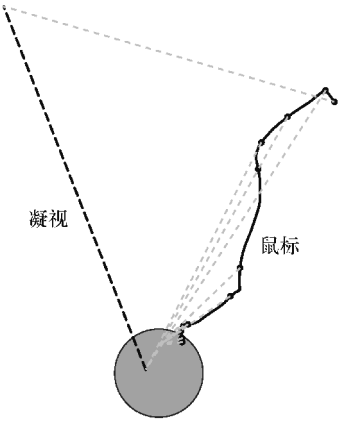


图 8.10　没有背景的典型鼠标任务中的视线（虚线）和鼠标移动（实线）。点状的灰色线连接同一时间的点

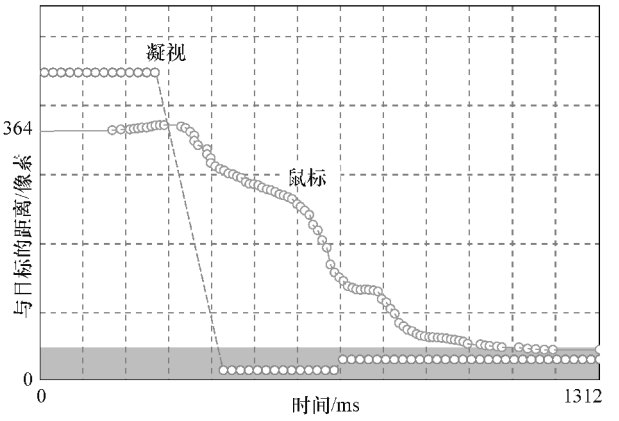


图 8.11　没有背景的典型鼠标任务中的视线（虚线）和鼠标移动（实线），随时间到目标的距离绘图

鼠标指针直接向目标移动意味着用户在开始动作之前就已经知道了鼠标指针的位置。在复杂背景下指向目标的任务将破坏用户提前感知的可能性，而且用户不知道鼠标指针的位置。通常情况下，人们最开始移动鼠标，让它运动来检测鼠标指针。图 8.12 和图 8.13 展示了此情况。

如图 8.11 和图 8.13 所示，眼睛和手具有大约相同的反应时间，但是视线到达目标的时间要早得多。因此，凝视指向绝对比鼠标指向更快。鼠标指向可能发生的情况是，我们不知道鼠标指针在哪里，必须先找到它，而这种情况凝视指向永远不会发生。凝视指向并按下凝视键是目前已知的最快的指向方式，通常需要大约 600ms，即 300ms 反应时间、100ms 视线移动到目标和 200ms 按下键。

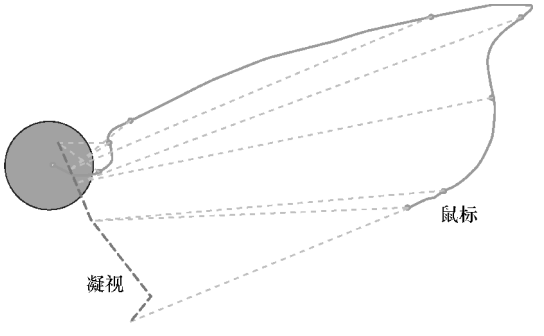


图 8.12　复杂背景下的典型鼠标任务中的凝视（虚线）和鼠标轨迹（实线）。一开始，使用者移动鼠标来检测鼠标的位置

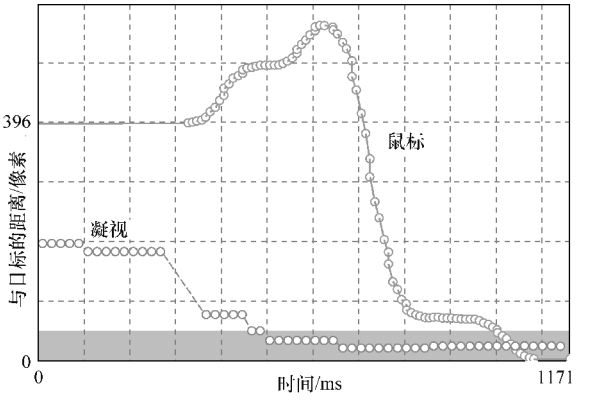


图 8.13　复杂背景下的典型鼠标任务中的凝视（虚线）和鼠标轨迹（实线），随时间到目标的距离绘图

8.6.5　凝视指向反馈

与凝视指向相关的一个有趣的问题就是反馈的提供。当然，用户知道她或他正在看的方向，但是无法达到单像素的精度。此外，可能存在的校准误差会造成的凝视位置和眼动仪所报告的位置不同。若提供一个凝视反馈的光标，则会导致凝视满屏追逐视线光标，或者就像 Jacob 所说的那样：

“如果有任何系统的校准误差，该光标就会从用户实际上看的地方稍微偏移，导致用户的视线被吸引到该光标处，这将进一步改变光标位置，创造一个正反馈回路。”[23]

然而，这样追逐视线反馈光标的情况通常不会发生。眼睛似乎并不关心光标是否正处于清晰视觉区域的正中心，因此，视线并没有被吸引到光标处。进一步解释这种现象不会发生的原因是眼动仪的过滤算法。原始的凝视数据通常是非常混乱的，因此，传输给应用程序的凝视数据通常会进行平滑处理。在许多情况下，会对原始数据进行一个扫视（和凝视）检测，凝视感知应用程序只会得到扫视的通知。在这种情况下，凝视方向的微小变化不改变所报告的坐标，反馈光标不移动；如果凝视位置变化超过一个阈值，那么反馈光标才会移动。

在原始数据的基础上提供一个反馈光标，这个光标会变成变形的光标。因为原始数据通常包含由凝视检测产生的噪声。数据平滑的反馈光标仍然是变形的而且显示延迟。引入阈值来表示凝视位置的变化，这是扫视检测的一个简单形式。它能产生一个稳定但跳动的反馈光标，因为光标移动至少是所述的阈值距离。所有例子与其说有用倒不如说更加令人不安，因此，不应该有任何的凝视光标。这并不意味着反馈是不必要的。通常情况下，系统使用凝视指向突出目光所聚焦的对象。如果系统使用停留时间的方法，提供对所用时间的反馈是一个好主意。至于目光感知的应用程序应不应该提供或者提供什么样的反馈，取决于应用程序，并且没有统一的答案。

8.7　凝视姿势

8.7.1　凝视姿势的概念

姿势是计算机交互的一种可行方式。智能手机是由手指在触摸感知的显示屏上触摸而运行的，随着智能手机的推行，这种交互的概念突然变得非常流行。3D 扫描仪可以检测手或身体的姿势，并提供另一种形式的姿势交互，这种交互主要用于游戏领域。

当然，姿势与眼神共同执行的想法更容易达成。我们当然在人与人的互动中使用眼姿势（例如，我们眨眨眼或滚动眼球）。这种眼姿势包括眼睑和眉毛的动作，属于面部表情的一部分。这里介绍的姿势仅限于眼球或凝视方向的运动。这种姿势是可以在眼动仪提供的数据中检测到的。

2000 年，Isokoski 建议使用屏幕外的目标进行文字输入[38]。凝视若要输入字符，必须以一定的顺序来看屏幕外的目标。虽然 Isokoski 没有使用“姿势”一词，但由此产生的眼球

运动就是凝视姿势。然而，屏幕外的目标迫使手势要在一个固定的位置、以一个固定的大小进行。这种姿势依然需要校准眼动仪。

2003 年，Milekic 使用了术语“凝视姿势”[39]。Milekic 概述了在博物馆环境中开发基于凝视的界面的一个概念性的框架，但是他来自艺术教育和艺术心理治疗的一个部门，所以他的方法并不是严格的科学——没有算法，没有用户研究。

与 Isokoski 的凝视姿势相反，由 Wobbrock 等人[40]以及 Drewes 和 Schmidt[41]提出来的凝视姿势规模可大可小，并且可以在任何位置进行。这种姿势的一大优势是，即使没有眼动仪的校准，它们也能够起作用。

8.7.2　姿势检测算法

流行的网络浏览器鼠标手势插件给凝视姿势提供了灵感。此手势插件跟踪鼠标移动并把它转换成代表 8 个方向动作的字符或记号。8 个方向是 U、R、D 和 L，分别代表上、右、下、左。根据键盘上的数字键盘的标准布局，1、3、7 和 9 为对角方向。鼠标手势检测算法接收 x 和 y 坐标。每当这一个或两个坐标都超过开始位置的阈值距离（或网格大小），算法输出一个字符为运动的方向，但它要与上一个字符不同。当前坐标成为新的起始位置，新动作的检测也开始了。其结果是一连串的坐标转换成一连串的字符（见图 8.14）。一串字符描述了一个手势，当算法在一连串的字符中发现手势序列能转换成手势，那么该算法就会给出手势出现的信号来显示手势的出现。

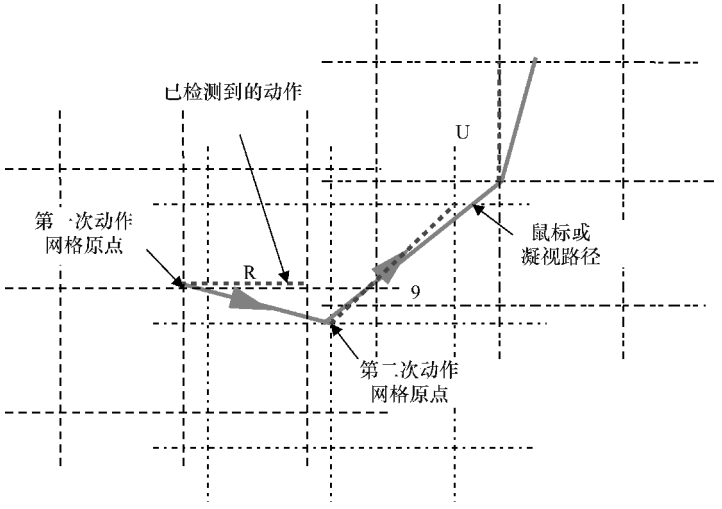


图 8.14　图中显示了鼠标或凝视路径如何转换成字符串 R9U。检测到的动作的终点是下一次动作检测的原点

鼠标手势算法也适用于凝视姿势。有趣的是，凝视运动比鼠标运动更加适合姿势检测。首先，手部动作的自然运动空间是弯曲的，而眼睛的扫视运动是直线。其次，鼠标轨迹是一个连续的坐标系，在同一时刻，这两个坐标都超过阈值是几乎不可能的，这意味着对角线上的笔画很难被检测到。对于对角运动的检测，只有运动的开始点和结束点就很好。凝视运动

的扫视检测提供的正是这一点。

8.7.3　执行凝视姿势的人类能力

作为一种主动的交互方式，凝视姿势最重要的问题是，人们是否能够执行它们。凝视姿势绝对不是很直观的。并不是所有试图执行凝视姿势的人都会立即成功的。要求受试者沿直线移动他们的目光，结果大多数受试者都感到很困惑，同时要求他们按一定顺序观看某些点，结果则更好。因此，为用户提供一些支持点（不是直线）是一个好主意。显示器的四个角能很好地让凝视执行凝视姿势。或者，对话框窗口的四个角也可以。

执行一个姿势的时间是由动作时间和停留时间组成，停留时间被叫作定睛时间。图 8.4 显示了扫视时间与视角的关系，可以转换为一定长度的动作时间。从图 8.4 中，我们得知，长扫视持续时间约为 100 ~ 150ms，而且对扫视的长度只有一点点依赖。因此，执行一个凝视姿势的时间并不取决于姿势的大小，除非该姿势非常小。如果 n 代表动作数，S 代表扫视时间，F 代表定睛时间，一个姿势的总时间 T 为

$$T = nS + (n-1)F$$

从理论上讲，定睛时间可能是零，执行一个姿势的最小时间可能是 120ms 乘以执行姿势时的动作数。实际上，特别是未经训练的用户需要几百毫秒的定睛时间。

8.7.4　凝视姿势字母表

显示屏的 4 个角完美地匹配了正方形姿势，正方形手势也被称为 EdgeWrit 姿势[42]。EdgeWrite 姿势使用的顺序是到达一个正方形的 4 个角的顺序（见图 8.15）。

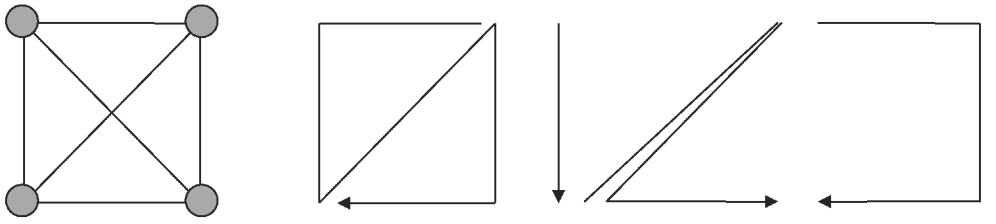


图 8.15　4 个角和用于 EdgeWrite 姿势的 6 条连接线，还有 EdgeWrite 姿势的例子（数字 0、1、2 和 3）

用在姿势检测中介绍的符号很容易描述所有可能的正方形姿势。例如字符串 LD9DL 代表的是图 8.15 的零姿势。另一方面字符串 URUR 不是一个正方形姿势，因此，正方形姿势是鼠标手势的一个子集。尽管如此，正方形姿势还是有能力定义 Wobbrock 等人[42]展示的大型字母表。Wobbrock 等人把拉丁字母表中的每一个字母和数字都分配了至少一个姿势。正方形姿势似乎提供了合适字母表的开始。图 8.16 展示了 4 个或 6 个动作的近似于正方形手势的凝视姿势的 3 个例子。

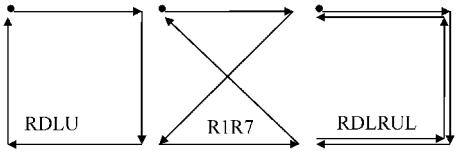


图 8.16　4 个或 6 个动作的正方形闭合凝视姿势的 3 个例子

EdgeWrite 字母表使用与拉丁字母和数字类似的姿势。这使得姿势比较容易被记住，但也暗示着使用姿势进行文本输入。选择字母表取决于姿势的应用。如后面所述，文本输入似乎不是姿势最好的应用。EdgeWrite 字母表来表示凝视姿势存在一个普遍的问题，就是要检测姿势什么时候开始什么时候结束。EdgeWrite 字母表是为手写笔用于手势输入而发明的，并且手势以抬起手写笔而结束。凝视姿势的这种情况就会发生点石成金问题。

凝视姿势的一种可能的应用是远程遥控电视机。对于电视机这样的应用，可由凝视姿势来上下切换频道或调节音量。在这种交互中，同一姿势重复发生是很可能的（例如，把频道向上切换 3 次）。在这种情况下，使用闭合的姿势会更舒适。闭合姿势中，姿势的结束位置也是姿势起始位置。否则，为了重复姿势，凝视必须从姿势的结束位置移动到开始位置。这意味着额外的动作。这也造成了姿势检测识别为另一个姿势的风险。

8.7.5　姿势从自然眼动中分离

鼠标手势算法需要一个手势键，通常是鼠标右键，来让系统检测到手势，否则鼠标移动的正常操作将与手势检测发生冲突。当然，凝视姿势也可以使用与鼠标手势相同的机制，但这会使凝视姿势几乎没有用处。如果执行凝视姿势来触发指令时，某个键一定要被按下，那么按下某个键来触发指令，根本用不上凝视姿势。也许在非常特殊的情况（例如，在只有一个键的移动环境），姿势键才是有意义的，但是，在一般情况下，姿势键会破坏凝视交互的所有好处。因此，有必要把凝视姿势从眼睛的自然运动中分离出来。然而，这似乎不是一件容易的事。

一种可能性是仅在特定的情况下检测姿势。例如，如果要用凝视姿势来关闭对话框，那么当对话框出现时姿势检测开始，对话框关闭时姿势检测结束。这意味着，姿势检测只发生在对话框打开的情况下，并且通常时间很短。如果姿势检测活跃的时间很短，那么自然眼动中出现不经意的凝视姿势的概率是很小的。

另一种把有意的凝视姿势和自然眼动分离开的可能性在于选择合适的姿势。姿势的动作越多，就越不可能发生在自然眼动中。然而，动作增加，凝视姿势需要更多的时间来执行并且姿势很难被记住。对自然眼动的分析显示，某些姿势的出现频率比其他姿势更为频繁。坐在电脑显示器前的人们的眼动包含了许多 RLRLRL 姿势，这是由阅读产生的姿势。姿势的发生频率取决于人们的活动；许多人在打字的时候发生 DUDUDU 姿势，因为他们要低头看键盘再看显示屏。

Drewes 等人引入了第 9 个标记，即冒号，来表示超时的情况。在超时这个概念背后的想法是，姿势应该在短时间内完成。如果凝视在一个网格单元内保持不动，那么检测算法不会报告任何标记。在这种情况下，改良后的算法将产生一个冒号隔开随后的标记和之前的标记。Drewes [43] 用人们上网和观看视频的凝视数据，即不同的参数，改变网格尺寸和不同的超时数据来检测凝视姿势识别。从自然眼动中分离凝视姿势的解决方法竟然是如此出奇的简单：当使用接近显示屏大小的网格时，几乎没有具有 4 个或更多动作的凝视姿势发生。跨越屏幕的长扫视很少发生在自然眼动中，连续 4 个或更多的长扫视更是几乎从来不会发生。

超时参数在价值上不是紧要的，但是很重要。在大的网格尺寸下，凝视通常会在转移到另一个单元前，在上一个单元停留很长一段时间。如果没有冒号，算法可能检测到了长时间的姿势，而这些姿势并不是用户有意为之。只要超时的间隔比凝视在一个单元里平均停留的时间要短，分离就起作用了。进一步减少超时并不提高分离。表 8.2 显示了记录的眼动使用不同值的参数转换为姿势字符串。

表 8.2　使用不同网格大小 *s* 和超时时间 *t* 将同一个眼动（上网）转换成一个姿势字符串

参数	姿势字符串结果
$s=80$	:3LUD:::7R1L9:73LR:73LR:7379RL:U:D:U3:LU:::RL:::R13U:::LR:R:73:73D:73:
$t=1000$	LRLRLR7373DU7LD:RUL13L:R:RL:RL:LRL:R7L3L9R1UR3DR:::7: URLRLRLRDLR7U:R3RL:LR
$s=80$	:3LU:D:::7R:1L9:7:3LR:::73L:R:::737:9:RL:U:D:U3:LU:::RL:::R13U:::LR:R:73:73D:73:
$t=700$	LR:LRLR7373DU7LD:RUL1:3L:R:RL:RL:LRL:::R:7L:3:L9R:1UR3DR:::7: URLRL:RLRDLR7U
$s=250$	:3L:::7R:1U:U3:::7RL:UDU:L:::::::::RD:R:::73:::73:::::L:::L:::R:::L:R:L:R:::R7:R:RL:
$t=1000$	DR:L:::::U:R:R:::LR:L:RL:UD:R:::L:::LR:L:3:L:::::::D:RL:RLRL:::DRL::: RLRLRLRLR:LRLRL:RLR:7:::
$s=250$	:3L:::::7R:1:U:U:3:::7RL:::UDU:::L:::::::::RD:::R:::::73:::73:::::L:::L:::::R:::L:R:::L:::::R7:::
$t=700$	R:::R:L:D:R:L:::::::::U:R:R:::LR:::L:::RL:::UD:R:::::L:::LR:L:::3:::L:::::::D:::RL:::RL: RL:::D:RL:::RL:RL
$s=400$	:3:::L:D:::U3:::::U:::L:::::::RD:::::73:::73:::::L:::::RL:::::U:::RL:R:::::::7:R:::::L:R:7:
$t=1000$	R:::L:::LR:::::L:::R:::::RL:::R:::L:RL:::::RL:R:::L:::RL:RL: RL:::RL:RL:R:::L:::RL:

8.7.6　凝视姿势的应用

通过凝视姿势进行文本输入是可能的，但值得商榷的是它是否有意义。对于一个没有经验的用户而言，输入一个字符花费 1～2s。即使是训练有素的用户要进行带有停留时间的标准的凝视打字，都可能存在问题，通常每个字符需要 5000ms 的停留时间。此外，执行凝视姿势没有直视一个键那么直观，因此没有理由认为用户会更喜欢凝视姿势输入法。

凝视姿势的最大优点之一是它们无需校准即可工作。因此，一个明显的想法是使用凝视姿势来为残疾人调用系统的校准程序。

凝视姿势给不同的人提供了一次性使用的界面，因为没有校准程序。此外，凝视姿势工作不需要接触任何东西，因此可以在高度卫生的环境中使用，如手术室或实验室。凝视姿势比视线接触传感器或其他非接触式的技术，如容量传感器或光电屏障提供更复杂的控制。

凝视姿势是否可以作为电视机的远程遥控器，是否能成为普遍使用的而不是只针对特殊用途的输入技术，这些问题很有趣，并且仍然处于开放的状态。凝视姿势远程遥控的最大优点不需要控制装置 ——不会找不着，也不用为电池充电。然而，一些电视机制造商现在卖的是可以通过手势来控制的电视机。手势控制更加直观并且能给用户带来同样的好处。

应用凝视姿势的另一个领域是移动计算。在移动环境中，解放双手来做其他任务是可取的。凝视界面能让这点变得方便。凝视指向需要在显示屏上有看的对象，因此出于这个原因，显示屏需要增强现实。然而，这种交互方式中凝视的对象会掩盖部分视觉内容。凝视姿势则不需要任何交互对象，并节省显示屏的空间。在增强现实显示屏和图形用户界面的背景下，顺时针方向看向窗口的四角来关闭一个对话框，这样的事情是可以想象得到的。一个 RDLU 姿势和鼠标点击 OK 按钮一样，都能关闭对话框。逆时针方向看窗口的四角就像用一个 NO 来关闭该对话框，如果有必要的话，交叉的姿势像 3U1U 意味着取消操作。

Bulling 等人表示凝视姿势对移动应用同样适用[13]。他们使用了眼电图为他们的研究进行眼动跟踪，这说明凝视姿势的概念不依赖于所使用的眼跟踪技术。

8.8　作为情境的凝视

我们不把用户有意触发指令时的凝视作为一种主动输入的方式，取而代之，我们可能把凝视数据作为情境信息或者使用用户的凝视进行隐式人机交互。计算机利用眼动仪的信息来分析用户的情况和发生的活动，并根据用户当前状态来调整自身行为。把用户的情况考虑到的想法追溯到 1977 年[44]。自那时起，对用户所处的环境和情形的考虑一直是人机交互研究的话题，该话题被称为“情境感知”。

8.8.1　活动识别

用户的当前活动是计算机交互的一个非常重要的方面。因此能否根据用户的眼动来猜测该用户的活动，是个很有趣的问题。

表 8.3 和表 8.4 展示的凝视数据是来自人们看视频和上网，如平均每秒进行的扫视次数、平均扫视的时间和长度、平均定睛时间。

表 8.3　所有参与者（看视频）凝视活动参数的平均值

视频	每秒扫视	每次扫视的像素	平均扫视时间	平均定睛时间	总时间
单位	1/s	像素	ms	ms	s
P1	3.68	101.9	67.8	204.0	216.9
P2	3.43	87.2	69.3	221.9	219.2
P3	3.45	94.1	71.7	213.9	231.3
P4	2.74	107.5	76.3	282.0	228.3
P5	3.24	131.2	73.5	225.6	216.6
P6	2.79	134.1	99.1	259.3	225.1
P7	3.49	102.0	73.9	212.5	217.4
P8	2.76	111.3	91.1	270.6	219.9
平均值	**3.20**	**108.7**	**77.8**	**236.2**	**221.8**
标准差	0.38	16.6	11.2	29.8	
标准差/平均值	11.8%	15.3%	14.4%	12.6%	

表 8.4　所有参与者（上网）凝视活动参数的平均值

上网	每秒扫视	每次扫视的像素	平均扫视时间	平均定睛时间	总时间
单位	1/s	像素	ms	ms	s
P1	5.36	73.0	44.9	141.5	234.0
P2	5.10	73.2	43.9	137.2	229.4
P3	4.54	109.4	70.3	150.0	228.9
P4	5.51	69.0	41.5	140.0	229.8
P5	4.58	105.9	70.2	145.6	291.1
P6	4.59	106.7	54.7	156.0	264.0
P7	4.78	108.0	66.0	143.3	238.2
P8	3.17	123.3	104.5	177.1	374.7
平均值	**4.70**	**96.1**	**62.0**	**148.8**	**261.3**
标准差	0.72	20.9	20.9	12.9	
标准差/平均值	15.4%	21.8%	33.7%	8.6%	

这是统计的特性，即测量值也会不同。因此我们必须回答这个问题，即两个表中的平均值不同是偶然造成还是本身就有显著的不同。这个问题的标准回答是进行 t 检验。t 检验显示了平均值的不同是偶然造成的。表 8.5 显示了一对学生比较两个任务的 t 检验的值。相差显著的值是每秒扫视次数和平均定睛时间。

数据结果对于使用凝视活动来感知情境而言是个好消息。结果的强显著性证明凝视感知系统能够很好地猜测用户的活动。凝视活动参数个体差异性小，让我们完全有理由期待活动识别只需要普遍的阈值就能正常工作，而不需要为个人用户做调整。

表 8.5　“看视频”和“上网”的 t 检验

	每秒扫视	每次扫视的像素	平均扫视时间	平均定睛时间
t 检验	0.00040	0.13136	0.05318	0.00003

起初，这看上去似乎有点矛盾。人们在观看全是动作的视频时眼动比上网浏览静态页面时眼动少。其原因在于在阅读的过程中，扫视和短时间的定睛相间。阅读时，眼球尽可能快地移动，但看电影的时候，眼球只是等待画面出现。

人类的大部分活动都涉及眼动。Land[45]描述日常活动的眼动，如阅读、打字、看图片、绘画、开车、打乒乓球和泡茶。检索情境信息意味着反过来从眼球运动到活动行为。使用这样的方法，Iqbal 和 Bailey 检测凝视模式来识别用户的任务[46]。他们的目标是建立一个注意力管理设备，通过识别心理负荷来在用户的任务序列中减轻破坏性影响。研究表明，每一个任务 —— 阅读、搜索、对象操作和数学推理 ——都有独特的眼动轨迹。

用平均值来识别活动的方法很简单，而超出了本书范围的更复杂的数学运算可以从凝视数据中获得更多信息。Bulling 等人描述过这样一种活动识别方法，即将特征选择的最大相关最小冗余算法（mRMR）和支持向量机（SVM）分类器相结合的识别方法[47]。他们用 6 个不同的活动测试他们的系统：复制文本、阅读印刷纸张、手写笔记、观看视频、浏览网页以及并非具体的活动。他们记录检测精度约 70%。

活动识别的方法很有趣，但普遍的问题是，眼动的分析可以告诉用户过去的任务是什么，而不能预测该用户将要干什么。提供分析所需的数据期间导致了延迟。不清楚任务识别工作有多可靠，最后，不需要社会智力来做正确的决定。

活动识别的路径很有趣，但其普遍问题在于眼睛运动的分析可以说明用户过去的任务，但却不能预测用户未来的任务。提供分析所需数据的期间会导致延迟。同时，任务识别的可靠性暂不清楚，最后，做出正确决定所需的社交智能概念并不存在。

8.8.2　阅读检测

阅读是我们常做的特定活动，在与计算机设备交互的情况下尤其常见。阅读时的眼动在 19 世纪就已经是一个研究课题。Javal（1897）和 Lamare（1892）观察了阅读时候的眼动并引用法语词“saccade（扫视）”来表达眼睛的突发运动。心理学家对于阅读过程有过深刻研究，对其细节理解深入[48]。

很多情况下，尤其是在上网时，人们通常不会仔细阅读，而且常常不会完整阅读文本。Jakob Nielsen 对几百个受试者进行过大型阅读网页习惯调查。他利用热度图视觉化了眼睛凝视活动，发现大多数热度图呈“F”形状。这意味着读者常常只读开始几行。鉴于此，网页应当将重要事情放在开始几行。

对阅读的凝视分析对于寻找交互的设计规则很有启发。然而，若分析是实时进行，则对用户更有利。例如，若系统了解用户正在阅读，就可以停止显示分心的动漫，并延迟扰人的通知。参考文献［49－51］中对于阅读检测推荐了数种算法。

阅读检测大体上不算太难。一系列的前向扫视和随后的一个后向扫视是阅读活动的有力指示。上一章介绍的姿势识别算法，在有人阅读时，会产出 RLRLRL 姿势。阅读检测的问题在于延迟和可靠度。阅读检测器需要几次扫视作为输入，来知道用户在阅读，因此在用户刚开始阅读时无法检测用户开始阅读。因此，阅读检测在检测短篇或单行文本的阅读时有问题。阅读检测可能在用户做其他事情的时候，显示在做阅读活动。例如，这种情况会出现在看群体照片中人们头部的时候，因为这时的凝视姿势与阅读时相似。

检测阅读活动是有用的，但若系统可以知晓阅读内容，则帮助更大。对此，值得对阅读时的凝视进行更仔细的了解。图 8.17 所示为阅读文本时的凝视路径。

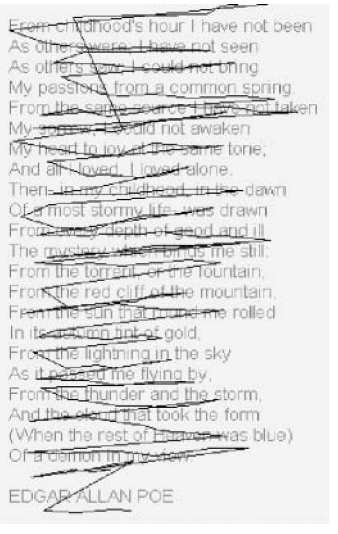


图 8.17　阅读文本时的凝视路径

很容易看到，扫视期间的凝视会有短暂的定睛。有时会有回看，尤其是在看困难的文本时。数据分析得出的扫视长度大约在每行 1°，而后向扫视长度比一行长度略短。前向扫视的角度与中央凹的角度相同，这很合理，因为这意味着对这一行的覆盖是最优

的，既没有空缺也没有重复。因此，要注意一个有趣的现象：一行中的第一批定睛点大约在第一行开始后半度，而最后的定睛与这一行末尾也是同样距离。在垂直方位，低精度导致的问题是，如果行高在正常范围内，则无法准确测出凝视阅读哪一行。图 8.17 中的行高大约是 0.5°。

很多职业都需要读许多很长的文件。智能系统能够掌握文档的阅读情况，并将这个信息送给用户。这便提出了如何将凝视活动转成说明文档阅读情况的数字。

参考文献［52］中呈现的想法就是将虚拟单元装满到文本中去。算法可以计算每个单元中的定睛次数。某文档中定睛的总次数提供有用信息，但是却无法显示文档是否已经读完，或文档的 1/3 已经读过 3 遍 。因此，一个数字不能够对文档阅读情况提供较好指示。需要获得第二个值，才能指示出定睛在整个文本中的分布。

第二个值的一个可能定义是单元内定睛方差。低方差表明凝视在文本中均匀分布。另一种可能的定义是文本内视线扫过的单元的百分比。用这个值来描述文件是否被完全阅读，是很容易被人理解的。但它不提供是否该文件被阅读数次的信息。

阅读质量的值对查找未读文档很有帮助。文档本身也能就凝视数据提供反馈。例如，文档可以把已经阅读过的文档用不同的背景颜色在显示屏上显示。

阅读检测是眼睛凝视情境信息并不简单也不太模糊的一个例子。当我们从心理学的角度上看精心制作的阅读模型[48]，我们就能清晰地看到阅读检测的潜力。阅读速度、向后扫视次数和阅读困难文字所需的时间，这些都有可能让我们得到用户的阅读能力的信息。通过用不同的语言或脚本显示文本，可以找出用户使用什么语言阅读。网上书店以后推荐书目的时候可以使用这些信息。这里肯定还有留待进一步研究的空间。

8.8.3　注意力检测

使用凝视情境信息最明显的方法就是凝视作为注意力的指标。在大多数情况下，我们看我们关注的对象。这听起来可能微不足道，但注意力是非常强大的情境信息，它可以为用户提供真正的好处。电子设备如台式计算机或移动电话的显示器给用户提供信息。然而，当用户不看屏幕的时候，就没有必要显示信息。现在，系统通常不能感知用户的注意力，它们会在用户离开的时候推送很重要的信息。然而当用户返回时，该信息可能已经被下一条信息覆盖。当用户的注意力回到系统的时候，注意力感知系统和眼动仪一起可以告诉用户刚刚发生的事情的概况。

记录用户的注意力，为她或他在特定的文件上花了多少时间提供了可靠的数据。文件被显示的时间是不可靠的，因为用户可以打开文档，然后离开去拿咖啡。关于哪个文件花了多少时间的统计数据是非常有用的。工作的时候，有人要做几个项目，有必要计算出每个项目的成本，这意味着需要知道每个项目花费多少时间。这样的统计数据也可用于电子学习。

8.8.4　应用凝视情境

凝视对人类互相交流而言非常重要。如果我们想要计算机很好地协助我们，计算机必须知道我们正在看的是什么。然而，诠释凝视是很困难的，因为即使是人类也不能总是从别人的眼中读出别人的期望。正确解读凝视的方式需要社会智力，而在计算机上不容易实现这点。

然而在许多情况下，简单的方法可以给用户带来好处。如果有人在看，显示屏就会自动开启，当没有人看的时候，显示屏就会自动关闭。这不仅是为了方便用户，还可以节省电量，移动设备的电量都是有限的。另一个例子是，如果系统识别出计算机显示的消息已被阅读，鼠标不点击的话，该消息就可能会消失。

另一个例子是由 Kern 等人[54]介绍的凝视标识。当我们看向其他地方又要看回原来位置的时候，凝视标识就是我们放在地图上或者是文档中的手指的替代。凝视标识是一个视觉占位符，突出我们在显示屏上看的最后一个位置。凝视标识在多显示器的情况下很有帮助，或者是我们需要在显示屏和实体文件中转换注意力的时候也很有帮助。凝视标识在汽车中的用户界面也许会更有用，我们与导航系统的交互可能会被交通情况中断。因为汽车移动时，导航系统显示的内容可能会发生改变，这意味着我们重新看向显示屏的时候需要一些时间找到新方向。在这种情况下，凝视标识必须随着地图移动，不能一直固定在显示屏上。由于我们的注意力应该主要集中在驾驶汽车上，如果能更快地在界面上找到定位并且交互的时间更短，这将是一个很大的优势。

上述例子清楚地表明，凝视界面不仅是指挥计算机的一种可行方式，也有很大的潜力成为一种新型的辅助系统。

8.9　展望

正如在引言中所提到的，有许多原因证明把视线作为交互方法是可取的。如果我们想要使用与人类相似的方式进行交互的计算机设备，特别是对于类人的机器人，那么眼跟踪技术是必需的。

纵观 20 年以来的眼跟踪研究，似乎我们期待利用视线信息的方式随着时间发生了改变。早期的研究主要集中在用视线来操作图形用户界面，目前的重点似乎是视线感知的应用程序。

凝视指向有一定的困难要克服——点石成金问题、精度问题和需要校准的问题。当处理文字或电子表格应用程序，又或者操作自动柜员机时，我们并不需要眼交互的速度优势。少数情况下，如玩射击游戏或军事上类似的射击活动，我们需要速度。大众市场的眼动仪硬件最有可能作为游戏机的附加产品。

只要有廉价的眼动仪，我们可以期待游戏领域以外的进一步应用。然而，我们用眼动仪

完全替代如今的鼠标操作，完成来操作图形用户界面，却似乎不可能。

眼动仪似乎给予我们更多的功能来使系统更加智能。眼动仪可以通过定位多显示器中的鼠标指针来协助我们；它可以跟踪和记录我们的眼球活动，并且能告诉我们已经读了哪个文件或者文件的哪部分；当我们阅读文本时，数字百科全书的动画形象可能会暂停。

最近几年我们与移动设备的交互急剧增加。移动设备用于眼跟踪有两种选择。一种选择是放置在移动设备中的眼动仪。这样的眼动仪面临的挑战是对手部动作和不断变化的光线条件的补偿。另一种选择是头戴式眼动仪。头戴式眼动仪似乎更容易实现，但肯定是很突兀的。然而，与眼镜式的增强现实显示相结合是有意义的。

视觉和操作界面之间可能存在冲突的基本问题在移动计算领域中特别严重。这可能也会与我们的社交礼仪相冲突。社交礼仪要求我们直接把凝视对准正在和我们谈话的人。移动计算对不用手控制设备有很强的要求。语音命令是我们的一个选择，但社交礼仪在很多情况不允许使用语音。凝视控制不需要手，并且是无声的。也许凝视姿势在增强现实显示的切换模式上很有用，如开启或关闭显示屏。

在把眼动仪作为界面技术介绍给大众的前一步似乎是介绍注意力传感器。注意力传感器不能报告凝视的方向，只能报告是否有人在看。注意力传感器更容易制作且不需要校准。一些手机厂商已经在市场上有凝视感知的设备了，它能在没人看的时候，关掉显示器省电。凝视感知已经存在。

参 考 文 献

1. Fitts, P.M. (1954). The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology* **47**, 381–391.
2. Ware, C., Mikaelian, H.H. (1987). An Evaluation of an Eye Tracker as a Device for Computer Input. *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics interface* CHI '87, 183–188.
3. Miniotas, D. (2000). Application of Fitt's Law to Eye Gaze Interaction. *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, 339–340.
4. Zhang, X., MacKenzie, I.S. (2007). Evaluating Eye Tracking with ISO 9241 – Part 9. *Proceedings of HCI International 2007*, 779–788.
5. Vertegaal, R.A. (2008). Fitt's Law comparison of eye tracking and manual input in the selection of visual targets. *Conference on Multimodal interfaces* IMCI '08, 241–248.
6. Carpenter, R.H.S. (1977). *Movement of the Eyes*. Pion, London.
7. Abrams, R., Meyer, D.E., Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the occulomotor system. *Journal of Experimental Psychology: Human perception and performance* **15**, 529–543.
8. Tian, Y., Kanade, T., Cohn, J.F. (2000). Dual-State Parametric Eye Tracking. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 110–115.
9. Ravyse, I., Sahli, H., Reinders MJT, Cornelis J. (2000). Eye Activity Detection and Recognition Using Morphological Scale-Space Decomposition. *Proceedings of the international Conference on Pattern Recognition* ICPR. IEEE Computer Society, **1**, 1080–1083.
10. von Romberg, G., Ohm, J. (1944). Ergebnisse der Spiegelnystagmographie. *GräfesArch Ophtalm* **146**, 388–402.
11. Robinson, D.A. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans Biomed Eng (BME)* **10**, 137–145.

12. Brown, M., Marmor, M., Zrenner, V., Brigell, E., Bach, M. (2006). ISCEV Standard for Clinical Electro-oculography (EOG). *Documenta Ophthalmologica* 2006 **113**(3), 205–212.
13. Bulling, A., Roggen, D., Tröster, G. (2008). It's in Your Eyes: Towards Context-Awareness and Mobile HCI Using Wearable EOG Goggles. *Proceedings of the 10th international Conference on Ubiquitous Computing, UbiComp '08*, **344**, 84–93.
14. Vertegaal, R., Mamuji, A., Sohn, C., Cheng, D. (2005). Media Eyepliances: Using Eye Tracking for Remote Control Focus Selection of Appliances. *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 1861–1864.
15. Kim, K.-N., Ramakrishna, R.S. (1999). Vision-based Eye-gaze Tracking for Human Computer Interface. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, **2**, 324–329.
16. Dongheng, L., Winfield, D., Parkhurst, D.J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **3**, 79.
17. Nguyen, K., Wagner, C., Koons, D., Flickner, M. (2002). Differences in the infrared bright pupil response of human eyes. *Proc. Symposium on Eye Tracking Research & Applications (ETRA 2002)*, 133–138.
18. Ohno, T., Mukawa, N., Yoshikawa, A. (2002). FreeGaze: a gaze tracking system for everyday gaze interaction. *Proceedings of the symposium on ETRA 2002: eye tracking research & applications symposium*, 125–132.
19. Ohno, T., Mukawa, N. (2004). A Free-head, Simple Calibration, Gaze Tracking System That Enables Gaze-Based Interaction. *Proceedings of the symposium on ETRA 2004: eye tracking research & application symposium*, 115–122.
20. Hennessey, C., Noureddin, B., Lawrence, P. (2006). A single camera eye-gaze tracking system with free head motion. *Proceedings of the 2006 symposium on Eye tracking research & applications*, 87–94.
21. Zhai, S., Morimoto, C., Ihde, S. (1999). Manual And Gaze Input Cascaded (MAGIC) Pointing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '99*, 246–253.
22. Morimoto, C.H., Koons, D., Amir, A., Flickner, M. (1999). Frame-Rate Pupil Detector and Gaze Tracker. *Proceedings of the IEEE ICCV'99 frame-rate workshop.*
23. Jacob, R.J.K. (1990). What You Look At is What You Get: Eye Movement-Based Interaction Techniques. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '90*, 11–18.
24. Majaranta, P., Räihä, K. (2002). Twenty Years of Eye Typing: Systems and Design Issues. *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications ETRA '02*, 15–22.
25. Bolt, R.A. (1981). Gaze-orchestrated Dynamic Windows. *Proceedings of the 8th Annual Conference on Computer Graphics and interactive Techniques, SIGGRAPH '81*, 109–119.
26. Bolt, R.A. (1982). Eyes at the Interface. *Proceedings ACM Human Factors in Computer Systems Conference*, 360–362.
27. Card, S.K., Moran, T.P., Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum.
28. Dickie, C., Vertegaal, R., Sohn, C., Cheng, D. (2005). eyeLook: using attention to facilitate mobile media consumption. *Proceedings of the 2005 ACM Symposium on User Interface Software and Technology 2005*, 103–106.
29. Drewes, H., Schmidt, A. (2009). The MAGIC Touch: Combining MAGIC-Pointing with a Touch-Sensitive Mouse. *Proceedings of Human-Computer Interaction – INTERACT 2009*, 415–428.
30. Salvucci, D.D., Anderson, J.R. (2000). Intelligent Gaze-Added Interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '00*, 273–280.
31. McGuffin, M., Balakrishnan, R. (2002). Acquisition of Expanding Targets. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '02*, 57–64.
32. Zhai, S., Conversy, S., Beaudouin-Lafon, M., Guiard, Y. (2003). Human On-line Response to Target Expansion. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '03*, 177–184.
33. Miniotas, D., Špakov, O., MacKenzie, I.S. (2004). Eye Gaze Interaction with Expanding Targets. *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, 1255–1258.
34. Špakov, O., Miniotas, D. (2005). Gaze-Based Selection of Standard-Size Menu Items. *Proceedings of the 7th international Conference on Multimodal interfaces, ICMI '05*, 124–128.
35. Ashmore, M., Duchowski, A.T., Shoemaker, G. (2005). Efficient eye pointing with a fisheye lens. *Proceedings of Graphics Interface 2005 (GI '05)*, 203–210.
36. Kumar, M., Paepcke, A., Winograd, T. (2007). EyePoint: Practical Pointing and Selection Using Gaze and Keyboard. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '07*, 421–430.
37. Miniotas, D., Špakov, O., Tugoy, I., MackKenzie, I.S. (2005). Extending the limits for gaze pointing through the use of speech. *Information Technology and Control* **34**, 225–230.

38. Isokoski, P. (2000). Text Input Methods for Eye Trackers Using Off-Screen Targets. *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, ETRA '00*, 15–21.
39. Milekic, S. (2003). The More You Look the More You Get: Intention-based Interface using Gaze-tracking. In: Bearman D, Trant J. (eds.). *Museums and the Web 2002*: Selected Papers from an International Conference. Archives & Museum Informatics, Pittsburgh, PA.
40. Wobbrock, J.O., Rubinstein, J., Sawyer, M., Duchowski, A.T. (2007). Not Typing but Writing: Eye-based Text Entry Using Letter-like Gestures. *Proceedings of COGAIN*, 61–64.
41. Drewes, H., Schmidt, A. (2007). Interacting with the Computer using Gaze Gestures. *Proceedings of Human-Computer Interaction – INTERACT 2007*, 475–488.
42. Wobbrock, J.O., Myers, B.A., Kembel, J.A. (2003). EdgeWrite: A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion. *Proceedings of the 16th Annual ACM Symposium on User interface Software and Technology, UIST '03*, 61–70.
43. Drewes, H. (2010). *Eye Gaze Tracking for Human Computer Interaction*. Dissertation an der Ludwig-Maximilians-Universität München.
44. Hull, R., Neaves, P., Bedford-Roberts, J. (1997). Towards Situated Computing. *Tech Reports*: HPL-97-66, HP Labs Bristol.
45. Land, M.F. (2006). Eye movements and the control of actions in everyday life. *Prog Retinal & Eye Res* **25**, 296–324.
46. Iqbal, B., Bailey, P. (2004). Using Eye Gaze Patterns to Identify User Tasks. *Proceedings of The Grace Hopper Celebration of Women in Computing*.
47. Bulling, A., Ward, J.A., Gellersen, H., Tröster, G. (2009). Eye movement analysis for activity recognition. *Proceedings of the 11th international conference on Ubiquitous computing*, 41–50.
48. Reichle, E.D., Pollatsek, A., Fisher, D.L., Rayner, K. (1998). Toward a Model of Eye Movement Control in Reading. *Psychological Review* **105**, 125–157.
49. Campbell, C.S., Maglio, P.P. (2001). A Robust Algorithm for Reading Detection. *Proceedings of the 2001 Workshop on Perceptive User interfaces, PUI '01*, **15**, 1–7.
50. Keat, F.-T., Ranganath, S., Venkatesh, Y.V. (2003). *Eye Gaze Based Reading Detection. Conference on Convergent Technologies for Asia-Pacific Region, TENCON '03*, **2**, 825–828.
51. Bulling, A., Ward, J.A., Gellersen, H., Tröster, G. (2008). Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography. *Proc. of the 6th International Conference on Pervasive Computing (Pervasive 2008)*, 19–37.
52. Drewes, H., Atterer, R., Schmidt, A. (2007). Detailed Monitoring of User's Gaze and Interaction to Improve Future E-Learning. *Proceedings of the 12th International Conference on Human-Computer Interaction HCII '07*, 802–811.
53. Kern, D., Marshall, P., Schmidt, A. (2010). Gazemarks – Gaze-Based Visual Placeholders to Ease Attention Switching. *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (CHI '10)*. ACM, Atlanta (GA), USA.
54. Wobbrock, J.O., Rubinstein, J., Sawyer, M.W., Duchowski, A.T. (2008). Longitudinal Evaluation of Discrete Consecutive Gaze Gestures for Text Entry. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ETRA '08, 11–18.

第9章

感知用户界面的多模态输入

Joseph J. LaViola Jr. , Sarah Buchanan, Corey Pittman
美国佛罗里达州中佛罗里达大学

9.1 引言

自从 Bolt 发表了开创性的论文《Put that there: Voice and Gesture at the Graphics Interface》，可用于计算机应用程序交互的多模态输入就成为了人机交互研究的一个活跃领域[1]。这种不同形式的输入组合（例如，语音、手势、触摸和眼睛凝视）被称为多模态交互模式，其目标是向用户提供与计算机进行交互的多种选择方式，以支持自然的用户体验。这些方式可以帮助简化界面，以便在使用识别技术时能有更稳定的输入，以及支持更逼真的交互场景，因为界面可以更精妙地协调人类通信系统。从理论的角度看，多模态界面以协调的方式处理两个或更多个输入模式，其目的是识别天然形成的人类语言和行为，一般包括一个以上的识别技术[2]。

随着更强大的感知计算技术的出现，多模态界面因为可以被动地感知用户正在做的事情而变得更加突出。这些界面也被称为感知用户界面[3]，它们的传感器装置在物理环境中而不是在用户身上，因而提供了支持非侵入的交互机制。本书前面章节已经重点介绍了各种输入技术和相关的交互模态。在本章中，我们将研究如何将这些不同的技术以及它们的输入模式——特别是语音、手势、触摸、眼睛凝视、面部表情和脑机输入——结合一体及其所能提供的交互类型。我们也将研究综合这些输入模式的策略，也被称为多模态整合或融合。最后，我们将探讨一些多模态界面的可用性问题和处理这些问题的方法。研究多模态界面跨越多个领域，包括心理学、认知科学、软件工程学和人机交互等。

本章的重点将是使用多模态输入的界面类型。更全面的调查详见参考文献［5，6］。

9.2 多模态交互类型

相比传统的单一界面，多模态界面可以被定义为多个输入模态的组合，以提供给用户更

丰富的交互集。输入模态的组合可以分为6种基本类型：互补型、重复型、等价型、专业型、并发型以及转化型[7]。在本节中，我们将逐一对其做简要定义：

- 互补型：当两个或多个输入模态联合发出一个命令时，它们便会相得益彰。例如，为了实例化一个虚拟对象，用户做出指示手势，然后说话。语音和手势相得益彰，因为手势提供了在哪里放置对象的信息，而语音命令则提供了放置什么类型的对象的信息。
- 重复型：当两个或多个输入模态同时向某个应用程序发送信息时，它们的输入模态是冗余的。通过让每个模态发出相同的命令，多重的信息可以帮助解决识别错误的问题，并加强系统需要执行的操作[8]。例如，用户发出一个语音命令来创建一个可视化工具，同时也做一个手势表示该工具的创建。当提供多于一个的输入流时，该系统便有更好的机会来识别用户的预期行为。
- 等价型：当用户具有使用多个模态的选择时，两个或多个输入模态是等价的。例如，用户可以通过发出一个语音命令，或从一个虚拟的调色板中选择对象来创建一个虚拟对象。这两种模态呈现的是等效的交互，且最终的结果是相同的。用户也可以根据自己偏好（他们只喜欢在虚拟调色板上使用语音输入）或规避（语音识别不够准确，因此他们改用调色板）来选择使用的方式。
- 专业型：当某一个模态总是用于一个特定的任务时它就成了专业的模态，因为它是比较适合该任务的，或者说对于该任务来说它是当仁不让的了。例如，用户希望在虚拟环境中创建和放置一个对象。对于这个特定的任务，做出一个指向的手势确定物体的位置是极具意义的，因为对于放置物体可能使用的语音命令范围太广，并且一个语音命令无法达到对象放置任务的特定性。
- 并发型：当两个或者两个以上的输入模态在同一时间发出不同的命令时，它们是并发的。例如，用户在虚拟环境用手势来导航，与此同时，使用语音命令在该环境中询问关于对象的问题。并发型让用户可以发出并行指令，其体现为在做晚餐的同时也可打电话这样的真实世界的任务。
- 转化型：当两个输入模态分别从对方获取到信息时它们就会将信息转化，并使用此信息来完成一个给定的任务。多模态交互转化的最佳例子之一是在多模态交互的一键通话界面里[9]，语音模态从一个手势动作获得信息，告诉它应激活通话。

9.3　多模态界面

本节中，我们研究在本书中讨论过的不同的技术和输入模态是怎样被用作多模态交互系统的一部分。需要注意的是，尽管语音输入是多模态界面的主要方式，但我们在本章中没有专门介绍语音部分。相反，语音是作为每种模态的一个子部分。

9.3.1　触控输入

近年来，随着多点触控手机、平板电脑、笔记本电脑、桌面电脑和显示屏等的日益普

及，多点触控设备变得越来越普遍。因此，多点触控手势成为用户的日常词汇的一部分，如滑动解锁或缩放屏幕。然而，复杂的任务，如 3D 建模或图像编辑，单独使用多点触控输入是很困难的。多模态交互技术的设计能将多点触控界面与其他输入融合，如语音，为复杂任务创建了更直观的交互。

9.3.1.1　3D 建模和设计

大型多点触控显示器和桌面电脑常常被营销成能够促进协作的自然界面。然而，这些产品往往针对在公共环境中的商业客户，并作为一个新奇物件作为宣传。因此现在的问题仍然是，它们是否能在具有有效性的同时也提供独特的用户体验。由于鼠标和键盘都不再可用，语音可以为之前应用的 WIMP 范式提供操作环境，如复杂的工程应用程序（例如，Auto CAD）。例如，MozArt 公司[10]结合了语音命令与可倾斜的多点触控桌面，创建了一个可以创造 3D 模型的简易界面，如图 9.1 所示。一项研究对 MozArt 公司的产品进行了评估，让新手使用 MozArt 和另一个多点触控的 CAD 软件，并进行了比较。大多数用户优选的是多模态界面，尽管该结论需要更多的用户进行测试才能考量其准确性和有效性。类似的界面可以通过结合语音和触摸来改善，正如在一项关于用单一多点触控界面执行 3D CAD 操作的项目中所提到的一样[11]。

图 9.1　MozArt 公司桌面硬件原型。来源：Sharma A，Madhvanath S，Shekhawat A and Billinghurst M 2011。经授权转载

9.3.1.2　协作

大型多点触控显示器以及桌面电脑用于协作是最为理想的，原因是它们有 360°触控界面，即一个大型显示屏桌面，并且它们也支持多种输入源。举个例子，Tse 等人（2008）[12]开发出了一个名为“设计师环境”的多模态多点触控系统，该系统能通过用户手势或语音发出指令控制一个设计应用。它是基于工业设计师常用于头脑风暴的 KJ 创意方法，该方法有以下四个步骤：

1）创建笔记。

2）小组笔记。

3）标记各组。

4）关联各组。

在“设计师环境”这一应用中，多个用户可运用触控结合手势和语音输入指令完成各种任务，如图 9.3 所示。然而，Tse 等人（2008）[12]发现，这一应用仍有一些未解决的问

题，如并行工作、模态转换、个人及集体领域，还有联合多模态指令等。Tse 等人对这些问题提出了相应对策，如在桌面创建个人工作区域来解决并行工作问题。

Tse 等人（2006）[13]也曾经开发了 GSI Demo（演示创建手势与语音基础结构系统）。这一系统通过在已有的鼠标/键盘应用上创建多用户语音或手势输入包装器，演示了多模态交互。GSI Demo 能够有效将单一用户桌面应用转化成多点触控桌上应用，例如地图、指令与控制模拟器、模拟与训练、游戏等。Tse 等人（2007）[14]特别讨论了使用这一多点触控桌面系统，用户可以协作共同玩暴雪公司的魔兽世界 3 和模拟人生游戏。他们提出的界面允许玩家使用手势或语音输入指令创造一种全新的多人参与的体验，更接近街机游戏对人们社会需求的满足，如图 9.2 所示。

图 9.2　两人在魔兽世界 3（左图）及模拟人生游戏（右图）中互动
来源：Tse E，Greenberg S，Shen C，Forlines C and Kodama R 2008。授权转载

协作情境还有另一个有趣的一面，那就是可追溯成员在协作过程中的行为和言语。协作数据清楚揭示了学习和协作的过程。这类数据也可作为机器学习以及数据挖掘算法的输入，以提供应景的反馈及个性化内容。

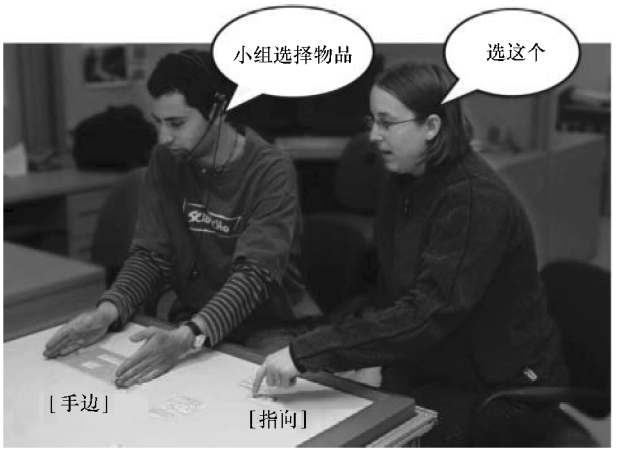


图 9.3　在“设计师环境”中的两人组合手势。来源：Tse E，Green－berg S，Shen C and Forlines C 2007。授权转载

“Collaid（协作学习辅助）”在桌面学习活动中是一个捕捉多模态数据的环境[15]。数据收集时使用了一组麦克风和一个传感器，与学习系统中其他部分形成了一个整体，最后经过转化，桌面协作过程被转化成为可视过程，展现了正在桌边发生的协作过程。图 9.4 是一个协作小组和另一个协作较少的小组之间的对比数据可视化实例。其他运用分布式白板进行多模态协作的研究可

见参考文献［16］。

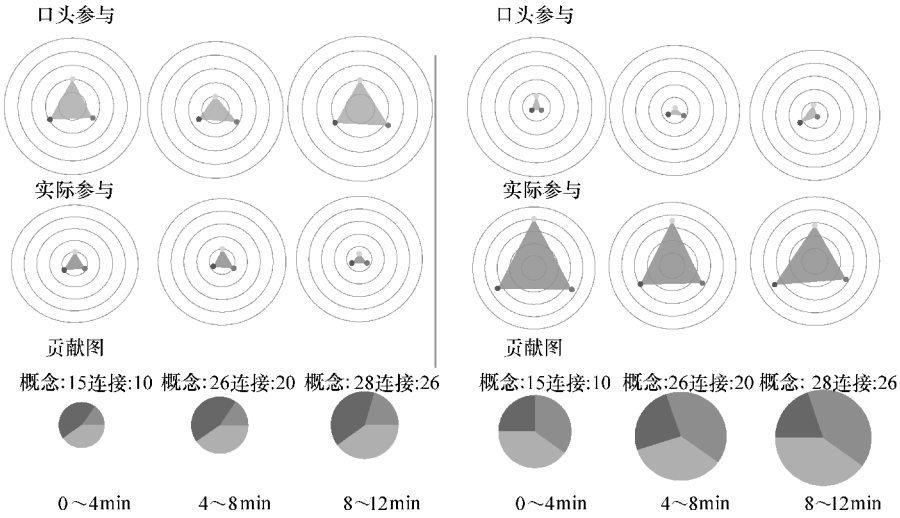


图 9.4　一个交流频繁的小组（左图）以及一个协作较少的小组（右图）12min 活动的协作可视化
来源：Martínez – Maldonado R，Collins A，Kay J and Yacef K 2011。授权转载

9.3.1.3　与残疾或老年病人交流

多模态桌面应用也能支持与听力残障病患交流的功能[17]，并且进一步研究发现，通过转录语音可以增进医生与老年患者的交流[18]。这个“共享语音界面”（SSI）是为多点触控桌面显示器开发的一项应用，用于支持听力残障病人与不会手语的听力复健医生之间的交流。听力残障的病人只需敲敲键盘，听力复健医生对着耳机上的耳麦说话。两人交流时，他们的话语会被转录，然后以可移动的对话框的形式出现在显示屏上，如图 9.5 所示。多模态界面技术还可以造福其他有不同交流需求的人群。例如，一位学外语的学生可以一边听到教师念词组的音频片段，一边可以获取其听到的语音表述的文字。

图 9.5　一位医生（图左侧）和病人（图右侧）在用“共享语音界面（SSI）”交流。同时可移动的对话框出现在多点触控屏幕上。
来源：Piper AM 2010。经 Anne – Marie Piper 授权转载

9.3.1.4　移动设备搜索

移动设备用户现在越来越精通他们设备的使用，他们希望在执行多个任务或奔忙之时能用到移动设备，像是开车或是想快速寻找资讯的时候。另外，现有的移动设备包含了广泛的输入技术，例如多点触控界面、麦克风、摄像头和全球定位系统，还有加速计。移动设备应用需要利用多种输入模态，可以不要求用户停下手中的事情，而在用户繁忙之时能够快捷地输入输出信息。

移动设备也面临着挑战，如数据转化更缓慢了，显示屏和键盘更小了，因此它们也在开发自己的应用，使桌面范式不再适合了。有人认为语音输入能轻易解决这些问题，然而单纯运用语音输入是不现实的，因为语音识别容易出错，尤其在嘈杂环境下，它不能提供精确的控制。许多多模态移动界面在不断涌现，它们运用了语音输入并巧妙结合了其他交互形式。例如，语音输入可以利用语言为操作提供语境信息，而把精确的控制问题交给直接的触控输入模式。或者语音输入可与文本录入同时进行，以保证录入文本的正确性。

语音输入是移动设备搜索的一个理想输入形式，快捷又便利。然而，由于语音输入容易出错，校正工作也应快捷方便。“声音搜索系统”为移动设备的搜索功能提供多模态校正[19]。用户说出查询内容之后，系统会根据查询内容给出识别结果的多元最佳列表（N - best list）。多元最佳搜索结果由字板组成，这一字板让用户能根据搜索结果，运用触控输入的方式轻松地重新排列并查询新内容。

移动设备搜索还包括局部搜索，这一搜索方式在现有的移动设备技术中让用户非常满意，它可以将搜索范围限定在当前位置。“搜话（Speak4it）”就是一款改进了语音搜索的移动设备应用，它让用户用手指在他们想查询的位置上书写[20]。“搜话（Speak4it）”支持多模态输入法，用户可以用语音或打字的方式输入搜索条件，在想要查询的位置用触控输入法轻划。一个“搜话（Speak4it）”的语境范例就是骑行者可以用语音或手势搜索路上最近的修车行，得到更为精确的搜索结果。例如，可以用语音输入查询：“斯泰弗森特镇修自行车的商铺”，再在显示屏上画下一条路，该应用就会给出反馈，告知显示屏上标记的这一路段上的各个搜索结果（见图 9.6）。

图 9.6　“搜话（Speak4it）”手势输入。经 Patrick Ehlen 授权转载

具备这些能力的研究技术原型已经存在多年，如 QuickSet[21]。然而，这些技术却是在能用触摸屏输入、能进行语音识别、能上网的移动设备被广泛使用之后才进入普通用户的视野。另外 Ramsay 等人（2012）的“火车系统（Tilt and Go system）”也对多模态交互的移动设备搜索进行了探索[22]。Feng 等人（2011）介绍了语音和移动设备搜索多模态交互的详细分析[23]。

9.3.1.5　移动设备文本录入

在触摸屏显示器上用软键盘打字录入文本对许多用户是再平常不过了，但这很费时间。有两个办法可以快速录入文字，分别是手势键盘输入和语音输入。手势键盘输入让用户可快速在熟悉的标准键盘上滑动划出文字路径，巧妙规避了打字过程。然而，要预测手势是非常模棱两可的。语音输入这个选项非常吸引人，它完全不需要打字。然而，语音输入依赖于自动语音识别技术，在嘈杂环境或非母语用户来说效果欠佳。“边说边滑”（SAYS）[24]是一个结合手势键盘和语音识别的多模态界面，用于改善文本录入的效率和准确性，如图 9.7 所示。滑动手势和语音输入为语言预测提供补充信息，让 SAYS 系统能从周围声音智能提取有用的线索，改进语言预测的准确性。另外，SAYS 是在之前研究的基础上[25]建立起来的，它使得持续同步的输入方式成为可能。

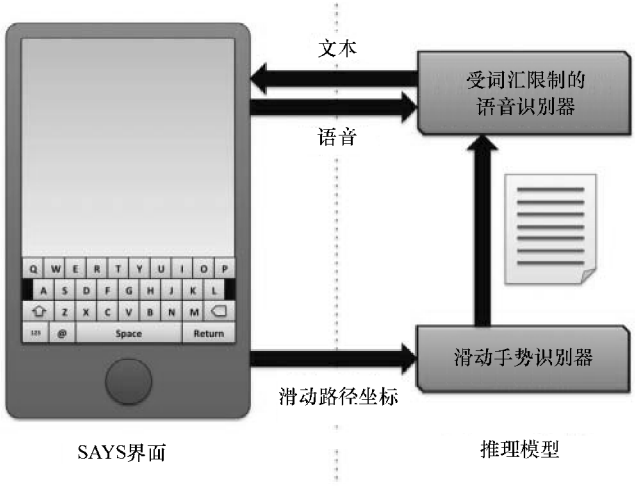


图 9.7　“边说边滑（Speak As You Swipe）”界面。来源：经 Khe Chai Sim 授权转载

Shinoda 等人（2011）开发了一个类似的界面[26]，能支持移动环境半同步语音和手写输入，如图 9.8 所示。语音和手写之间有固有的时间差，很难应用传统多模态识别算法。要解决这个时间差，他们开发了一个多模态识别算法，运用了分段式统一方案以及适应用户个人时间差特性的方式。该界面

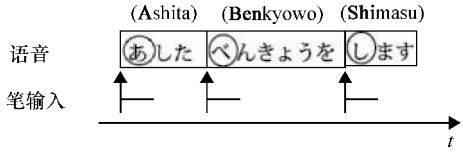


图 9.8　语音与手写输入的关系。来源：Shinoda K，Watanabe Y，Iwata K，Liang Y，Nakagawa R and Furui S 2011。授权转载

也可支持键盘输入，可通过多种不同方式检验：

1）用户在一句话中写下每个短语的初始字符。

2）用户写下如（1）中初始字符的第一划。

3）用户输入笔触提示每个短语的开头。

4）用户敲击每个短语首字符所属字符表。

5）用户用标准键盘输入每个短语初始字符。

这 5 个不同笔头输入界面都使用过嘈杂的语音数据进行评估，系统识别准确率比在 5 个界面中单纯使用语音要更高。它们也为每个界面进行了可用性测试，找到了识别可用性与性能改进之间的平衡。

其他研究比较了在不同的多模态交互策略的指导下使用触控输入进行文本录入，具体见参考文献［27］。

9.3.1.6　移动设备图片编辑

另一个创新融合多模态输入的移动设备应用是“像素色调（PixelTone）”[28]。“像素色调（PixelTone）”是一个多模态图片编辑界面，结合了语音和直接操作，让新手能简单地使用移动设备编辑图片。该应用能使用自然语言表达想要如何修改图片，也能直接操作定位指定位置的修改，如图 9.9 所示。“像素色调（PixelTone）”不仅仅为编辑图片提供了便利的界面。该界面允许模糊指令，新手可以使用例如“让它好看些”这种指令，也可以用更高级的指令，如“锐化顶部的中间色调”。虽然相比简单的触控界面，用户运用多模态界面进行的也是一样的操作，但他们总体更倾向于多模态界面，并且能够有效运用该应用完成实际工作量。

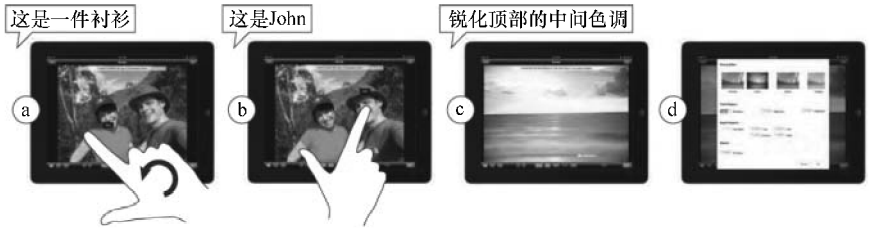


图 9.9　像素色调（PixelTone）。来源：Laput GP，Dontcheva M，Wilensky G，Chang W，Agarwala A，Linder J and Adar E 2013。授权转载

9.3.1.7　汽车控制

虽然美国大多数州禁止开车发短信，但司机通勤时仍要进行许多活动。正在进行的研究可帮助司机在进行更高级的活动，如在进行导航、通信、换音乐、控制环境等操作的同时，有效完成基本驾驶任务[29]。美国汽车工程师学会建议，在非驾驶状态需要用超过 15s 来完成的任务，在汽车行驶时应禁止执行。语音控制是 15s 规则的例外，因为它们不要求用户把视线从道路上移开，也许能够显著解决这一问题。

然而，有些数据显示某些语音界面会导致高识别负荷，可能对驾驶情况有负面影响。这一负面影响是由于语音识别有一定技术限制，还有一些可用性方面的问题，例如混淆的或不

一致的指令集合，还有多余的深奥复杂的对话结构。根据驾驶情况，通过结合最好的输入模态，多模态界面可能是应对这些问题的一种好方法。

语音、触控、手势和触控板都分别被用作驾驶界面的输入。然而，单凭某一项输入是无法完全解决问题的。Pfleging 等人（2012）[30]创造了一个多模态界面，使用语音结合手势操控方向盘，尽量防止司机分心，如图 9.10 所示。Pfleging 等人指出只用语音输入无法进行精确控制，而只用触控输入又要求较多的视觉交互，只用手势输入不能很好地缩放[31]。他们提出一种结合语音和手势的多模态交互方式，可用语音指令选取可视的对象或功能（镜子、窗户等），并且简单的触控手势可用于控制这些功能。有了这种方法，用户能看见他们需要说什么，就能较简单地想起语音指令。通过运用简单的触控手势，这种交互方式降低了对视觉交互的需求，同时也能即时反馈，取消操作也变得很简单。其他关注汽车控制多模态输入的内容可参见 Gruenstein 等人（2009）的研究[32]。

图 9.10　将语音和姿势相结合的多模态汽车方向盘
来源：Pfleging B，Kienast M，Schmidt A and Doring T 2011。经授权转载

9.3.2　3D 姿势

有一些设备，例如由微软公司出品的 Kinect 和由英特尔公司根据感知计算软件开发包出品的深度相机，比如 Creative 的 Senz3D 相机，已经得到稳定而广泛的普及，并用于以 3D 姿势为基础的新的交互科技中。结合使用深度相机和标准色彩模态相机时，这些设备可以提供精准的骨骼追踪和手势检测。相比较 WIMP 的人机交互界面，这样的 3D 姿势可以更自然地完成一些任务。为了丰富用户的体验，可以结合多种方式来使用姿势，例如语音和面部追踪。更加有趣的是，Kinect 和 Senz3D 相机已经运用于微软手机中，这使得微软手机可以在多模态界面上做得更好。

另外一种普遍使用的姿势检测技术是立体相机，立体相机采集姿势后通过机器识别和过滤后将姿势进行分类。在这些科学技术得到发展之前，在多模态界面中使用 3D 姿势仅限于

使用简单的相机来检测用于选择或类似任务的手势。语音也是一种普遍使用的与手势相结合的方式，这也体现了当下的3D姿势倾向于同时调动使用人体的多个部位，而不是像其他一些方式一样在同时使用姿势的方面受到很大局限[5]。

众多应用程序已经实现了对Kinect传感器的使用，其中的很多游戏都兼具了语音和姿势识别功能，这些功能的发展得益于微软公司的第一方开发商。除了游戏行业，类似Kinect这类的传感器还被用于各类仿真模拟中，这也使得科技交互可以通过更多比较自然的动作得以实现。在人机交互以及医疗领域中已经做了部分工作，以实现免提、姿势控制的应用。

9.3.2.1　游戏和仿真

通过肢体语言进行的仿真交互在虚拟情景中得到普遍使用，尤其是当语音功能也可以与姿势结合允许同时输入的情况下。Williamson等人（2011）根据这些开发出了一套可以为士兵进行全方面身体训练的系统[33]，这套系统结合运用了Kinect、语音控制以及索尼游戏平台上的动作控制器（见图9.11）。这套“真实边缘”系统的原型可以使用户通过前进行走、微微倾斜等动作实现在机器上的同步情景。用户还可以通过连接到类武器设备的移动控制器环顾四周的环境。除此之外，用户还可以通过语音对虚拟情景中的角色发出指令。

图9.11　结合了Kinect和PS移动控制器的“真实边缘”系统。来源：经Brian Williamson允许转载

目前可使用的基于深度摄像机的姿势识别装置存在一个缺点，即用户必须面对装置才能使其准确跟踪用户的体态。“真实边缘”融合系统是“真实边缘”原型的一个延伸，由于添加了多个环绕用户的Kinect，使得该系统可以提供360°无死角的姿势识别，同时由于在数据层面加入了融合检索骨骼的技术，也使得该系统可以对任意方向的用户姿势进行识别[34]。骨骼跟踪信息通过Kinect从用户客户端传入电脑服务器，并在电脑服务器进行数据融合。也就是说，相对原型而言，这套融合系统只需要更多的Kinect传感器、电脑客户端，以及安装好的可以为用户提供关于虚拟环境的正确数据的头盔。

已经有大量研究强调了关于从语音到3D面部识别的分割以及选择。Budhiraja等人明确提出，关于指示姿势存在的一个问题是大量密集或阻塞的对象会使选择变得困难[35]。为了解决这一问题，专家将语音作为一种添加模态以帮助指定所需对象的属性，例如对象存在的空间位置、相对位置或物理特性等。正是因为有了这些属性我们才能对对象进行特定的描述，比如“左边蓝色的那个”可以用于帮助人们选择需要的到底是哪个。如果要进行精确的定义，那么对象的物理属性和方位都必须清楚明了。

有许多实例证明，在人机交互过程中，3D 姿势并非是最主要的交互方式。SpeeG 输入系统是一种基于姿势的键盘替换系统，它结合了语音增强以及 3D 姿势交互技术[36]。这项系统是基于语音和姿势相结合的 Dasher 交互体系[37]，可以取代鼠标的功能。该系统使用调节语言的功能使得用户可以通过姿势使软件选择正确的指令。

图 9.12 向我们展示的是虚拟场景以及指令手势。尽管由于语音识别的延迟导致系统原型无法实现实时同步的信息输入，但用户在进行体验后觉得相对单一由微软 Xbox 360 控制器、语音控制器或微软 Kinect 键盘控制的屏幕键盘而言，SpeeG 仍是最有效的交互方式。

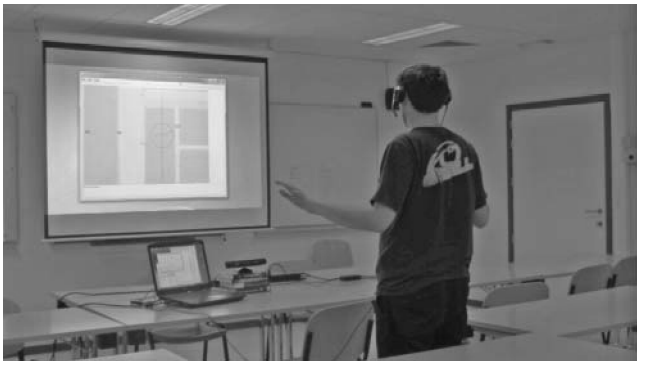


图 9.12　SpeeG 交互以及示例场景。来源：经 Lode Hoste 许可转载

3D 姿势不受身体整体移动的限制。Bohus 和 Horvitz（2009）[38]研发了一套用于检测对话中的头部姿势、面部表情和一定数量的自然语言的系统。通过一个基础的广角相机以及商用软件可以实现头部姿势的跟踪和凝视估算。一个线性麦克风用于采集用户的声源。这些方式通过融合和分析后会向用户做出一个适当的反馈。

这个多方系统被用于对话的观察性研究，在这一研究中，系统会提出问题等待用户进行回答。除了接收来自用户的回答，该系统还会口头询问用户是否确认答案。系统还将视用户的答题情况判断是否继续答题或切换问题等。这套系统的行为是基于一个轮换会话模型，也因如此，该系统具有四种行为模态，即保持、结束、接受、无效。图 9.13 给出了该系统的功能。

Hrúz 等人（2011）[39]为两位残疾人士的交流情景开发出了一套多模态的姿势和语音识别系统，两位交流者一位为失聪者，另一位为失明者。该系统利用事先训练好的识别器对其中一位用户的手势进行识别。识别器只使用了一个相机对手势进行捕捉，所以用户必须身着深色的衣服，与双手和背景形成明显对比，从而使识别器能够更加准确地检测到信号。这些信号被采集之后会被转化为文字形式，然后再由文字被转化为语音形式后传达给失明用户。另一位失聪用户则可以通过将语音信号转化为文字信号的方式进行交流。这个语境中的每一位用户都是该系统建构中的一个独立输入模态，如图 9.14 所示。这套系统同样适用于其他两个无法找到合适媒介进行交流的人。

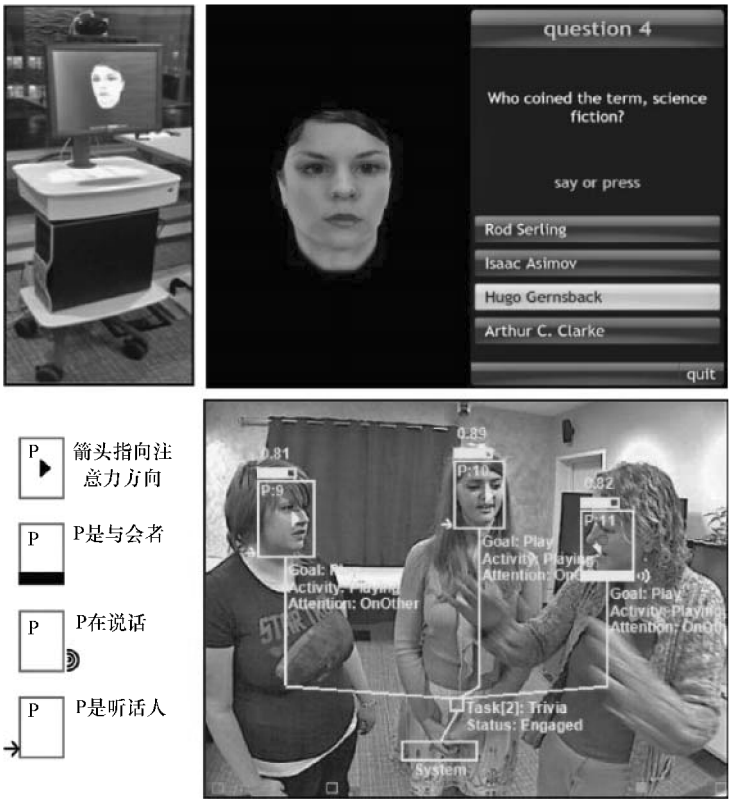


图 9.13　多方轮换会话模态的示例。来源：经微软公司许可转载

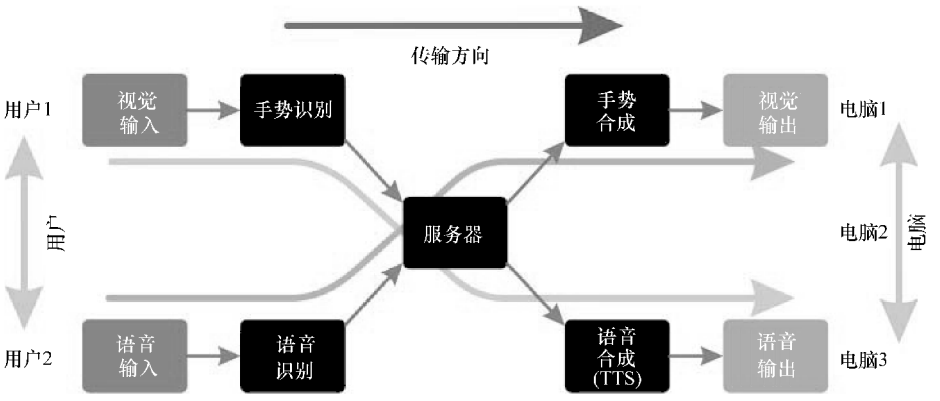


图 9.14　两位残障人士之间的交流原理示意图。来源：经 Marek Hruz 许可转载

9.3.2.2　医学应用

Gallo 等人（2011）[40]根据 Kinect 研发出了一套医学影像数据导航系统。尽管这是虚拟的系统，但用户仍可以在这些数据中自由切换，例如用户通过手势就可以实现缩放、翻译、旋转以及指向等功能。除此之外，用户还可以在此环境中选择并提取自己感兴趣的数据。这套系统支持一些常用的成像系统，包括人体横断面扫描成像、核磁共振以及断层扫描成像等。

在医学中采用 3D 姿势的最大好处在于它可以避免手与医疗器械的接触，使医疗环境保持无菌的状态。这类交互可以在手术环境中加以使用，以提供图像信息而无需重复消毒除菌。在计算机化的医疗系统中往往存在着有关无菌环境的问题。一般来说，手术室里需要一位助理来对这些图像信息进行处理，并管理其终端和病人的图像。这些助理通常不具备同级主刀医生的培训水准，而且可能会误解医生本来能够正确解读的信息。

9.3.2.3　人机交互

Perzanowski 等人（2001）[41]设计出了一种人机交互的方式，使用这种方式时人们的语言、姿势都显得更加自然。这种交互方式的实现是通过摆放一个立体摄像机来查看人们所做出的手势动作，并判定这些动作是否具有意义。通过接收来自用户的姿势和语音信息，机器人可以正确完成用户的指示。使用的语音指令包括“去那边”“走快一点”等简单的语句。

用户给出的姿势信息可能是从手指指向的某一个位置移动到一段距离之外的另一个位置，或是将握紧的双手打开。机器人可以精确地做出判断并决定移动的距离。用户还可通过掌上电脑等实现对机器人的远程控制。用户可以在掌上电脑中结合使用语音和姿势指令代替对机器人直接的语音和姿势指令。指派给机器人的任务可能会受到干扰而中断，但是机器人最终可以还原其原始任务并完成。类似的结合了 3D 姿势和语音识别控制机器人的交互技术在参考文献［42，43］中有详细介绍。

9.3.2.4　电子消费品

多模态输入技术的发展日益成熟，现已成为电子消费品，尤其是大屏显示器和电视机的界面功能之一。其中一个商业例子就是三星公司的智能电视机系列，其包含 3D 姿势、语音输入以及面部识别。在研究界，Lee 等人（2013）[44]将 3D 姿势和面部识别结合运用于智能电视机，其中 3D 姿势用于调频和控制音量，面部识别则用于用户身份验证。Takahashi 等人（2013）[45]还利用深度相机对带有脸部跟踪的 3D 姿势技术进行了研究，以辅助用户观看电视。

Krahnstoever 等人（2002）[46]研发了一个类似的系统，将 3D 手势、头部追踪、语音输入以及面部识别结合并用于一个大画幅显示器。这个系统可配置于商场内，帮助顾客选择合适的目标商店。由于研发技术日益进步，产品体积变小以及价格有所降低，这些多模态输入界面将作为主要的用户界面被应用于更多的电子消费品，如台式电脑、笔记本电脑等。

9.3.3　眼动跟踪和凝视

从电子游戏到广告分析，在各种不同的应用软件里，商业性眼动跟踪和凝视定位设备都

层出不穷[47]。因此在过去几年，提高用户凝视位置识别的能力成为了研究领域中一个很重要的方面。在多模态输入系统中，凝视主要用于对象的选择：将凝视与键盘结合进行选择[48,49]或者将凝视作为初步选择，再结合鼠标进行更精确的识别[50,51]。

多模态界面将凝视与手势一体化，适用于大屏设备的多显示器环境[52]。在凝视跟踪的基础上使用鼠标滚轮、倾斜手持设备或者利用触摸输入等各种不同的功能可实现大型信息空间的平移和缩放操作[53]。将凝视跟踪和语音输入结合则可用于文本输入[54]。尽管可用性测试表明，传统的键盘输入速率更快，信息更准确，但在这样的多模态界面中，用户只需聚焦于一个感兴趣的特征，发出语音命令，就可将该关键字输入目标文件。带有眼动跟踪功能的多模态界面还被用于娱乐领域。例如，Heo 等人（2010）[55]研发了一款结合凝视、手势以及生物反馈的游戏（见图 9.15），表明了多模态界面比传统的键盘和鼠标控制更具吸引力。

图 9.15　一款结合凝视、手势以及生物反馈的游戏。由 Hwan Heo 许可转载

另外，最近眼动跟踪还被用于脑机接口（BCI），以方便残障人士使用[56, 57]。在这样的界面里，脑机接口部分模仿选择特定对象的目光停留时间，凝视则用于指向这一特定对象，两者协作即可便于残障人士进行选择。考虑到单独使用脑机接口时人体动作感知的有限性，这种将凝视与脑机接口技术一体化的多模态界面能令用户产生更直观的感觉（见 9.3.5 节）。

9.3.4　面部表情

面部表情识别可看作感知计算应用的重要组成部分，且是一个具有挑战性的热点研究问题。而且目前在计算机视觉领域已进行了大量有关面部表情识别的研究[58]。在多模态交互中，面部表情主要用于两个方面。

第一种方式，面部表情与人体其他特征相结合，增强识别的准确性，最终达到人类情绪识别。例如，De Silva 等人（1997）[59]将视觉信息与听觉信息结合，以确定哪种信息能更好地识别某些情绪。结果表明，视觉信息能更准确地识别人类的喜、怒、惊和恶，而听觉信息则更易识别哀和惧。

而 Busso 等人（2004）[60]在情绪识别探测系统中结合语音和面部表情也发现了类似的结果。Kessous 等人（2010）[61]在多模态识别器中利用面部表情、语音识别以及肢体语言来探测人类情绪。另一个将面部表情和语音输入结合的情绪探测系统的例子见参考文

献［62］。

面部表情用于多模态界面情景的另一种方式就是建立情感计算系统，以确定情绪或者心情状态，进一步调整应用软件的界面、难度以及其他参数，以增强用户体验效果。例如，Lisetti 和 Nasoz（2002）[63]开发了一个多模态用户情感界面——MAUI 系统。该系统通过结合面部表情、语音以及生物反馈来探测用户的情绪状态。

又如，Caridakis 等人（2010）[64]利用递归神经网络发出的视觉与听觉信息研发出了情感状态识别器。其识别率高达 98%，已达到多模态感知计算系统的标准：它可以观察并理解用户的情感状态，不论用户正在主动发送命令还是被动接受监控。

9.3.5　脑机接口

现代脑机接口只需利用脑电图便能监测人类的心理状态。为了利用现代技术追踪信号，必须将多个电极连接至人脑特定部位。然而这些连接限制了人体某些特征进入可交互的系统。如果用户头戴脑机接口器的同时身体稍有移动，传输信号中就会有噪声，从而降低了信号准确度。不过由于脑机接口通常用于残疾人士的交流和运动，信号噪声便不是大问题。

包括 Emotiv（见图 9.16）和 Neurosky 在内的多家公司已开始研发低成本脑机接口，用于诸如电子游戏等以前非常规应用软件。由于脑机接口成本降低，数款多模态应用软件已被提议使用脑机接口。目前，使用最广泛的脑机接口形态是语音和凝视，因为两者的应用不需要身体的移动。Gürkök 和 Nijhol（2012）[65]特列举了多项例子，表明脑机接口可通过将人脑控制的界面作为多模态界面的一种形态来增强用户体验和工作效率。

图 9.16　Emotiv 脑电图神经头盔。来源：Corey Pittman

脑电图通常与额外的神经图像，比如用来测量肌肉活动的肌电图进行结合。与单独使用脑电图或肌电图相比，结合使用使 Leeb 等人（2010）[66]在识别性能效果方面取得了显著提高。两种信号的贝叶斯融合则可产生混合信号，脑电图与近红外光谱结合也被证明能有效提高信号的分类精度[67]。不过由于自身明显的延迟性，近红外光谱对实时脑机接口造成了阻碍。

Gürkök 等人（2011）[68]研究了在用户自创的电子游戏中各种不同输入模态对用户表现

的影响。在一款名叫“照看绵羊!”的游戏中，用户需要移动一群小狗，让小狗将羊群赶入围栏（游戏系统设置见图 9.17）。该游戏由鼠标和一两种其他形式相结合使用。游戏者通过语音或者脑机接口来选择小狗。若用语音，只需说出待选小狗的名字；若用脑机接口，则需要专注于待选小狗所在的位置，然后在小狗被指定的目的地位置释放鼠标按钮。游戏者被要求在三种情况下进行游戏：自动语音识别，脑机接口以及在多模态配置中结合自动语音识别和脑机接口。研究发现，与只能使用一种特定游戏形式相比，有机会选择游戏模态并没有显著提高游戏者表现能力，因为部分游戏者整个过程一次都没有改变过游戏模态。

图 9.17　“照看绵羊!”游戏界面。
来源：Hayrettin Gürkök 许可转载

脑机接口的一个扩展应用是建模接口。Sree 等人（2013）[69]设计了一个软件框架，将脑机接口作为 3D 建模的额外辅助模态。在这个软件框架里，Emotiv 脑电图神经头盔是主要的应用装置，再次结合脑电图与肌电图，并且连至键盘和鼠标，共同控制建模过程。这个带有 Emotiv 的软件将为装置信号设置参数，并根据特定用户的需求对装置进行调整。软件的肌电图模块用于探测脸部动作，包括向左看、控制画弧、对鼠标左键眨眼等。软件的脑电图模块则用于控制鼠标动作，以探测用户的行为意图。

Emotiv 应用程序可用于解释 12 种动作，包括 6 种定向动作和 6 种转体动作，且都可用于计算机辅助设计环境。然而该程序系统有参与者疲劳这一普遍问题，还有一些有关脑电图信号强度的问题。因此，系统可添加如语音在内的其他输入形式以提高可用性。

Zander 等人（2010b）[70]让用户自由使用想象动作或视觉焦点或两者的结合来控制脑机接口。他们认为，如果可供选择的控制方式或者混合控制技术可明显提高准确性，那么脑机接口在只有一种控制方式的情况下就不适用于所有用户。Maye 等人（2011）[71]在利用脑机接口增加用户可控制的外界刺激（不同的触觉和视觉刺激）时，保持相似的脑力活动，从而提出了一个可用的方法。用户在不同外界刺激中进行转换就可更加容易地对大脑活动进行分类。而 Zander 等人（2010b）[70]则将人机交互中的脑机接口分为三类：主动活动、反应活动以及被动活动。

9.4　多模态集成策略

多模态界面中最重要的模块之一就是集成部分。集成通常被称为融合引擎，它将不同的输入模态结合，创造出有意义指令的连贯界面[72]。但在建立多模态集成引擎时会有很多外

来的技术性挑战：

- 第一，不同的输入模态在数据格式、输入频率、语义意义等方面有不同的特点，从而难以进行结合。
- 第二，一个交互序列的不同时间需使用不同的输入形式，要求集成引擎反应灵活以推进输入过程，因此定时很重要。
- 第三，是和定时有关的重大挑战：消除模糊。当集成引擎约束力不足（引擎信息不足，无法做出融合决定）或者约束力过大（引擎产生信息冲突，需做出几项融合决定）时，就会导致融合模糊。
- 最后，多模态界面所使用的输入模态通常来源于自然交流渠道（例如，3D 姿势、语音、面部表情等），在这些渠道中，需运用识别技术对接收的数据进行切割与分类。因此，所有这些输入模态都存在概率的不确定性，使得集成引擎运行更加复杂。

在融合引擎中执行多模态集成有两个基本方法：前期集成与后期集成。两种方法都有各自不同的集成方式[73]。前期集成的前提是数据要早于任何主要处理过程（低阶处理除外）而首先被集成。与之对比，后期集成分别通过每个模态进行数据的处理，并在集成开始前将数据单模态化。后期集成的优势在于，因各种输入模态可以被单独分析，那么就不存在时间同步的问题，软件开发也更加简单。

然而，后期集成自身有一个问题，它可能丢失潜在的跨模态交互作用信息。例如，前期集成中来自姿势识别器的结果和语音识别器的结果可互相补充与纠正。而在后期集成中，每一个识别器只能独立做出形式运用的决策。目前，对前期集成或后期集成的选择问题仍是研究热点，这取决于所使用的输入形式以及应用软件所支持的多模态交互形式。需要注意的是，在某些情况下，可折中使用两种方式来执行多模态集成。例如，将前期集成中的 3D 姿势和凝视与后期集成的语音结合。在前期集成和后期集成的背景下，对于任何的接收数据流都有 3 个不同的集成级别：数据级、特征级和决策级[74]。数据级和特征级适用于前期集成，其中数据级集成主要关注低阶处理，通常用于相似的输入模态，如嘴唇和面部表情。这种处理方式还被用于最小集成。因为最接近原始数据源，数据级集成便可以提供最详细的信息，但它的运行易受噪声的影响。

特征级集成用于各种模态紧密结合或者同步运行时。示例形式包括来自声音和嘴唇动作的语音识别，示例策略包括神经网络和隐马尔可夫模型。与低级集成相比，特征级集成不易受噪声影响，但无法提供大量细节信息。

决策级集成（例如，对话水平融合[72]）属于后期集成，是多模态集成最普遍的形式。其主要优势在于处理松散结合的模态（例如，触控输入和语音输入）的能力，但还要取决于各输入模态独立完成信息处理的准确性。

框架式、合并式、程序性和符号/统计集成是在决策级集成下的最普遍的集成策略。

9.4.1　框架式集成

框架式集成着重于属性 – 值对的数据结构。这种框架收集各种输入模态的值对，并做出

全局性解释。以语音输入为例，一个属性－值对可能是“操作”，其含义可能是“删除”“添加”和“修改”等。每个框架支持一个独立的输入形式，集成则作为框架含义群的集合。每一个属性都有分值，集成属性的总分就代表最好的行动方案。Koons 等人（1993）[75]是第一个研究结合了3D 姿势、凝视以及语音的特征型集成的团体之一。最近，Dumas 等人（2008）[76]研发了 HephaisTK 多模态界面工具包，将框架式融入多模态集成。其他通过不同输入形式而使用框架式集成的多模态界面见参考文献［77－79］。

9.4.2　合并式集成

合并式集成的主要理念是使用合并操作符。该理念源于自然语言处理[80]，控制两个部分信息的一致性，若信息一致，就可组合为一条信息[81]。例如，Cohen 等人（1997b）[82]首先在 QuickSet 系统中结合了一致性和类型性特征结构将笔式手势和语音输入进行集成。最近，Taylor 等人（2012）[83]选择了一个合并式集成方案，将语音和 3D 指向手势与带触摸手势的语音连接，支持与无人控制机器人车辆的交互。Sun 等人（2006）[84]也使用合并式集成，并与多模态语法句法结合，该语法句法存在于运用 3D 姿势和语音的交通管理工具中。合并式集成在一次性融合两种输入模态的情况下运行状况更佳，并且绝大多数合并式集成研究都更倾向于输入对。更多合并式多模态集成见参考文献［85，86］。

9.4.3　程序性集成

程序性集成技术通过算法管理明确表示了多模态状态空间[72]。程序性集成的常见例子有扩展转移网络和有限状态机。例如，Neal 等人（1989）[87]和 Latoschik（2002）[88]在程序性集成中都运用了扩展转移网络，Johnston 和 Bangalore（2005）[89]以及 Bourguet（2002）[90]则对程序性几何运用了有限状态自动机。其他使用程序性集成的方式还有 Petri 网[91]和引导传播网络[7]。在这些系统里，语音输入可与鼠标、键盘、笔式输入、触控输入或者 3D 姿势结合。

9.4.4　符号/统计集成

符号/统计集成使用更多传统合并式方法，并将这些方法与统计处理结合，形成混合多模态集成策略。这些策略也从机器学习中引进相关概念[92]。尽管主要和特征级集成共同被应用，机器学习在决策级集成方面也有被研究的事例[4]。以 Pan 等人（1999）[93]为例，他们利用贝叶斯推理得出了一个公式，以估算多感信号的联合概率。其中多感信号利用合适的映射函数以反映信号之间的关联，映射则由最大互信息量引导。

关于符号/统计集成技术的一个更早的例子是 QuickSet 应用中的 MTC（小组委员会）架构[94]。在 MTC 中，各种输入形式根据后验概率而被集成。各种模态的识别器作为 MTC 统计积分器的成员，组成不同的团队，经训练而互相协作并衡量不同模态的输出。当前输入一经接收，团队就建立后验估算机制，发出多模态指令。

MTC 积分器将后验概率的经验分布进行分析，然后将每个待选指令标记为最高级别指

令。Flippo 等人（2003）[95]利用了与 MTC 相似的方法作为多模态交互框架的一部分，他们使用并行代理来估算每一个模态识别结果的后验概率，然后衡量之，综合决策出可适用的多模态指令。

最近，Dumas 等人（2012）[96]开发了一个统计型多模态集成方案，该方案通过时间关系属性，使用隐马尔可夫模型进行语义级相关输入形式的探测。更多有关多模态集成中机器学习的信息见参考文献［97，98］。虽然这些方法在多模态集成中能有效应对建模的不确定性，但它们有一个主要的缺陷：需要大量训练数据的支持。

9.5　多模态交互的可用性问题

考虑到感知计算情境中多模态交互的本质特点以及为提供直观有力的用户体验而紧密结合不同输入形式的目的[99]，多模态交互的可用性就成为了多模态界面设计中至关重要的一部分。为方便讨论多模态输入的部分可用性问题，我们以 Oviatt 关于多模态交互的十大迷思作为讨论的开端[100]。尽管这些迷思著于几年前，但至今仍可适用。

如果建立一个多模态系统，那么系统用户就能进行多模态交互。然而一项应用支持多模态输入并不表明用户会利用它们发出所有指令。因此，指令结构的灵活性对于人机之间的自然交流形式非常重要。换言之，多模态界面应该具有灵活性，能以不同方式发出指令。例如，系统用户应能同时使用语音和 3D 姿势发出指令，并且还可选择同时使用语音和凝视或者 3D 姿势和凝视或者单独使用语音。就输入模态集成的方式而言，这种选择设计会使整体多模态用户界面更复杂，但具有最广泛的概括性。

语音和指向是主要的多模态集成形式。从可用性角度看，语音和指向有利于直观的多模态输入组合，尤其当用户要选择虚拟对象并对这些对象执行操作（例如，将［这个］圆筒漆蓝）时。但本章所讨论过的内容中，还有多种可用的多模态输入组合。而从可用性角度看，存在一个关键问题：特定的输入组合真能适用某一特定任务吗？

总的来看，为给定任务提供支持简单自然的交互隐喻的多模态输入组合非常重要。例如，在应用触控输入或者 3D 姿势的移动设备中，语音和指向就可能不是最佳输入组合。

多模态输入包含同时信号。并非所有多模态输入策略都要求用户同时执行各种输入形式，特定输入形式需要时间整合，而许多情况下，各种形式以互补输入模态交替进行（例如，先说话，然后执行 3D 姿势，反之亦然）。实际上，多模态输入策略还可以对一些特定任务使用一种模态，而对其他任务使用另外的模态。因此，从可用性角度看，重要的是输入形式可以多种不同方式结合，且并不是所有输入形式都需支持同时输入。

在任何带有语音的多模态系统中，语音就是主要的输入模态。虽然语音输入是人类用以交流的主要输入模态，但它在多数情况下并不是多模态界面的主要模态。可惜的是，语音识别的效果在喧闹环境中会减弱，对语音输入也就更不利。另外，用户可能因顾及隐私而不愿使用语音输入。对于其他情况，当其他输入形式的结合更有利于执行给定任务，语音就可能仅是一种备用输入模态。因此，设计多模态交互情境时，并不需要将语音设置为主要的输入

模态，而只在最合理的情况下使用。

多模态语言和单模态语言在语言学上并无差异。多模态交互的优点之一在于简化了输入模态。设想一位用户想要移动对象位置的情境便知。而单独使用语音则要求用户不仅要描述关键对象，还要描述出对象将被放置的地点。然而，将语音和姿势结合，用户就可简化描述过程，因为他们同时也在使用第二种输入形式（在这种情况下是指向），既可识别对象，也可将对象放置于不同地点。这种输入组合表明，当用户在执行并发多模态交互时，可使用简单的输入指令来控制单个输入模态。从可用性来看，重点是了解单模态语言有时可能比多模态语言更加复杂，而且多模态输入可消除这种复杂性，有利于界面的简化使用。

多模态集成包含各种输入模态间的信息重复。多模态集成中一个关键理念是重复的输入模态有助于增强用户体验，原因是输入模态可强化彼此。从计算角度来看这无疑是正确的假设，且在多模态集成中占有一席之地。但是，从可用性角度看，多模态输入的补充性质不应因为它的优点而被忽视。因此，确保合适的多模态集成以达到补充的效果在用户看来很重要。

单个错误识别技术经结合成多模态技术，可能导致更多的错误。多模态输入，尤其是感知计算的一项有意思的挑战在于，使用的各输入模态需要识别技术以理解输入进程。不足的是，由于识别器的精准度不确定，识别结果也会出错。然而，结合多种识别性输入确有助于提高指令的整体精准度，创造出更可靠的使用界面。而精准度提高的关键在于多模态集成策略。另外，如果可以自由选择，用户就会使用他们认为精准度更高的输入形式。所以，从可用性角度看，这一使用模式也可说明确保多模态界面灵活性的原因。

所有用户发出的多模态指令都以相同方式集成。多模态界面用户会识别集成模态，以确定早期将如何使用界面，并保持这种使用方式。然而，正如我们所看到的，人类有很多种不同的方式来使用多模态界面。因此，多模态集成方案要灵活，能识别基于用户的主要集成模态。由于融合引擎可感知用户如何与不同输入模态交互，这一方案可以提高识别率。

不同输入模态可用于传输相似的信息，但是并非所有的输入模态都是平等的。换句话说，根据用户想要传达信息类型的不同，针对这些类型，输入模态也各有强势和弱势。例如，凝视能产生与语音几乎完全不同类型的信息。所以，从可用性角度看，重点在于了解哪些输入模态可用，且可用于哪些情况。也就是说，如果一个输入模态用于执行不相符的任务，将只会使得界面操作更加复杂。

高效是多模态系统的一个主要优势。不过，速度和效率并不是多模态界面的仅有优势。多模态交互的其他重要优势还包括能降低单个识别系统的错误率以及能提高按照用户意愿与应用软件进行交互时的灵活性。

9.6　结语

在本章中，我们已经探索了如何组合不同的输入模态可以形成自然和表现力强的多模态界面。我们已经研究了多模态输入策略，并提出了各种能够提供触摸输入、语音、3D 姿势、

眼睛凝视与跟踪、面部表情和脑机接口的不同组合的多模态界面。我们还研究了多模态整合或融合，这是能集成不同模态的多模态结构的重要组成部分，通过检测不同的方法和集成水平形成一个有凝聚力的界面。最后，我们已经提出了一些可用性问题，因为它们与多模态输入相关。显然，多模态界面距离 Bolt 的“放在那里”系统[1]还有很长的路要走。

然而，各种领域还需要更多的努力，包括多模态集成、识别技术和可用性，以全面支持感知计算应用，从而提供强大、高效、表现力强的人机交互。

参 考 文 献

1. Bolt, R.A. (1980). *"Put-that-there": voice and gesture at the graphics interface*. Proceedings of the 7th annual conference on computer graphics and interactive techniques, SIGGRAPH '80, 262–270. ACM, New York, NY, USA.
2. Oviatt, S. (2003). Advances in robust multimodal interface design. *IEEE Computer Graphics and Applications* **23**(5), 62–68.
3. Turk, M., Robertson, G. (2000). Perceptual user interfaces (introduction). *Communications of the ACM* **43**(3), 32–34.
4. Dumas, B., Lalanne, D., Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In: Lalanne, D., Kohlas, J. (ed.). *Human Machine Interaction,* 3–26. vol. 5440 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg.
5. Jaimes, A., Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding* **108**(12), 116–134. Special Issue on Vision for Human-Computer Interaction.
6. Oviatt, S. (2007). Multimodal interfaces. In: Sears, A., Jack, J. (eds.). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Second Edition, 413–432. CRC Press.
7. Martin, J.C. (1998). Tycoon: Theoretical framework and software tools for multimodal interfaces In: Lee, J, (Ed.). *Intelligence and Multimodality in Multimedia Interfaces*. AAAI Press.
8. Oviatt, S., Van gent, R. (1996). *Error resolution during multimodal human-computer interaction*, 204–207.
9. Bowman, D.A., Riff, E., Lavolta, J.J., Papyri, I. (2004). *3D User Interfaces: Theory and Practice*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.
10. Sharma, A., Madhvanath, S., Shekhawat, A., Billinghurst, M. (2011). *Mozart: a multimodal interface for conceptual 3D modelling*. Proceedings of the 13th international conference on multimodal interfaces, ICMI '11. ACM, New York, NY, USA, 307–310.
11. Radhakrishnan, S., Lin, Y., Zeid, I., Kamarthi, S. (2013). Finger-based multitouch interface for performing 3D CAD operations. *International Journal of Human-Computer Studies* **71**(3), 261–275.
12. Tse, E., Greenberg, S., Shen, C., Forlines, C., Kodama, R. (2008). *Exploring true multi-user multimodal interaction over a digital table*. Proceedings of the 7th ACM conference on Designing interactive systems, DIS '08. ACM, New York, NY, USA, 109–118.
13. Tse, E., Greenberg, S., Shen, C. (2006). *GSI demo: multiuser gesture/speech interaction over digital tables by wrapping single user applications*. Proceedings of the 8th international conference on Multimodal interfaces, 76–83.
14. Tse, E., Greenberg, S., Shen, C., Forlines, C. (2007). Multimodal multiplayer tabletop gaming. *Computers in Entertainment (CIE)* **5**(2), 12.
15. Martínez, R., Collins, A., Kay, J., Yacef, K. (2011). *Who did what? Who said that? Collaid: an environment for capturing traces of collaborative learning at the tabletop*. Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, 172–181.
16. Barthelmess, P., Kaiser, E., Huang, X., Demirdjian, D. (2005). *Distributed pointing for multimodal collaboration over sketched diagrams*. Proceedings of the 7th international conference on Multimodal interfaces, ICMI '05. ACM, New York, NY, USA, 10–17
17. Piper, A.M., Hollan, J.D. (2008). *Supporting medical conversations between deaf and hearing individuals with tabletop displays*. Proceedings of the 2008 ACM conference on Computer supported cooperative work, CSCW '08. ACM, New York, NY, USA, 147–156.

18. Piper, A.M. (2010). *Supporting medical communication with a multimodal surface computer*. CHI '10 Extended Abstracts on Human Factors in Computing Systems, CHI EA '10. ACM, New York, NY, USA, 2899–2902.
19. Paek, T., Thiesson, B., Ju, Y.C., Lee, B. (2008). *Search VOX: Leveraging multimodal refinement and partial knowledge for mobile voice search*. Proceedings of the 21st annual ACM symposium on User interface software and technology, 141–150.
20. Ehlen, P., Johnston, M. (2011). *Multimodal local search in speak4it*. Proceedings of the 16th international conference on Intelligent user interfaces, 435–436 ACM.
21. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J. (1997a). *Quickset: Multimodal interaction for distributed application*. Proceedings of the fifth ACM international conference on Multimedia, 31–40.
22. Ramsay, A., McGee-Lennon, M., Wilson, G.A., Gray, S.J., Gray, P., De Turenne, F. (2010). Tilt and go: exploring multimodal mobile maps in the field. *Journal on Multimodal User Interfaces* **3**(3), 167–177.
23. Feng, J., Johnston, M., Bangalore, S. (2011). Speech and multimodal interaction in mobile search. *IEEE Signal Processing Magazine* **28**(4), 40–49.
24. Sim, K.C. (2012). *Speak-as-you-swipe (says): a multimodal interface combining speech and gesture keyboard synchronously for continuous mobile text entry*. Proceedings of the 14th ACM international conference on Multimodal interaction, 555–560.
25. Kristensson, P.O., Vertanen, K. (2011). *Asynchronous multimodal text entry using speech and gesture keyboards*. Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011), 581–584.
26. Shinoda, K., Watanabe, Y., Iwata, K., Liang, Y., Nakagawa, R., Furui, S. (2011). Semi-synchronous speech and pen input for mobile user interfaces. *Speech Communication* **53**(3), 283–291.
27. Dearman, D., Karlson, A., Meyers, B., Bederson, B. (2010). *Multi-modal text entry and selection on a mobile device*. Proceedings of Graphics Interface 2010, 19–26 Canadian Information Processing Society.
28. Laput, G.P., Dontcheva, M., Wilensky, G., Chang, W., Agarwala, A., Linder, J., Adar, E. (2013). *Pixeltone: a multimodal interface for image editing*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13. ACM, New York, NY, USA, 2185–2194.
29. Muller, C., Weinberg, G. (2011). Multimodal input in the car, today and tomorrow. *IEEE Multimedia* **18**(1), 98–103.
30. Pfleging, B., Schneegass, S., Schmidt, A. (2012). *Multimodal interaction in the car: combining speech and gestures on the steering wheel*. Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '12. ACM, New York, NY, USA, 155–162.
31. Pfleging, B., Kienast, M., Schmidt, A., Döring, T. (2011). *Speet: A multimodal interaction style combining speech and touch interaction in automotive environments*. Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI, 65–66.
32. Gruenstein, A., Orszulak, J., Liu, S., Roberts, S., Zabel, J., Reimer, B., Mehler, B., Seneff, S., Glass, J., Coughlin, J. (2009). *City browser: developing a conversational automotive HMI*. CHI '09 Extended Abstracts on Human Factors in Computing Systems, CHI EA '09. ACM, New York, NY, USA, 4291–4296.
33. Williamson, B.M., Wingrave, C., LaViola, J.J., Roberts, T., Garrity, P. (2011). *Natural full body interaction for navigation in dismounted soldier training*. The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), NTSA.
34. Williamson, B.M., LaViola, J.J., Roberts, T., Garrity, P. (2012). *Multi-kinect tracking for dismounted soldier training*. The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), NTSA.
35. Budhiraja, P., Madhvanath, S. (2012). *The blue one to the left: enabling expressive user interaction in a multimodal interface for object selection in virtual 3D environments*. Proceedings of the 14th ACM international conference on Multimodal interaction, 57–58.
36. Hoste, L., Dumas, B., Signer, B. (2012). *Speeg: a multimodal speech-and gesture-based text input solution*. Proceedings of the International Working Conference on Advanced Visual Interfaces, 156–163.
37. Ward, D.J., Blackwell, A.F., MacKay, D.J. (2000). *Dashera data entry interface using continuous gestures and language models*. Proceedings of the 13th annual ACM symposium on User interface software and technology, 129–137.
38. Bohus, D., Horvitz, E. (2009). Dialog in the open world: platform and applications Proceedings of the 2009 international conference on Multimodal interfaces, 31–38, ACM.
39. Hrúz, M., Campr, P., Dikici, E., Kındıroğlu, A.A., Krňoul, Z., Ronzhin, A., Sak, H., Schorno, D., Yalçın, H., Akarun, L. *et al.* (2011). Automatic fingersign-to-speech translation system. *Journal on Multimodal User Interfaces* **4**(2), 61–79.

40. Gallo, L., Placitelli, A.P., Ciampi, M. (2011). *Controller-free exploration of medical image data: Experiencing the Kinect*. IEEE International Symposium on Computer-Based Medical Systems (CBMS), 1–6.
41. Perzanowski, D., Schultz, A.C., Adams, W., Marsh, E., Bugajska, M. (2001). Building a multimodal human-robot interface. *IEEE Intelligent Systems* **16**(1), 16–21.
42. Burger, B., Ferrané, I., Lerasle, F., Infantes, G. (2012). Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots* **32**(2), 129–147.
43. Stiefelhagen, R., Fugen, C., Gieselmann, R., Holzapfel, H., Nickel, K., Waibel, A. (2004). *Natural human-robot interaction using speech, head pose and gestures*. Proceedings of International Conference on Intelligent Robots and Systems, IROS, **3**, 2422–2427.
44. Lee, S.H., Sohn, M.K., Kim, D.J., Kim, B., Kim, H. (2013). *Smart TV interaction system using face and hand gesture recognition*. IEEE International Conference on Consumer Electronics (ICCE), 173–174.
45. Takahashi, M., Fujii, M., Naemura, M., Satoh, S. (2013). Human gesture recognition system for TV viewing using time-of-flight camera. *Multimedia Tools and Applications* **62**(3), 761–783.
46. Krahnstoever, N., Kettebekov, S., Yeasin, M., Sharma, R. (2002). *A real-time framework for natural multimodal interaction with large screen displays*. Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, 349.
47. Duchowski, A.T. (2007). *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
48. Kumar, M., Paepcke, A., Winograd, T. (2007). *Eyepoint: practical pointing and selection using gaze and keyboard*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07. ACM, New York, NY, USA, 421–430
49. Zhang, X., MacKenzie, I. (2007). Evaluating eye tracking with ISO 9241– part 9. In: Jacko, J. (ed.). *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*. vol. 4552 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg, 779–788.
50. Hild, J., Muller, E., Klaus, E., Peinsipp-Byma, E., Beyerer, J. (2013). *Evaluating multi-modal eye gaze interaction for moving object selection*. The Sixth International Conference on Advances in Computer-Human Interactions, ACHI, 454–459.
51. Zhai, S., Morimoto, C., Ihde, S. (1999). *Manual and gaze input cascaded (magic) pointing*. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, CHI '99. ACM, New York, NY, USA, 246–253.
52. Cha, T., Maier, S. (2012). *Eye gaze assisted human-computer interaction in a hand gesture controlled multi-display environment*. Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Gaze-In '12. ACM, New York, NY, USA, 13, 1–13:3.
53. Stellmach, S., Dachselt, R. (2012). *Investigating gaze-supported multimodal pan and zoom*. Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12. ACM, New York, NY, USA, 357–360
54. Beelders, T.R., Blignaut, P.J. (2012). *Measuring the performance of gaze and speech for text input*. Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '12. ACM, New York, NY, USA, 337–340.
55. Heo, H., Lee, E.C., Park, K.R., Kim, C.J., Whang, M. (2010). A realistic game system using multi-modal user interfaces. *IEEE Transactions on Consumer Electronics* **56**(3), 1364–1372.
56. Vilimek, R., Zander, T. (2009). Bc(eye): Combining eye-gaze input with brain-computer interaction. In: Stephanidis, C. (ed.). *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*. Vol. 5615 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg, 593–602.
57. Zander, T.O., Gaertner, M., Kothe, C., Vilimek, R. (2010a). Combining eye gaze input with a brain-computer interface for touchless human-computer interaction. *Intl. Journal of Human-Computer Interaction* **27**(1), 38–51.
58. Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing* **30**(10), 683–697.
59. De Silva, L., Miyasato, T., Nakatsu, R. (1997). *Facial emotion recognition using multi-modal information*. Proceedings of 1997 International Conference on Information, Communications and Signal Processing (ICICS), **1**, 397–401.
60. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S. (2004). *Analysis of emotion recognition using facial expressions, speech and multimodal information*. Proceedings of the 6th international conference on Multimodal interfaces, ICMI '04. ACM, New York, NY, USA, 205–211.
61. Kessous, L., Castellano, G., Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces* **3**(1–2), 33–48.

62. Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., Narayanan, S.S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. Proceedings of Interspeech, Japan, 2362–2365.
63. Lisetti, C.L., Nasoz, F. (2002). *Maui: a multimodal affective user interface*. Proceedings of the tenth ACM international conference on Multimedia, MULTIMEDIA '02. ACM, New York, NY, USA, 161–170.
64. Caridakis, G., Karpouzis, K., Wallace, M., Kessous, L., Amir, N. (2010). Multimodal users affective state analysis in naturalistic interaction. *Journal on Multimodal User Interfaces* **3**(1–2), 49–66.
65. Gürkök, H., Nijholt, A. (2012). Brain-computer interfaces for multimodal interaction: a survey and principles. *International Journal of Human-Computer Interaction* **28**(5), 292–307.
66. Leeb, R., Sagha, H., Chavarriaga, R., del R Millan, J. (2010). *Multimodal fusion of muscle and brain signals for a hybrid-BCI*. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 4343–4346.
67. Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.R., Blankertz, B. (2012). Enhanced performance by a hybrid NIRS-EEG brain computer interface. *Neuroimage* **59**(1), 519–529.
68. Gürkök, H., Hakvoort, G., Poel, M. (2011). *Modality switching and performance in a thought and speech controlled computer game*. Proceedings of the 13th international conference on multimodal interfaces, ICMI '11. ACM, New York, NY, USA, 41–48.
69. Sree, S., Verma, A., Rai, R. (2013). *Creating by imaging: Use of natural and intuitive BCI in 3D CAD modelling*. ASME International Design Engineering Technical Conference ASME/DETC/CIE ASME.
70. Zander, T.O., Kothe, C., Jatzev, S., Gaertner, M. (2010b). Enhancing human-computer interaction with input from active and passive brain-computer interfaces. *Brain-Computer Interfaces*, Springer, 181–199.
71. Maye, A., Zhang, D., Wang, Y., Gao, S., Engel, A.K. (2011). Multimodal brain-computer interfaces. *Tsinghua Science & Technology* **16**(2), 133–139.
72. Lalanne, D., Nigay, L., Palanque, P., Robinson, P., Vanderdonckt, J., Ladry, J.F. (2009). *Fusion engines for multimodal input: a survey*. Proceedings of the 2009 international conference on Multimodal interfaces, ICMIMLMI '09. ACM, New York, NY, USA, 153–160.
73. Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters* **36**, 189–195.
74. Sharma, R., Pavlovic, V., Huang, T. (1998). Toward multimodal human-computer interface. *Proceedings of the IEEE* **86**(5), 853–869.
75. Koons, D.B., Sparrell, C.J., Thorisson, K.R. (1993). Integrating simultaneous input from speech, gaze, and hand gestures. In: Maybury, M.T. (Ed.). *Intelligent multimedia interfaces*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 257–276.
76. Dumas, B., Lalanne, D., Guinard, D., Koenig, R., Ingold, R. (2008). *Strengths and weaknesses of software architectures for the rapid creation of tangible and multimodal interfaces*. Proceedings of the 2nd international conference on Tangible and embedded interaction, TEI '08. ACM, New York, NY, USA, 47–54.
77. Bouchet, J., Nigay, L., Ganille, T. (2004). *Icare software components for rapidly developing multimodal interfaces*. Proceedings of the 6th international conference on Multimodal interfaces, ICMI '04. ACM, New York, NY, USA, 251–258.
78. Nigay, L., Coutaz, J. (1995). *A generic platform for addressing the multimodal challenge*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '95. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 98–105.
79. Vo, M.T., Wood, C. (1996). *Building an application framework for speech and pen input integration in multimodal learning interfaces*. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, **6**, 3545–3548.
80. Calder J. (1987). Typed unification for natural language processing. In: Kahn, G., MacQueen, D., Plotkin, G. (eds.). *Categories, Polymorphism, and Unification*. Centre for Cognitive Science University of Edinburgh, Edinburgh, Scotland.
81. Johnston, M., Cohen, P.R., McGee, D., Oviatt, S.L., Pittman, J.A., Smith, I. (1997). *Unification-based multimodal integration*. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98. Association for Computational Linguistics, Stroudsburg, PA, USA, 281–288.
82. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J. (1997b). *Quickset: multimodal interaction for distributed applications*. Proceedings of the fifth ACM international conference on Multimedia, MULTIMEDIA '97. New York, NY, USA, 31–40.

83. Taylor, G., Frederiksen, R., Crossman, J., Quist, M., Theisen, P. (2012). *A multi-modal intelligent user interface for supervisory control of unmanned platforms*. International Conference on Collaboration Technologies and Systems (CTS), 117–124.
84. Sun, Y., Chen, F., Shi, Y.D., Chung, V. (2006). *A novel method for multi-sensory data fusion in multimodal human computer interaction*. Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments, OZCHI '06. ACM, New York, NY, USA, 401–404.
85. Holzapfel, H., Nickel, K., Stiefelhagen, R. (2004). *Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures*. Proceedings of the 6th international conference on Multimodal interfaces, ICMI '04. ACM, New York, NY, USA, 175–182.
86. Pfleger, N. (2004). *Context-based multimodal fusion*. Proceedings of the 6th international conference on Multimodal interfaces, ICMI '04. ACM, New York, NY, USA, 265–272.
87. Neal, J.G., Thielman, C.Y., Dobes, Z., Haller, S.M., Shapiro, S.C. (1989). *Natural language with integrated deictic and graphic gestures*. Proceedings of the workshop on Speech and Natural Language, HLT '89. Association for Computational Linguistics, Stroudsburg, PA, USA, 410–423.
88. Latoschik, M. (2002). *Designing transition networks for multimodal VR-interactions using a markup language*. Proceedings of Fourth IEEE International Conference on Multimodal Interfaces, 411–416.
89. Johnston, M., Bangalore, S. (2005). Finite-state multimodal integration and understanding. *Natural Language Engineering* **11**(2), 159–187.
90. Bourguet, M. (2002). *A toolkit for creating and testing multimodal interface designs*. Proceedings of User Interface Software and Technology (UIST 2002) Companion proceedings, 29–30.
91. Navarre, D., Palanque, P., Bastide, R., Schyn, A.,Winckler, M., Nedel, L.P., Freitas, C.M.D.S. (2005). *A formal description of multimodal interaction techniques for immersive virtual reality applications*. Proceedings of the 2005 IFIP TC13 international conference on Human-Computer Interaction, INTERACT'05. Springer-Verlag, Berlin, Heidelberg, 170–183.
92. Wu, L., Oviatt, S.L., Cohen, P.R. (1999). Multimodal integration – a statistical view. *IEEE Transactions on Multimedia* **1**, 334–341.
93. Pan, H., Liang, Z.P., Huang, T. (1999). *Exploiting the dependencies in information fusion*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **2**, 412.
94. Wu, L., Oviatt, S.L., Cohen, P.R. (2002). From members to teams to committee – a robust approach to gestural and multimodal recognition. *IEEE Transactions on Neural Networks* **13**(4), 972–982.
95. Flippo, F., Krebs, A., Marsic, I. (2003). *A framework for rapid development of multimodal interfaces*. Proceedings of the 5th international conference on Multimodal interfaces, ICMI '03. ACM, New York, NY, USA, 109–116.
96. Dumas, B., Signer, B., Lalanne, D. (2012). *Fusion in multimodal interactive systems: an hmm-based algorithm for user-induced adaptation*. Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems, EICS '12. ACM, New York, NY, USA, 15–24.
97. Damousis, I.G., Argyropoulos, S. (2012). Four machine learning algorithms for biometrics fusion: a comparative study. *Appl. Comp. Intell. Soft Comput.* 242401-1-7.
98. Huang, X., Oviatt, S., Lunsford, R. (2006). Combining user modeling and machine learning to predict users multimodal integration patterns. In: Renals, S., Bengio, S., Fiscus, J. (eds.). *Machine Learning for Multimodal Interaction,* vol. 4299 of Lecture Notes in Computer Science Springer Berlin, Heidelberg, 50–62.
99. Reeves, L.M., Lai, J., Larson, J.A., Oviatt, S., Balaji, T.S., Buisine, S., Collings, P., Cohen, P., Kraal, B., Martin, J.C., McTear, M., Raman, T., Stanney, K.M., Su, H., Wang, Q.Y. (2004). Guidelines for multimodal user interface design. *Communications of the ACM* **47**(1), 57–59.
100. Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM* **42**(11), 74–81.

第10章

生物计量学中的多模态交互：技术与可用性挑战

Norman Poh[1]，Phillip A. Tresadern[2]，Rita Wong[3]
1. 英国萨里大学计算机系
2. 英国曼彻斯特大学
3. 英国萨里大学

10.1 引言

在我们这个互相联系日益紧密的社会中，确立个人身份变得越来越重要。传统的身份认证都是基于人们所持有的（例如，令牌/门禁卡）或人们所知道的（密码），而现在这些已经不够用了。生物计量学是利用人的生理特征或行为特征，来进行个人身份鉴定的科学。这项技术有效考察人体固有特征，为身份安全提供了重要的解决方法。单一模态的生物特征识别系统拥有其自身的局限性，因而我们往往需要一个多模态生物特征识别系统。这为多模态交互提供了一个极好的案例研究。

本章将描述关于多模态生物计量技术在不同领域运用中所涉及的技术设计和可用性，这包括从大型多站点边界控制到确保个人便携式设备安全等诸多方面。

在移动生物计量工程（MOBIO）背景下，我们结合人脸和语音生物特征来确保安全、快速的用户验证，保证请求访问数据的人得到授权。此外，我们也陈述了能够使盲人用户在他们的移动设备上使用人脸生物计量方面的经验。为了帮助他们捕捉到良好的人脸图像，我们设计了一个由他们面部图像质量和盲人用户技术可用性评估系统驱动的音频反馈机制。

10.1.1 身份确认动机

每天各处都会有以下这类的问题："他（她）真的是他（她）所说的那个人吗？""他（她）有权访问这个场所/资源/信息吗？""他是官方所寻找的那个人吗？"传统上，一个人的身份一般是由驾照、护照或国家身份证件进行验证。要想访问受保护的资源，只有该人知道他/她的密码或个人识别号时才会被授权。基于令牌或基于密码的身份验证授权手段会很

容易被犯罪分子运用日益复杂的技术而利用。这已经造成了重大的经济损失，也造成了我们这个现代社会的信任损失。例如，根据 Javelin Strategy & Research 报告㊀，美国 2012 年身份欺诈造成的损失达 210 亿美元，而在英国，据估算达 13 亿英镑。每年因身份欺诈，世界各地企业的损失可达 2210 亿美元。由此可见，对可靠的用户身份验证技术的需求尤为重要。

10.1.2　生物计量学

在确定个人身份方面，生物计量技术作为一种合法的方法越来越被广泛地接受。今天，通过使用虹膜、人脸和/或指纹以及旅行证件提高了边境控制的安全性。然而，单独使用一个生物计量系统往往是不够的。单模态生物计量系统必须应对各种问题，例如噪声数据、对象内部变化、采集过程中设置的自由度限制、非普遍性（即不是每个人都能提供清晰的指纹）、欺骗攻击和一些用户不可接受的错误率[1]。噪声的出现是由于生物计量特征的改变（例如，寒冷所致的声音改变），不完美的传感器（例如，弄脏的传感器）或采集生物计量特征的传感器所处的环境（例如，人脸图像受照明条件的影响）。

改善系统鲁棒性的一种方法是使用多模态生物计量技术。因为不同的生物计量方式受不同噪声源的影响，与任何单一的生物计量系统相比，多模态生物计量系统通常会实现性能的显著增益。

10.1.3　多模态生物计量学的应用特征

以下所列的是一些需要用多模态生物计量方法解决的相关应用的标准。

- 录入要求：当一个生物计量技术大规模推出时——例如，在人口层面——必须考虑多个生物计量模态。这是因为由于工作、健康或残障原因，用户人口中有一小部分可能无法提供可用的指纹。例如，没有右手的人不能提供任何右手手指的指纹。出于这个原因，大型生物计量项目，如美国访客和移民身份指示技术（US－VISIT）和唯一标识（UID）项目必须要考虑生物计量的多种模态来确保该项技术在目标人群中可以被所有用户使用。美国访客和移民身份指示技术（US－VISIT）项目要求进入美国的访客在入境处必须提供左手和右手食指的指纹图像，还有面部图像，而唯一标识（UID）项目使用指纹、虹膜和面部生物计量技术[2]。
- 欺骗的风险和可行性：在涉及重要基础设施的应用程序中，入侵的风险是极大的，因而可以使用多模态生物计量技术。因为多个生物计量特征的运用会使得利用受害者所有生物计量模态的仿造或行骗来非法获得设施的访问权变得非常困难。例如，手指静脉和虹膜的生物计量方法很难收集，因此，难以用来行骗。将这些模态与其他的生物计量模态相结合可以阻止针对生物计量传感器的恶意袭击。然而，最近的研究[3,4]表明，即使受攻击的风险降低了，一个多模态的生物计量系统仍然可以容易受到恶意袭击。这是因为如果一个或多个生物计量子系统被破坏，那么多模态系统的性能将会被影响。

㊀　见 https：//www. javelinstrategy. com/brochure/276。——原书注

- 完整性要求：一旦用户成功地通过验证和被授予安全资源的访问权，通常还要确保同一用户实际中在使用该系统。由此，在经过初步验证后，还需要一个连续的、非侵入性的认证解决方案。例如，在逻辑访问控制中，当用户登录到一个安全资源时，用户使用的终端将尝试连续地验证用户。这可以防止攻击者访问系统，而真正的用户却时不时缺席。Altinok 和 Turk（2003）[5] 运用人脸、声音和指纹描述了这条思路的研究方法；同时 Azzini 等人（2008）[6] 和 Sim 等人（2007）[7] 使用了人脸和指纹的识别模态。Niinuma 和 Jain（2010）[8] 运用面部和服装颜色识别模态来进行连续认证。当面部不可观察时，后者所提供的信息是最有用的。
- 精度要求：用来证实多模态生物计量系统运用的最苛刻的应用之一是消极识别或记录的重复数据删除。此举的目的是防止身份的重复条目。与积极识别不同的是，消极识别确保一个人在数据库中的不存在。该应用的实例是防止某人以两个不同的身份领取两倍的社会福利；或防止被列入黑名单的人员进入一个国家。十指的生物计量和图像可以确保重复数据删除精度高于 95%，而依赖于数据采集的质量，虹膜模态的加入可以提高精度高达 99%[2]。
- 不受控制的环境：大多数生物计量应用程序需要用户的合作，然而也有利基应用程序并不需要用户之间的合作。解决这一问题的生物计量研究思路被称为非合作生物计量技术。通常情况下，受试者与传感器之间的距离有几米，他们可能并没有意识到自己正在被监视。该生物计量传感器在关键位置不断跟踪和识别所有通过此地的受试者。对于这个应用程序，通常需要几个生物计量模态或视觉线索，但实际的识别机制可能只依赖于一个或两个可用的生物计量模态。Li 等人（2008）[9] 用一个宽视场（FOV）相机加上两个窄视场相机来进行海上监视中的对象跟踪。如果宽视场相机检测到人的剪影，一个窄视场相机将被激活拉近到人的面部，另一个将试图获得人的虹膜。

在法医学的应用程序中，Nixon 等人（2010）[10] 建议结合步态和耳朵来确认罪犯。因为罪犯往往会通过伪装或遮蔽来试图避开自己的身份，而步态往往是可用的自然生物计量候选项。同时，耳朵的形状随着时间的变化几乎不会改变，从而使步态和耳朵的结合成为了在非合作（如取证监视的身份识别）的情况下一个潜在的有用法医学工具。

在我们的研究中，我们运用配有麦克风和摄像头的通用移动设备解决了生物计量认证的问题。被称为“移动生物计量”的该领域，在实际应用中具有重要的作用，因为该生物计量系统可以防止其他人访问可能非常隐私和高度敏感的数据，如存储在手机上的信息。此外，生物计量系统也可以被用于电子交易的认证机制。这会使得交易服务更加具有价值，也会赢得更多的信任。我们的研究[11] 表明将说话的人脸图像与同时记录的语音相结合的认证性能要比单独使用任何生物计量方式更好。上述三个例子表明，多模式生物计量技术是有助于在不受控制的环境中进行身份识别。

10.1.4　2D 和 3D 人脸识别

最普遍的生物计量方式之一大概就是脸部了。甚至在摄影变得普遍以前，面部的画像已

经被用于通缉犯人。1882 年 Alphonse Bertillon 发明了罪犯识别系统，现在被称为 Bertillon 系统，该系统采集人体尺寸，以及面部图像。

如今，摄像头已经具备了人脸检测功能。脸谱网提供了人脸标签服务，在上传的照片中会自动识别人脸。其面部识别引擎，由 Face. com 正式提供，但现在也成了脸谱网的一部分，它能够在不受约束的环境中识别图像[12]。正因为此，以及大多数移动设备上都装有人脸识别软件的事实，我们在本节会简短讨论自动 2D 和 3D 人脸的识别。

人脸识别的过程通常包括三个主要阶段：检测，特征提取（降维）和分类（存储）。检测阶段一般会包括面部定位与规范化两部分，以便处理视角和光线的变化。

人脸识别技术可以依照所使用的传感器类型：图像传感器、摄像机和深度传感器来进行分类。2D 面部识别是目前最常见的面部识别类型。图像传感器，如数字电荷耦合器件（CCD）或互补金属氧化物半导体（CMOS）有源像素传感器，可以低成本生产，并且体型小到可以适合所有个人装置。

早期的 2D 人脸识别算法是基于整体法，如主成分分析（PCA）——该方法的图像显示被称为 Eigenfaces；或，线性判别分析（LDA）——由此的面部显示称为 Fisherfaces。虽然这两种方法的结果是令人满意的，但由于图像的高变化率，使它们都受到了 2D 普遍存在的缺陷的影响，导致它们都缺乏鲁棒性。

因此，最终使用的是基于分块的方法[13]。这些方法将图像分解成局部区域或分量，以便通过分块来识别图像。随后分量方式结果相结合形成最终输出假说。

基于视频的人脸识别研究[14,15]通过在一定时间内考虑多幅图像延伸了 2D 人脸识别，以消除在单一 2D 人脸图像中显示的不确定性。虽然基于帧的人脸识别方法采用时间投票方案很常见，但较为强有力的方法或以结合来自图像帧级的假设为目的，或以获取比任何单一图像具有更高的分辨率的图像为目的。后项技术被称为超分辨率人脸识别[16]。

尝试克服头部姿势所带来的几何失真的另一种方式是通过图像弯曲的方式将一个非正面图像弯曲成一个正面图像。主动外观模型（AAM）是运用这种方式的一种先进的方法。AAM 把感兴趣的图像构成有形状和文本（外观）的模型。当应用到面部图像时，兴趣点往往在人脸图像周围做手动标记，从而使面部特征随时被跟踪。正是由于跟踪点，非正面的人脸图像可以被弯曲到一个正面的图像。所得到的弯曲图像往往产生比原来非正面的图像更好的识别性能。

受到 AAM 的启发，这条思路的研究为 2D 人脸识别开发出了 3D 模型。Blanz 和 Vetter (1999)[17]的探索性研究提出了一种 3D 形变模型，该形变模型可以使 3D 模型适用于 2D 图像，这样面部图像就可以重新呈现任何视角。这大大地提高了无所约束的 2D 人脸识别。据 Face. com™（现属于 Facebook™）报道，在这种方法的基础上延伸出的另一个方法使得面部识别从任意角度都可以实现。

市场上最近推出的“2. 5D” Kinect™传感器为人机交互，以及人脸识别开辟了一个新的时代。其软件开发包，被称为“Kinect Identity”，使实时玩家的人脸瞬间识别成为了可能。该传感器提供深度信息，以及一个总是对齐和同步的可见性图像。这可以使包含 3D 数据的

4D 无约束人脸识别[18]随着时间的推移而实现。

10.1.5　多模态案例研究

本章的其余部分，我们将对移动生物计量做深入的案例研究。移动设备是无处不在的，对于个人通信来说，它是用户容易随时随地掌握的。

在移动设备上实现时，生物计量有几个潜在的应用情形。首先，如果移动设备丢失或被盗，生物计量技术可以阻止其被非法利用。其次，它也可以用于数字式记录音频、文本或图像文件，为它们的来源和真实性提供证明[19]。我们把移动设备上的生物计量认证称为“移动生物计量”，或 MoBio。

Mobio 项目提供了一个软件验证层，运用移动设备所捕获的你的脸和声音，确保你是你所说的这个人（见图 10.1）。该软件层不仅验证脸和声音，而且使它们相结合以使系统具有更强的鲁棒性。它还会更新系统模型，允许其随着时间的推移改变条件——这都在消费级移动平台的硬件限制中。虽然其他研究已经调查了人脸和语音的认证[20,21]，但 Mobio 是首个在移动架构提出的挑战性条件下（例如，有限的处理能力，摇晃的手持相机）评估了双模态认证。

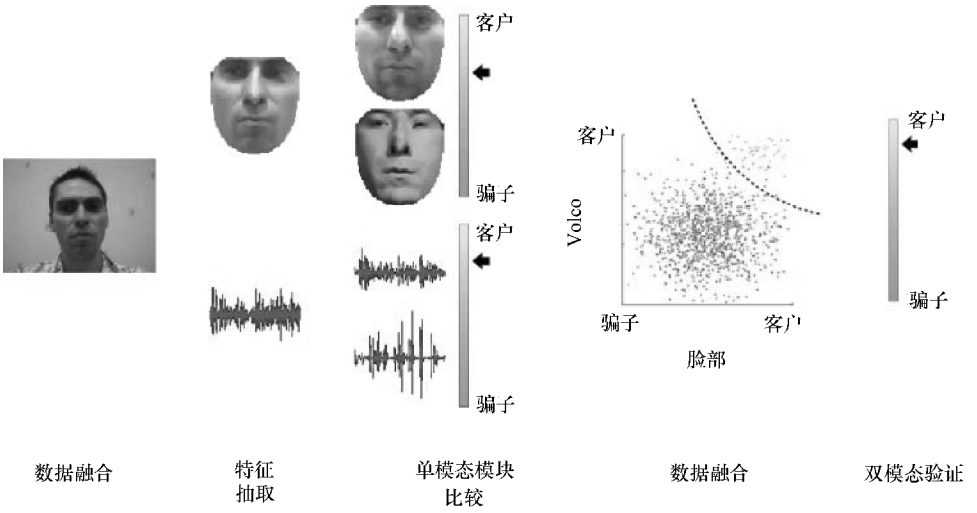


图 10.1　Mobio 身份验证系统计算捕获的正常化的人脸和声音的特征向量，与存储模型中的特征进行对比，为改进的鲁棒性进行评分，并执行双模态验证

该项目的支撑性的关键技术组件包括：

- 误报指数降低的快速人脸检测[22]。
- 在诺基亚 N900 移动设备上实现的高效面部特征定位算法，按照帧速率性能运行[23]。
- 为改进了的人脸验证提供图像描述符[24,25]。
- 使用空间因子方法的一种全新特点的说话人验证[26]，使从会话变异中解耦核心说话人的识别方法与有限的训练数据相结合[27]。
- 一种基于参考文献［28］中更大范围捕获的分级分类器融合算法 EER。

10.1.6　适应于盲人对象

有几个工程，以及用户交互，都对移动生物计量具有挑战性。从工程的角度来看，与台式计算机相比，相对降低的计算能力会使实现移动生物计量变得困难（即内存小，计算能力低，有限的支持浮点计算）。由于设备的便携性和随时随地的使用方式，所捕捉的生物计量数据可能质量不会很好。例如，大家都知道，在一个嘈杂环境录制的语音，语音识别的性能会严重退化[29]。

从用户交互的角度来看，以下因素会使移动生物计量问题困难重重：

- 依赖用户的技能：整个捕捉脸和语音生物计量的过程都依赖于用户的技能。
- 身体残疾的用户：有视觉缺陷的用户很可能会被生物计量验证的移动设备所排除。

关于最后一点，据世界卫生组织统计，全球有超过 1.61 亿的人有视障，它们中的大多数都是老人㊀。考虑到很多用户可能会受到视力缺陷的影响，我们将在本章的 10.2 节阐述这个问题。

因此，在平行于 Mobio 的发展中，我们还探讨了盲人用户要如何适用该平台。从一开始，我们发现面部生物计量将极具挑战性，因为盲人用户无法运用视觉提示得知相机捕捉到自己脸部的图像是如何的。

虽然人脸识别技术的改善表明了光照不好可以补救[30,31]，非正面的姿势可以矫正[32]，但图像复原过程始终都不能使人满意。

此外，面部表情的变化也会对识别性能产生负面影响。在这些因素中，头部姿势可以说是最难以纠正的，因为一个完美的恢复过程非常复杂，计算代价也相当高[33]。

图 10.2 对于为什么头部姿势可能会严重影响人脸识别系统给出了直观的解释。根据所

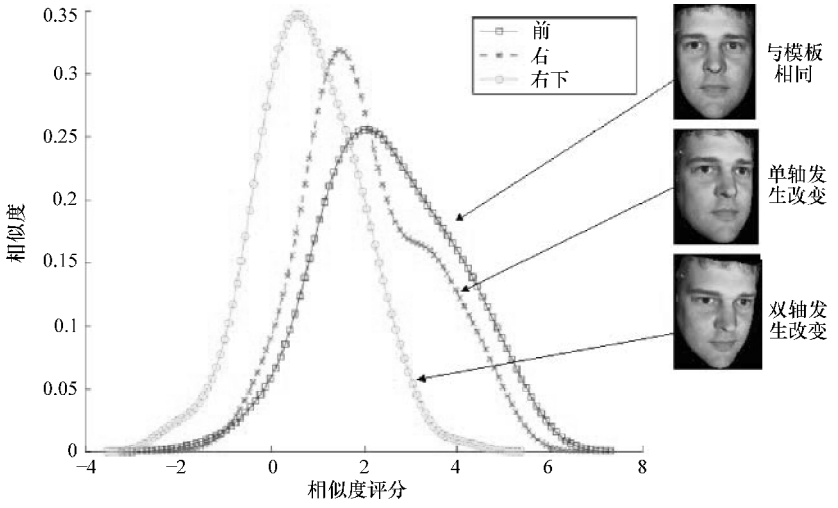


图 10.2　头部姿态对人脸识别系统的影响。更高的相似度分数意味着与真实身份更相似。头部姿势变化，自由度就增加，相似度评分就很可能降低，最终导致身份要求的误拒

㊀ 来源：http：//www. who. int/blindness/causes/magnitude/en。——原书注

给出的数据，对于一个完全正面的姿势，系统给出的所加工的面部图像属于称之为自己的真人具有很高的可能性。然而，只要姿势有所改变，可能性评分值就会降低，并接近骗子的可能性（低分数意味着图像属于骗子）。可以观察到，两个轴上头部姿势的变化（平移和倾斜）对分数的影响超过一个轴上头部姿势的改变（倾斜，本例中）。

10.1.7　本章结构

我们将在 10.2 节介绍 Mobio 平台的设计挑战，然后在 10.3 节中阐述了盲人对象通过音频反馈运用平台的问题。紧接着是 10.4 节的讨论和结论。

10.2　对移动生物计量平台的应用剖析

10.2.1　面部分析

10.2.1.1　面部检测

为了获得用户的外形，我们以包括（某些位置）用户的面部的一个图像为开端，并且对在图像中的面部进行定位，这样就可以对它的位置和大小进行大概的估算（见图 10.3）。这个过程很困难，因为在图像中的外形差异很大，并且我们的系统一定是在不考虑形状、大小、身份、肤色、表情以及光照等条件下对面部进行检测。理想状态下，它应该处理不同的定位与遮蔽问题。但是，在移动验证中，我们假定这个人大多数时间差不多都在直视摄像头。

通过对图像的每一个区域进行分类，把它分成面部与非面部，并且使用现代的模式识别方法从而学会区别面部与非面部的图像特征，这样我们就可以很好地解决这一问题。同时还要考虑两点：一个是如何概述以一种压缩结构形成的图像区域（即计算它的特征矢量），另一个是如何把基于它的特征的图像区域进行分类。

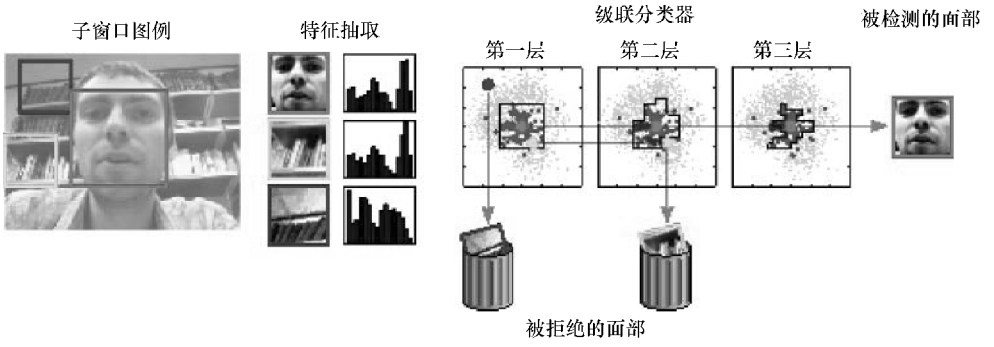


图 10.3　一个窗口在整个图像中滑动，然后在下面的区域被取样并分解成一个特征矢量。这个特征矢量注入到一个简化的分类器，这个分类器会拒绝明显的非面部。接着，被接受的子窗口注入一系列更复杂的分类器直到所有的非面部被拒绝，最终留下真正的面部

当我们搜索图像的时候，面部可能存在的位置成千上万，并且很重要的是，每一个图像区域都应该很快被总结。通过使用局部二进制图案（Local Binary Pattern）[34]中的一个变量，我们可以总结出围绕在每个像素的局部图像统计，同时二进制码可以指示关于它的 8 个邻近的图像梯度方向。然后，对于每一个补丁产生变换后的数值用柱状图来计算，并将它放入一个分类器去选择这一补丁是“面部”或者“非面部”。在实践中，我们使用非常复杂的分类器串联[35]来拒绝大部分图像区域（那些看起来像面部实际什么都不是的图像）。在研究早期，我们使用的是简单但又非常有效的各种分类器。对于看起来和面部最接近的更加有挑战性的图像区域，我们保留了更加精准并对运算要求非常高的分类器以备之需。

我们对关于标准的数据集（例如，BANCA 和 XM2VTS）的实验表明了上述这些方法对真实面部的检测正确率超过 97%。然而，在我们的应用中也会提示用户把自己的面部呈现在图像正中央。这样我们就可以把研究缩小到更小的一个区域，由此进一步降低测试错误率，并且允许更多的判别图像表示来提高检测率。

为了拓展这一基线系统，我们开发了一个原则系统。这一系统可以呈指数级降低检测错误率（背景区域被错误地认为是“面部”），并在同一张真实面部周围形成很多检测集群，而且几乎不会降低真实的接受率[21]。

10.2.1.2　面部标准化

虽然我们能够大概地从图像中的面部周围的长方形图像区域来识别用户，但是还是有一些因素会影响到性能，例如，背景杂波、照明和面部表情等。因此，我们通过把面部标准化来去除所有可能产生的影响，以便使其与用户的存储模式具有相似的属性（在形状和纹理方面）（见图 10.4）。首先，我们对个人的面部特征进行定位，例如，眼睛、鼻子、嘴巴和下巴，并且使用这些特征来去除所有不相关的背景。其次，我们把这一面部进行拉伸到适合之前定义好的形状，由此来弥补由这些人朝向的方向、他们的表情、他们面部的形状（一个弱的验证提示）所产生的差异。最终，我们通过调整亮度使照明标准化并且与一些固定值进行对比。为了进行精准鉴定，生成的图像能够直接与类似标准化的模型图像对比。

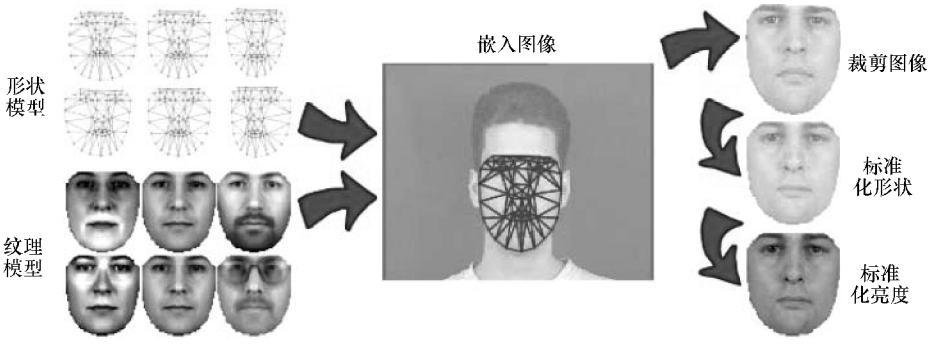


图 10.4　从训练数据中估计得出的形状和纹理的数据模型，并嵌入一个使用主动外观模型（Active Appearance Model）的图像。然后，该基础图像可以被抽样进而去除背景信息，也可被弯曲来去除不相关的形状信息（例如，由于表情产生的形状信息），并标准化亮度和对比度

为了定位面部特征，我们通过使用一个主动外观模型（Active Appearance Model）的新版本把可变形模型放进图像里。主动外观模型使用了现代机器学习技术[23]，是专门为移动架构开发的。主动外观模型使用形状和纹理两个变量的数据模型来描述仅使用了部分参数的面部，这些数据模型是从一系列有着手工标记的特征部位的训练模型学到的。同时，它还学习检测模型何时处于错误的位置，并且去纠正各种参数以便使模型和图像保持对齐。为了预测这些修改，我们训练了一个线性回归来学习样本图像数据与真正参数值之间的关系，这是通过使用已知错位量的各种图像样本来完成的。

当把该模型嵌入新图像时，我们首先把这一模型与粗略的面部检测结果校准，然后抽样并标准化图像的对应部分（见图 10.4）。之后，我们预测并对形状进行修正，使用各种参数来使模型与图像对齐。通过反复多次“样本 - 预测 - 修正”的循环，我们聚合于真实特征的位置，为鉴定提供一个标准化的纹理样本。

与主动外观模型相比，我们的方法能达到类似的效果甚至精度更高（通常两眼之间的距离在 6% 的范围内）。然而，用诺基亚 N900 能够达到三倍的加速比，把过程时间从 44.6ms 降到 13.8ms，因此最终达到帧率性能[23]。虽然该性能使用了由公开数据库训练的模型，可以通过反复训练预测器（线上或者线下）而适应特定用户，但是我们的结论表明相比额外增加的计算成本，该性能并未得到很大的改善。

10.2.1.3　面部确认

考虑到面部的标准化图像，最后的一个步骤就是要给予一个分数来描述它对于声称身份的存储模型的匹配程度，并使用这一分数去决定是否接受或者拒绝当事人的要求（见图 10.5）。同样的，我们把这个看作是一个分类问题，但是基于总结关于他们外形的图像特征，我们想要把当事人标记为一个客户或者骗子。客户可以获得他们所需要的资源，而骗子不可以。

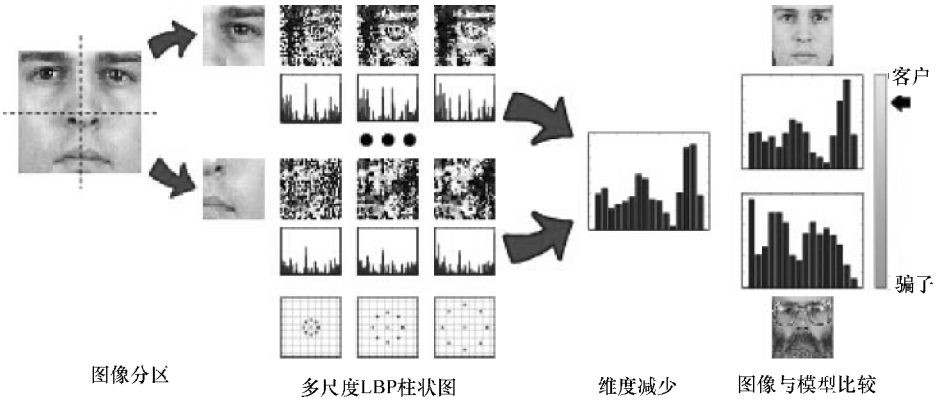


图 10.5　剪裁的面部窗口被细分为很多模块，每一模块使用局部二进制图案在不同尺度进行处理。之后，我们可以在柱状图里捕捉到局部二进制图案值的分布；在与存储模型对比之前，我们拼接并降低柱状图维度（例如，通过主成分分析）

因为光照条件会影响外形，因此我们应用了伽马校正、高斯差滤波与方差均衡来去除尽

可能多的照明影响。对于增加的鲁棒性，我们随后把处理过的图像再分成非重叠子窗口，使描述符对遮挡的处理更稳定，并在 3 个尺度上对每个像素的局部二进制图案（LBP）值进行计算。然后通过使用 LBP 柱状图对每一个窗口进行总结，并使用串联的柱状图作为整个图像的特征向量（见图 10.5）。

为了对观测到的特征向量进行分类，我们计算出它与声称身份的存储模型的差异。虽然我们能够仅仅基于类似测量而做出决定，但是我们使用了鲁棒似然比，其中与背景模型的距离说明了观察结果与声称身份的匹配程度高于平均值，这就印证了我们对分类的信心。通过这一方法，我们能够达到 BANCA 数据集大约 5% 的总错误率的一半（其中错误接受率可能和错误拒绝率相同）。

同时我们还开发了一系列可以提高识别性能的新式图像描述符。其中之一是基于局部相位量化（Local Phase Quantization）的一个图像描述符，用于散焦图像并且对一个模糊的人脸图像实现了 93.5% 的识别率（与之相比，局部二进制图案只能达到 70.1%）[24]。进一步开发这个描述符使之包含多尺度的信息，我们在一个含有大范围不同照明情况的更具挑战性的数据集里把识别率从 66% 提高到 80%[25]。

10.2.2　语音分析

虽然人脸验证技术日臻成熟，但是同时我们也发现一个事实：我们可以纳入基于语音的扬声器验证来更好地利用有可供支配使用的麦克风。

10.2.2.1　语音活动检测

给定一个使用手机麦克风录制的声音样本，我们的首要任务是从背景噪声（对声音识别无用）中分离出语音（对声音识别有用）。然而，与人脸识别相比，语音识别会因为受到各种因素的影响而复杂化，这些因素包括说话者的声道、生活习惯以及使用语言等方面的特征。同一说话者不同时域的语音输出也不相同（例如发生感冒）。

为了对一位说话者的声音进行综述，我们通过一个特征向量概述了在给定时间内的一个小窗口（以数十毫秒为单位）的频率特征，并以声道形状呈现这一变量。具体来说，我们运用倒谱分析来计算经由一个傅里叶变换（Fourier Transform）得出的频谱，并且通过第二个傅里叶变换分解它的对数，在第二次分解之前把频谱映射到梅尔尺度中（其中距离更密切地匹配音高感知差异），求出梅尔频率倒谱系数（MFCC）。

然后我们使用高斯混合模型（GMM）来对一个特征向量进行分类，分为言语与非言语，不考虑特征向量的时间顺序和低通过率影响的输出。虽然这个已经被证明对高信噪比是一个有效的方法，但是有大量背景噪声的环境还需要使用耗费更多信号能量的更复杂方法。

因此我们使用了人工神经网络（Artificial Neural Network）对 MFCC 向量进行分类，这些向量来源于约 300ms 的更长时间语境，被分类成 29 个音素之一，或是被分类成非言语，最终得出与 30 类对应的后验概率向量结果。这些向量随着时间推移变得平滑，通过使用隐马尔可夫模型（Hidden Markov Model）来检测从训练数据中习得的（具体语言的）已知频率的音素顺序，之后 29 个音素类得到整合，进而形成“言语”样本。

由于这一方法在计算上的要求非常高（因此对于嵌入式导入不大适合），我们提出建立一个更简易的特征集，记为“升级二进制特征”（Boosted Binary Features）［Roy et al.（2011b）］，它是基于过滤反应对之间的关系，它也实现了至少与现存各方法一样的不错性能（40 多种可能音素的正确分类大约为 65%），但是仅要求适度的运算量。

10.2.2.2　说话者验证

不考虑背景噪声，我们能够使用有用的言语分段来计算这个人的声音与声称身份的匹配程度，决定是否接受还是拒绝他们的申请。

为了描述这一声音，我们运用了 19 种 MFCC（超过 20ms 窗口计算）和一个能量系数，每个系数与其第一个和第二个导数增长。在通过声音活动检测移除了沉默帧之后，我们对 300 多个帧实施了短时倒谱均值法和方差归一法。

为了对申请人的特征向量进行分类，我们运用了基于高斯混合模型参数的联合因子分析法（Joint Factor Analysis）作为基准。其中混合构成物的加权与协方差在开始时就被最优化，但是中心被设定为数据的函数。这些加权值、协方差和平均值是通过学习一个大型的多人语音集合而获得的，且主体子空间是通过使用已知说话者的数据库习得的，包括综合不同会话时期的话语以减少时期间的差异。而时期子空间则是从剩下的部分习得。

测试时，我们使用每一个训练例子来估算说话者与会话时期，并且使通用模型适用于特定用户的模型。然后我们不考虑时期估算（因为时期匹配并非我们的目标，我们的目标是说话者匹配），并且根据具体说话者的模型来计算出测试例子的相似度。然后使用量化归一法作为分类的手段。

在 BANCA 数据集里，该基准系统达到了对说话者验证的约为 3% ~4% 的等错误率。但是我们证明了我们能够改进相关的 i 向量的估算方法（说话者识别技术的最新发展），使说话者建模的速度加快 25 ~50 倍，而仅仅使用 10% ~15% 的内存，而且仅会对性能造成微小的影响（通常增加的等错误率为 3% ~4%[26]）。

同时，我们还演示了从时期变化模型中去除核心说话者识别模型的过程，这样我们就可以分别最优化两种模型，并且在有限的训练数据条件下得到一个更为稳定的系统，且对性能造成很小或者零影响[29]。最终，我们展示了使用成对的特征实现了 17.2% 的半总错误率（HTER），高于跨越了 17 个其他系统的平均 15.4% 的 HTER，但是比其有效 100 ~1000 倍[37]。

10.2.3　模型适应

生物计量验证中的一个挑战是适应随着时间变化而改变的人的外貌——不管是主观的（如个人打扮）还是客观的（如皱纹），以及适应环境中的会影响识别性能的外部影响（如光照、背景噪声）。因此，在初创时的用户模型并不是固定的——它必须适应当前情况并调整相应的标准来做出接受或是拒绝的正确判断。

在面部验证的实验中，我们开始于从包含多人的训练数据中建立一个外形通用的模型。这有助于构建出没有出现在个人录入数据中的光照和头部姿势的模型。然后我们根据各个特定用户改变该通用模型，调整基于用户具体的训练数据的模型参数。在我们的案例中，使用了高斯混合模型来呈现容貌，因为它可以容忍定位误差。同时，我们又再一次地改编了该模

型来响应任何在拍摄条件中可能发生的变化。

为了说明拍摄环境的变化（例如 BANCA 数据库含有在可控的、不利的和降级的条件下拍摄的例子），在训练中，我们对每一个条件计算了错误分布的参数 q，并且使用了分数归一法，如 Z – norm，

$$Z_{\mathrm{q}}(y) = \frac{y - \mu_{\mathrm{q}}}{\sigma_{\mathrm{q}}} \tag{10.1}$$

或者基于贝叶斯归一法（通过逻辑回归完成）：

$$P(q|y) = \frac{1}{1 + \exp(\alpha_{\mathrm{q}} y - \beta_{\mathrm{q}})} \tag{10.2}$$

来减少拍摄环境的影响（式中，μ_{q}、σ_{q}、α_{q}、β_{q} 是通过学习估算的参数）。测试中，我们计算了与当前环境最接近的、可以被信号质量识别的已知环境，并且根据情况调试了分类器评分。

在实验[28]中，分数标准化在一些测试中降低了 20% ~30% 的等错误率（对于面部从 19.5% 降到 15.31%；对于言语从 4.8% 降到 3.38%）。然而使模型适应拍摄条件对性能产生了更大的效果，在一些实验中降低了高于 50% 的等错误率（对于面部从 19.37% 降到 9.69%；对于言语从 4.8% 降到 2.29%）。

10.2.4　数据融合

至此，视频序列中的每个样本得出的评分都可以说明申请人与其声称的身份的相似度，另一个分数说明他们的声音与声称身份的相似度。为了给出一个生物计量本身表现更为出色的系统，我们融合这两个模态，通过对每个模态单独地进行分类并且把分数结果对馈入到另一个分类器中（分数级融合），或者通过把特征融合并传送到一个单独的分类器中（特征级融合）。因为我们关注的是视频序列，所以在一段时间内对融合分数（或者特征）有益处。

一个天真的方法就是通过求序列的平均值而融合分数级数据。更有理论依据的方法是对观察序列的分数分布建模，并将其与各种从训练数据中获得的与正确和错误匹配对应的分布进行对比。我们对分数分布的非参数统计（例如，均值、方差和内部的四分位范围）进行了计算，以此作为基准，并通过使用逻辑回归获得的分类器把正确和错误的匹配分开。同理，我们运用分数归一法确保来自于不同感知模态的输出具有可比性，同时还要把信号质量考虑进去[38]。

虽然使用专属软件（其中内部分类器操作被隐藏）时分数级融合很受欢迎，但是特征融合法能够捕捉到两种模态之间的关系。然而特征融合可能导致产生一个大型的联合特征空间，其中“维度灾难”成为难题。并且在不同采样率（例如，视频和音频）进行融合源时我们必须要非常谨慎。

因此，我们开发了一个新的特征级融合方法，命名为“升级层次分类器”，它能搜索特征对空间（一个面孔和一个言语片段）来找到二次判别分析（QDA）最小化的错误分类率，在该过程中迭代地对训练样本重复加权。虽然这一方法在控制条件下只会产生轻微效果，但是在一种模态被破坏时它会优于基准的分数级融合系统，表明融合确实会增加系统的鲁棒性。

在另一个实验中，如检测误差权衡曲线（Detection Error Tradeoff curves）（见图 10.6a）

所示[㊀]，融合模态的益处是更为明显的。这说明对于变化的分类器分数的阈值，在错误拒绝率和错误接受率之间做出的权衡关系——接受更多的申请人可以降低错误拒绝率但是会增加错误接受率（反之亦然）。

10.2.5　移动平台实施

为了在移动设备上运行这一系统，我们需要考虑可用的硬件的各种缺陷，例如，低功率处理、单一固定点结构和有限的内存。因此，我们需要开展关注精准性效果的实验，从而计算出能使这个系统更为有效的近似值。

一个非常有效的修正就是通过运用固定点（而不是浮点）运算来实施尽量多的方法。虽然一些现代设备配备有浮点单元，但是它们并不常见而且效率低下。其他可以减少计算的方法还包括对于面部检测运用早期停止准则，并且减少面部特征定位中没有的循环量。由于减少内存消耗同时也对性能有益处，所以我们通过减少参数来进一步完善，例如，LBP 尺度的数量、特征向量的维度以及语音识别所用的高斯混合分量的数量。作为这些近似值的量化评估，我们按照两个标准对 1296 个尺度体系（48 张面孔 ×27 个言语片段）进行评估：一个既可以反映内存消耗还可以反映速度的抽象成本标准，以及一个由等错误率测量的结果泛化性能。果然，有效性的提高以牺牲精准度为代价，然而激增的复杂性导致了受益甚小（见图 10.6b）。

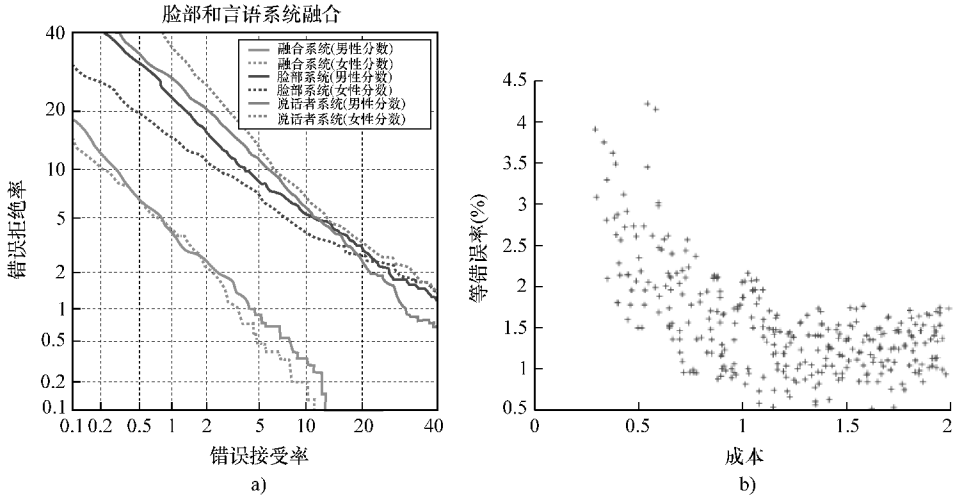


图 10.6　a）测试 MoBio 数据库得出的单模态和融合双模态系统的检测误差权衡（Detection Error Tradeoff）曲线；图示表明对数刻度上的错误接受率和错误拒绝率，这是就范围内一系列决定阈值而言的，其中处于下方的左边点是最理想的。在给定的曲线上的等错误率（EER）位于曲线和直线 $y = x$ 相交之处。b）EER 与各尺度系统的有效性的对比，确认了获取更高的精度是要付出代价的，定义为两个比例（内存消耗以及时间花费）中的相对于基准线系统下方的值

㊀ DET 曲线显示与接收器工作特性（ROC）曲线相同的变量，但是在对数尺度上；这使得曲线几乎是线性的，并给出了更均匀的点分布，使得解释更容易。——原书注

为了在真正的条件下测试这一系统，我们为诺基亚 N900 开发了一个原型应用（见图 10.7）。这款诺基亚手机包含可以进行视频拍摄的前置 VGA 摄像头，一个德州仪器公司 600MHz ARM Cortex – A8 核的 OMAP3 微处理器，以及 256MB 随机存储器。在用户界面和 gstreamer 中使用 GTK 来处理视频拍摄，我们完成了身份验证系统的近帧速率操作。

图 10.7　移动生物计量界面显示的面部检测、面部特征定位（对于形状标准化）以及很多受欢迎的网站中的自动登录和注销的用户界面，例如，电子邮件和社交网络。

10.2.6　MoBio 数据库和协议

Mobio 项目与其他相关项目之间的一个主要差异在于 MoBio 系统是一个使用面部和声音的双模态系统，因此，它需要一个双模态数据集来评估性能。然而，很多公开可用的数据集只包含面部数据或者只包含声音数据，而不是两者均有。即使少有能够做到的也只是包含了在严格控制条件下，使用高质量相机和麦克风录制的视频和音频数据[20,21]。因此对于我们的应用来说并不现实。我们受限只能使用低品质的手持相机。那些很接近的数据集（例如，BANCA 数据集）使用了一个静态的相机，因此它就没有产生我们必须要处理的手部轻微抖动造成的图像抖动现象。

因为我们预期未来会有其他移动识别和验证应用，所以为了研究的目的，我们使用了一个手持式移动设备（诺基亚 N93i）去收集真实并且公众可用的一个新数据库㊀（见图 10.8）。该数据库收集持续了 18 个月的周期，跨越了欧洲六地，包含了 150 个受试者，并且对每一个受试者进行了两个阶段的数据采集。第一个阶段有 6 个部分，当中每个部分包含 21 个视频；第二个阶段包含 6 个部分，当中每个部分包含 11 个视频。每一份测试协议和数据一并提供，定义了数据库应该如何分为训练、开发和测试集，以及如何对评估分数进行计算。该测试协议随后被使用在一个由 14 个地方参与的比赛中：9 个应用于面部验证，5 个应

图 10.8　来自于数据库的屏幕截图，显示出不受约束的室内环境性质和不受控制的照明条件

㊀　见 http：//www. idiap. ch/dataset/mobio。——原书注

用于说话者验证[39]。

10.3　案例研究：为视觉缺陷者进行可用性研究

本节探究了对于有视觉缺陷用户来说的面部识别系统的可用性。各类人机界面的应用尤其是音频界面成功帮助了许多视觉缺陷用户获取信息。受之启发，我们试图通过音频反馈来为视觉缺陷者获得改善的可视图像质量。

这一问题将会用几个阶段来呈现。在第一阶段，我们广泛地评估了头部姿势对图像质量和面部验证性能的影响。在第二阶段，我们开发了一个原型系统把头部姿势评分与频率和节奏整合在一起来提供一个用户交互机制和反馈。最后一个阶段以视觉缺陷者作为受试者开展实验，让他们与一个面部验证系统交互，该系统是由头部姿势驱动的音频反馈改进的。

10.3.1　头部姿势变化对性能的影响

为了量化头部姿势如何影响一个面部识别系统的性能，我们需要一个基于事实的姿势信息注释的数据库。这一数据库必须只能包含一个降级因素，例如，头部姿势变量，而不包括其他因素，例如，照明情况、面部表情以及背景变量。为了这个目的，我们使用了由萨里大学168个受试者组成的3D模型的数据库[40]。对于每一个受试者，我们从不同的倾斜角度和平移角度对他们的2D图像进行解读。这样一来，样本在前额面部图像周围的角度就会更密集，而对于极端的姿势则更加稀疏。研究中使用的81个姿势中的各个姿势平均图像如图10.9所示。

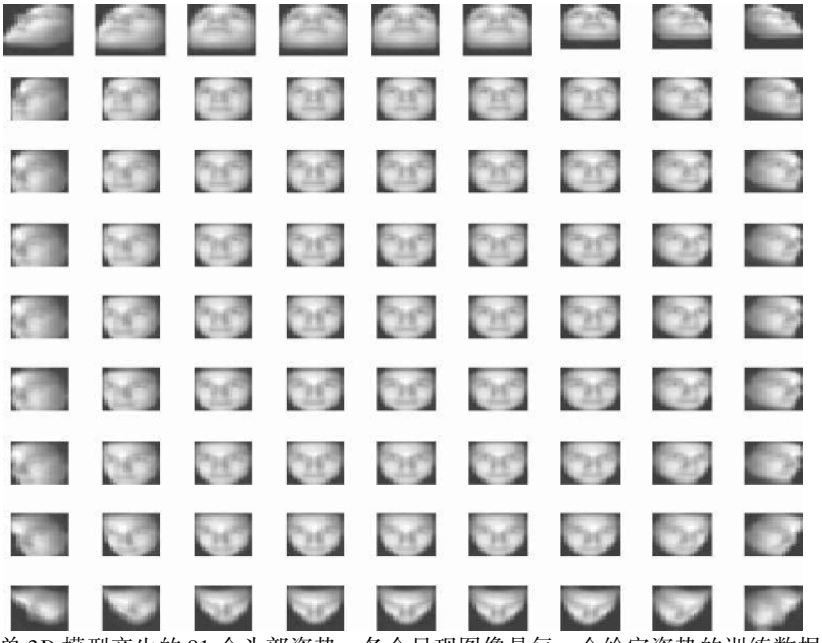


图10.9　单3D模型产生的81个头部姿势。各个呈现图像是每一个给定姿势的训练数据的平均图像。所有平移和倾斜方向中的采样角度都在以下的对数尺度中作为样本：{-45，-16.7，-5.8，-1.6，0，1.6，5.8，16.7，45}

下面的 θ 是来描述平移和倾斜角度的一个向量，$P(\text{error}|\theta)$ 是依存 θ 的系统错误。形式上，我们想要找到容许偏差的 θ_* 集，如此一来，可接受水平的识别错误 δ，在可容忍的范围内：

$$\theta_* \in \{\theta \mid P(\text{error} \mid \theta) < \delta\}$$

式中，δ 是一个很小的数。

在结果中，$P(\text{error}|\theta)$ 是由 EER 近似得到的。EER 是错误接受率与错误拒绝率一致的发生点。对于一个完美的面部证实模块，错误率为 0；对于一个性能比较差的系统，它的错误率最多可能达到 50%（超过这个标准，系统可能会接受骗子并且拒绝一个真实的用户）。如此的一个近似值暗示了错误估算强化了等先验类概率。这样是理想的，因为在这一个典型的生物计量实验中会有比匹配（真实的）更多的非匹配（骗子）进入。

为了估算 ERR，150 个合法用户中的每一个进入都与剩下的 18 个用户（充当骗子）进行匹配，同理对 81 个每个可能的头部姿势进行同样的操作。

我们期望 EER 将随一个倾斜和平移的函数变化。图 10.10 肯定了这一猜测，并且作为我们原型系统的规范基础。举个例子，基于以上的结论，为了能够使等错误率低于 5%（因此，设置 $\delta = 0.05$），头部姿势变量应该处于平移和倾斜方向（θ_* 的范围值）上的 5°之内。另一个方面来看，如果精度小于 15% EER，更大的头部姿势变量就可以得到处理。

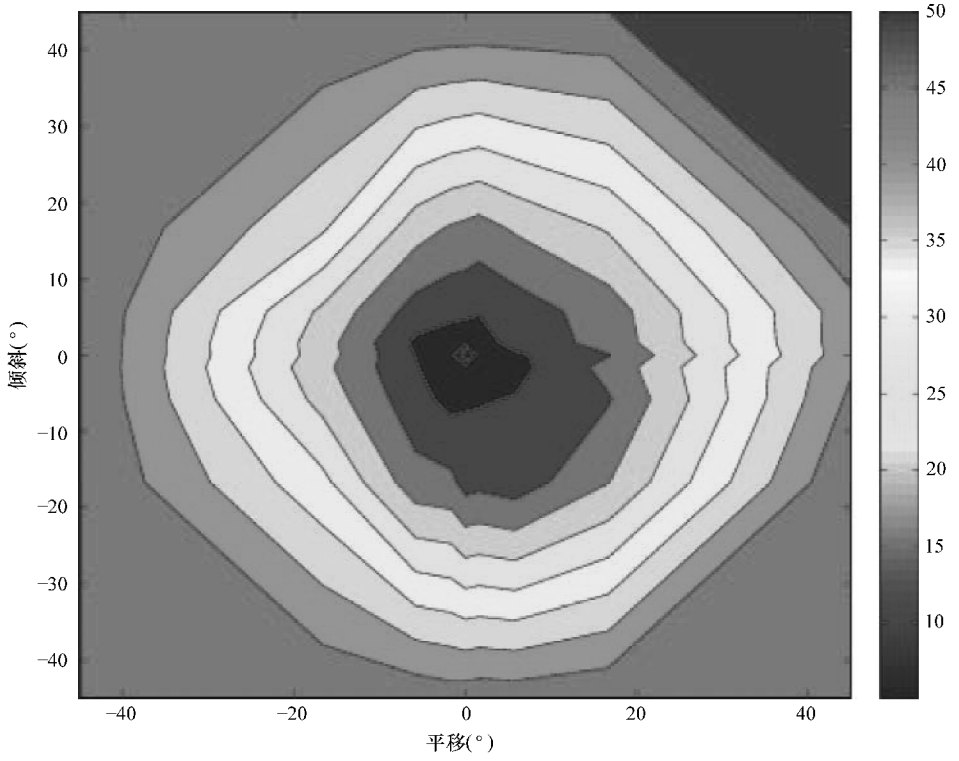


图 10.10　EER 作为平移和倾斜各个方向的一个函数

10.3.2　用户交互模块：头部姿势质量评估

相较于在前一节介绍了头部姿势是如何影响系统性能的具体方法，本节将探讨能够驱动用户反馈的机制。

本研究中我们探索了两种方法：使用面部检测可信度以及估算头部姿势。

10.3.2.1　面部检测方法

近十年来人们已经开始广泛地研究面部检测，并对此提出了很多解决办法[41]。其中包括特征降维方法（例如，主成分分析法，线性判别分析法）、肤色分析法、滤波技术法以及基于图像的方法（例如，AdaBoost 和神经网络）。本研究运用了使用分类器级联的基于图像的面部检测模块[42]，称为“WaldBoost”。这是 AdaBoost 的一个变体，这一最新的方法在已经由参考文献［43］提出。这个检测器非常吸引人，因为它能够进行实时操作，包含变化分辨率的图像，并且它不会被杂乱的背景所影响。而且，我们能够使用这种面部检测器的输出作为质量评估。其输出是一个暗示脸部检测相似度的对数似然比。

用 f 来指代面部检测输出。然后，通过使用之前的同一数据库，我们对 $p(f|\theta)$ 进行估算，其中 θ 是平移角度和倾斜角度的一个向量。图 10.11 呈现了这一分布的中值。我们注意到图 10.11 在一定程度上与 EER 等值线图相关联。这表明了基于面部检测输出的驱动反馈是可行的。

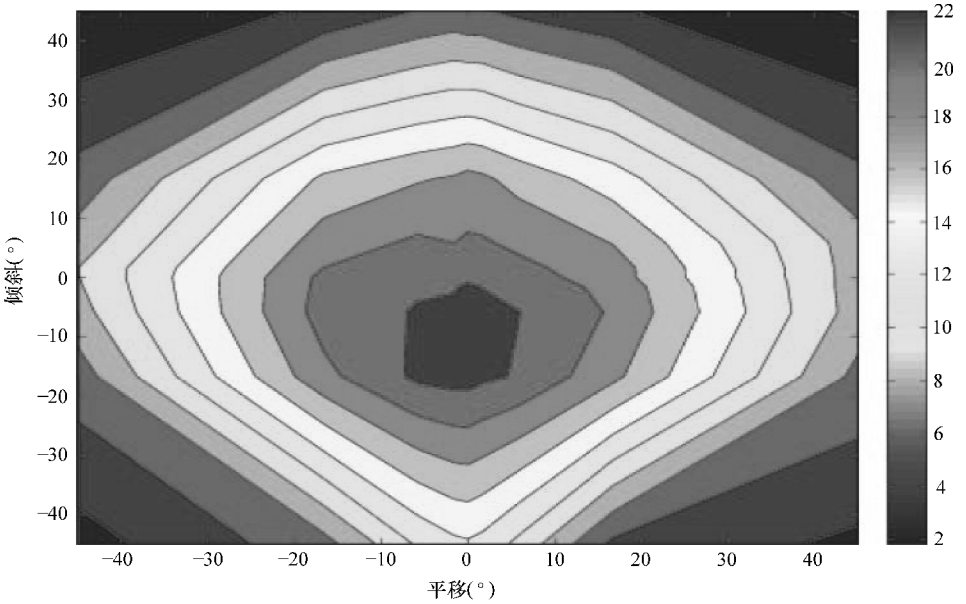


图 10.11　面部检测输出的中值作为平移和倾斜方向函数的等高线图

10.3.2.2　头部姿势估算方法

尽管对于头部姿势估算的算法很多[44]，但是我们的选择还是严格限定于一些手持设备

的应用要求：实时、轻量计算、小内存消耗和粗略头部姿势估算。为了把姿势信息反馈给用户，实时要求非常重要。

在本研究中，我们运用了可以满足这些要求降维方法，特别是学习判别投射和原型（LDPP）算法[45]。同时，这一算法学习了直线投射基础以及对于最近邻（Nearest - Neighbor）分类的一系列原型。但是，因为我们的任务是回归而不是分类，所以为了满足我们的需要算法被稍微修改了一下。这里的描述简要地解释了算法以及引入的修改。

$x_{N\times 1}$为一个随意姿势的剪裁图像，代表一个列向量有 N 个图像像素（灰度水平）。而投射的图像（其中尺寸 $b\times 1$）可以表示为

$$\widetilde{x} = B^{\mathrm{T}}x$$

式中，$B_{N\times b}$是一个投射基础矩阵，并且 T 是一个矩阵转置操作。要用的基数，也就是 b，是由牺牲泛化性能来实现运算速度的要求决定的（b 值较小暗示了运算量较少）。注意投射基础矩阵 B 不一定要正交，因为它是由 LDPP 通过梯度下降获得的。

θ 为包含一个头部姿势的倾斜度和平移度的双变量向量。通过使用被角度向量 $\theta_i(i=1,2,3,\cdots,81)$ 限定的 81 个离散姿势，我们能够有效地覆盖整个视野范围的头部姿势范围的连续。而且，视 p_i 为 81 个头部姿势（见图 10.9）其中之一的原型（平均图像），并且 $\widetilde{p}_i = B^{\mathrm{T}}p_i$ 为其对应的投影向量。

LDPP 的原始公式通过最近邻原则解决了分类问题，即一个求解样本$\widetilde{x}$被分配了类别标签，标签的原型 $\widetilde{p}_i$ 是与求解样本最接近的。然而，因为我们这里的问题是回归，所以最近邻原则在这里不适用。我们需要一个能够量化求解样本$\widetilde{x}$与给定的原型 $\widetilde{p}_i$ 的相似度的函数，对于所有可能的姿势范围 i。当 $\widetilde{x}$接近 $\widetilde{p}_i$ 时，类似测量应该有很高的响应，最后在 $\widetilde{x}=\widetilde{p}_i$ 时达到一个峰值。相反地，当$\widetilde{x}$远离$\widetilde{p}_i$ 时，测量值应该很小，最终到达 0。

可以证明以上特征的一个可能的测量是径向基函数（RBF），它同时普遍地被称为高斯内核（Gaussian kernel），以形式 $\exp\left(-\dfrac{\|\widetilde{x}-\widetilde{p}_i\|^2}{2\sigma^2}\right)$呈现，其中 σ 是内核宽度，当样本$\widetilde{x}$被定位在远离矩心$\widetilde{p}_i$ 时，它是一个能够控制这一测量急剧下降的参数。σ 的最理想值是数据依存与问题依存（因为它是被定义在数集$\widetilde{p}_i$，$\forall i$ 范围上的），并由多次实验决定。我们发现 $\sigma=1$ 对于我们的任务非常适合。当径向基函数被运用在其他姿势的场景中时，它可以被解读为头部姿势的后验概率，即

$$P(\theta_i \mid \widetilde{x}) = \frac{1}{Z}\exp\frac{(-\|\widetilde{x}-\widetilde{p}_i\|^2)}{2\sigma^2}$$

式中 Z 是一个正则化因子，遵守概率公理，即 $\sum_i P(\theta_i \mid \widetilde{x}) = 1$。由此可知，它遵循 $Z = \sum_i \exp\left(-\dfrac{\|\widetilde{x}-\widetilde{p}_i\|^2}{2\sigma^2}\right)$。然后，期望得到的头部姿势是

$$\hat{\theta} = \sum_i \theta_i P(\theta_i \mid \widetilde{x})\ \forall_i \text{s. t. } P(\theta_i \mid \widetilde{x}) > \eta$$

这在本质上是头部姿势的后验分布中的一个期望运算（就平常统计来看）$P(\theta_i \mid \widetilde{x})$，考

虑到在降维当中的观测值 x。

就条件 $P(\theta_i|\widetilde{x}) > \eta$，其中 η 是一个很小的值，安排在这里是由于原始图像 x 不是一个面部图像。因此，RBF 的响应，也就是 $P(\theta_i|\widetilde{x})$ 很可能是随机的，以至于 $P(\theta_i|\widetilde{x})$ 对于所有 i 都会很小。这个的结果是 $\hat{\theta}$ 会趋于平均值。通过设定 η，对应的 RBF 响应过小的头部姿势就能够被有效地排除。

另一个我们考虑到的完整性检查是为了确保 x 的确是在头部姿势估算之前的面部。这是通过在之前 10.3.2 节已经讨论的经过质量（f）函数完成的面部检测可信度而实现的。由于头部姿势估算细节在本章中不是很重要，这一方法的效用就不再进一步讨论了。

对本节进行总结之前，对由 $\widetilde{x}$ 覆盖的样本分布进行可视化呈现是很有指导意义的。为了这一目的，我们选择了 5 个独特的头部姿势，包含了基本的正面部分、左上部分、右上部分、左下部分和右下部分的姿势。图 10.12 显示了关于测试数据集中这些姿势的散布式图示。如图所示，所有的姿势在一定程度上已经被很好地隔开了。

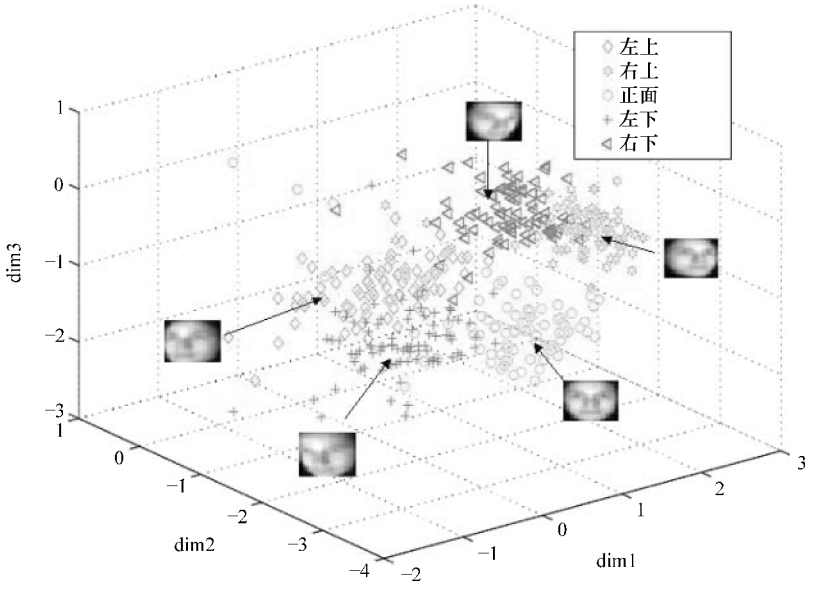


图 10.12　3D 坐标系中 81 个头部姿势中的 5 个散布式图示

10.3.3　用户 – 交互模块：音频反馈机制

一个能够让头部姿势的质量信息反馈给用户的通用办法就是连续评估错误概率并对反馈机制即时控制（见算法 1）。

让我们来定义 $q = [f, \hat{\theta}]$ 的质量信息，它是组成面部检测和头部姿势的一个量。质量条件的错误概率由 $P(\text{error}|q)$ 代表。在文献中，有很多方法可以用来估计 $P(\text{error}|q)$，例如，广义线性混合模型（GLMM）[46] 和逻辑回归（注意逻辑回归是前者的一种特殊情况）。使用 GLMM 的优点是确定不同因子或同时协变量的可能性，例如，性别、出现的类别以及民

族等。

尽管向量 q 的信息量很大（即包含估算的头部姿势和面部检测可信度），但是在如何使大量的信息以有意义的方法传达给用户这一点上不是特别清晰，如使用 3D 声音引导用户或者给他们明确的指示。两种情况下，向用户传达这个信息可能会对其造成一些心理上的负担。应该注意的另一点是，只有面部检测可信度和头部姿势关联很强，即 $P(f \mid q)$ 与 $P(\text{error} \mid q)$ 紧密相关（对比图 10.10 与图 10.11）。基于前面的推理和观测，我们没有使用一个单独的数据库对 $P(\text{error} \mid q)$ 进行估算，而是选择以下这一更为简便的确定性函数：

$$\text{quality}(f) = \begin{cases} \text{不可知}, & f \leqslant \Delta_{\text{lower}} \\ \text{非正面}, & \Delta_{\text{lower}} < f < \Delta_{\text{upper}} \\ \text{正面}, & \Delta_{\text{upper}} \leqslant f \end{cases} \tag{10.3}$$

这显示了被检测的面部质量是由面部检测输出的一个较低的阈值（_ lower）和一个较高的阈值（_ upper）所决定的。

算法 1　头部姿势驱动音频反馈模块

$\delta \in \mathfrak{R}$：一个容错阈值

```
while true do
    获取一个样本
    估算质量 q
  if P（error | q） <δ then
        执行匹配
        退出
    else
        产生反馈
    end if
end while
```

上述公式的一个即时效应就是根据面部质量状态，不同地驱动用户反馈机制。我们都知道尽管现在已存在设计有反馈的生物计量机制，但它们都非常的基本。举个例子，反馈机制信息包括两种状态，一种标记着数据获取过程的开始，另一种标记着结束。在我们提出的反馈机制中，更加丰富的信息（头部姿势）被传达给用户。然而，该信息没有像使用估算头部方法那样获取的信息丰富，如前一节讨论过的。这是因为我们没有办法把更加丰富的信息以一种有意义的方式传达给用户。所以，这部分是未来研究的方向。

已经决定了要给予用户高质量信息水平，接下来的问题就是反馈模式的实际形式，这可以通过不同方法传送给用户—— 屏幕显示的就是视觉反馈；或声音就是音频反馈；或振动就是触觉反馈。

在本研究中，音频反馈将会被采用。我们为了增加频率，创建了 3 个不同的正弦信号波，而且用 3 种不同节奏来表明不同的定性阶段（从不可知到非正面部分到正面部分）。

在我们的研究中，“不可知”定性阶段与一个慢节奏演奏的低频声音有关，“非正面”与较快节奏演奏的中等频率的声音有关，最后“正面”与一个以最快节奏演奏的最快频率

的声音有关。使用的频率分别是400Hz、800Hz、1200Hz。

在获取过程中，反馈得以即时并且连续地提供。图10.13展示了头部姿势质量评估与反馈机制（虚线）整合成的生物计量系统的新架构。当一个用户获得一个生物计量数据的时候，质量在质量评估模块中得到了检查。如果质量被认为很高，生物计量数据会被传送到特征提取模块，否则系统将把质量反馈给用户，并且为了获得新的生物计量数据，需要进行另一次交互。该过程将会一直持续到超时结束，或直到获得足够质量的头部姿势为止。

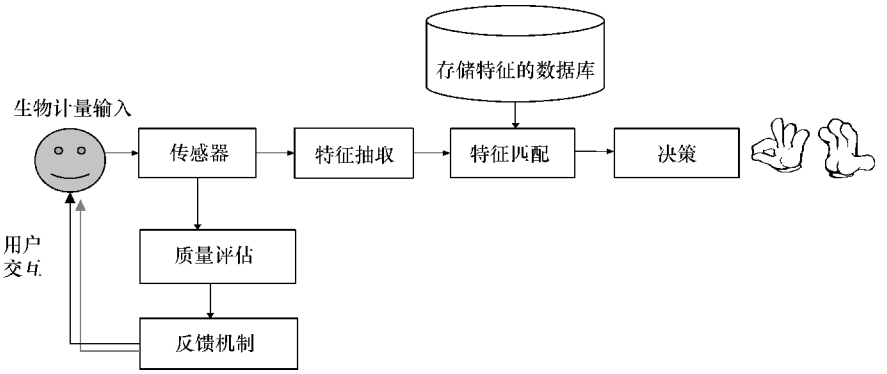


图10.13　生物计量系统中的质量评估和反馈机制的系统架构

10.3.4　视觉缺陷者的可用性测试

本节探讨了视觉缺陷用户如何与移动生物计量系统交互的可用性问题。为此，我们根据不同年龄群体、性别以及缺陷程度，从马来西亚槟榔城圣尼古拉斯视觉机构招募了40个视觉缺陷用户受试。图10.14描述了对象的人口统计结构。

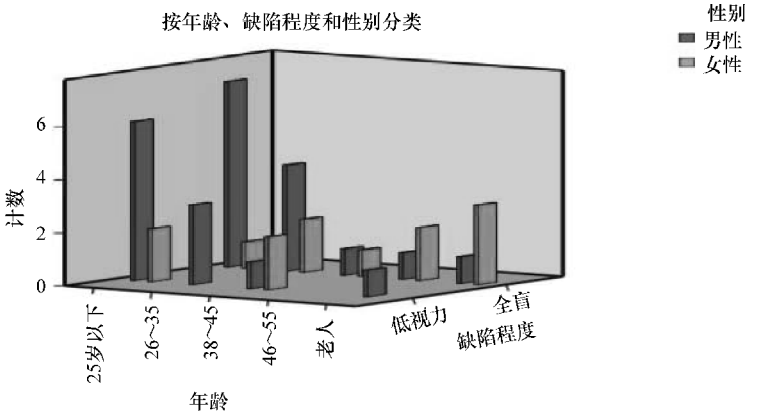


图10.14　受试者的人口统计信息

虽然参与者可以使用移动电话，但是他们不熟悉相机的功能。他们一些人已经有过计算机培训的经历。他们非常配合，并且对声音变化非常敏感。

在测试期间，参与者被安排在一个安静的房间，并且被要求使用我们的原型系统去完成一次图像获取任务。每一个受试者都通过一个视频短片获知他们需要拍摄下其面部位置尽可

能正面的图像。他们必须在 3 个条件下完成拍摄：无反馈，有音频反馈和在口头指示后给予音频反馈。通过音频反馈，不同频率音调的声音将被播放，以提供对他们头部姿势的指示，如式（10.3）所示。音频反馈之后的指示模式——音频 + 指示是指给予参与者音频反馈与口头讲解，告诉他们频率和节奏与图像质量的联系。参与者被特别告知，在图像获取过程开始之前，其手持相机应与自己保持一臂距离。

在一次实验中可能会出现两种结果，即没检测到面部，或是面部在序列中被检测到。第一种情况是失败的，而第二种情况推定为是成功的。但是后者的成功程度仍需要进一步区分，取决于面部检测可信度。因此，从每次实验事件中，我们可以得出两种统计方法：图像获取成功或失败，以及在成功事件中面部检测器产生的可信度值。

图 10.15 总结了给定的条件下关于视觉缺陷用户的成功率或者是失败率。图示表明了音频反馈成功率以及指示 + 音频反馈的成功率，会比在基准线条件下没有反馈的成功率（只有 49%）要高得多（同时分别有 65% 和 94%）。因此，当系统补充音频反馈和指示的时候，视觉缺陷的用户就可能使用移动生物计量技术。

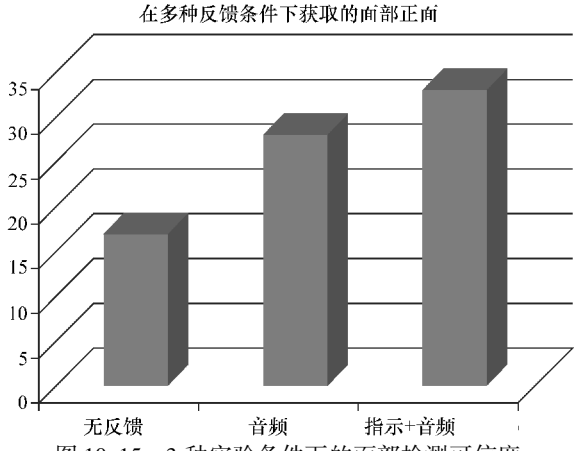


图 10.15　3 种实验条件下的面部检测可信度

每个实验阶段获得面部检测可信度的程序如下：当序列没有检测到面部，面部检测可信度被设置默认值为 0。当面部被检测的时候，序列中的面部检测可信度的最高值被使用。图 10.16 显示了一些检测到的面部图像。

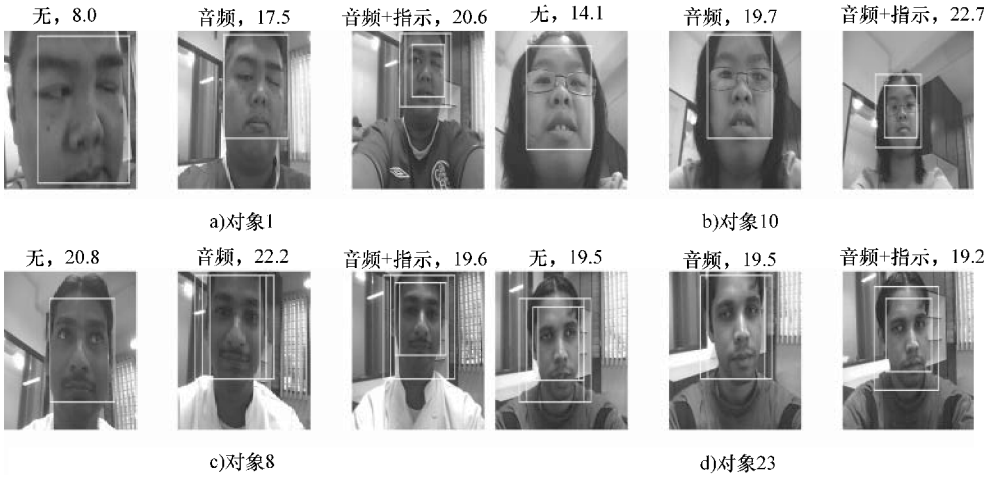
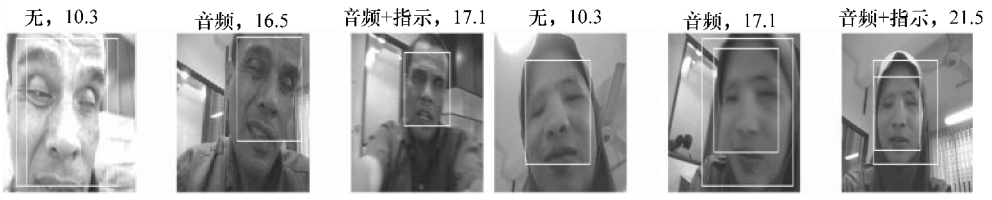


图 10.16　取自第一阶段的 6 个随机抽取的受试者，4 个实验配置中每一个里面可信度最高的面部检测图像的例子。如图所示，我们能观测到有音频 + 指令的图像有可能比只有音频或者只有指令的面部检测得到的图像的可信度更高



e)对象35　　f)对象28

图 10.16　取自第一阶段的 6 个随机抽取的受试者，4 个实验配置中每一个里面可信度最高的面部检测图像的例子。如图所示，我们能观测到有音频 + 指令的图像有可能比只有音频或者只有指令的面部检测得到的图像的可信度更高（续）

同时，为了考查 3 个条件下被检测到的面部可信度的平均值是否差别很大，我们完成了对面部检测可信度的连续值配对 t 检验。表 10.1 显示了结果。如图所示，在配对 t 检验的每一次对比中，结果是全部都很显著的（5% 以下）。这进一步证明了在提高可用性方面，我们设计的机制是非常有效的。

表 10.1　在不同条件下实施的显著的配对 t 检验

实验	配对差异					t	df	Sig.（2 - tailed）
	均值	标准差	标准误差均值	95% CI				
				低	高			
无 FB 与 A	-4.64	8.42	1.40	-7.49	-1.79	-3.31	35	**0.002**
无 FB 与 A + I	-7.74	7.58	1.26	-10.31	-5.17	-6.124	35	**0.000**
音频与 A + I	-3.01	4.86	0.81	-4.74	-1.45	-3.824	35	**0.001**

注：FB 代表反馈，A 代表音频，I 代表指示，CI 代表置信区间。

10.4　讨论与结语

身份验证在我们今天的日常生活中发挥着重要的作用。生物计量学作为一种实现这种验证的技术，仍然面临着许多挑战。人类能在不受限制的环境下识别熟悉的面孔，但是生物计量系统可以识别千万张面孔，不过需要在严格控制的环境下。这一系统很容易受到环境噪声源的影响，这一点人类能够轻易自然地克服。

本章已经探讨了多模态生物计量作为一种可利用的方式促进人类与系统进行交互。我们已经对需要多模态生物计量的大量场景进行了调研。此外，同时我们还对使用面部和声音特征的移动平台进行身份验证的一个新系统进行了概括。具体来看，最新发展的视频模块可以检测、标准化和验证面部，而音频模块则可以分段言语并验证说话人。为了确保系统的鲁棒性，我们使模型适用于评估拍摄条件并融合了多信号模态，这些都是在一个有诸多局限的消费级移动设备中实现的。

在一个独立的个案研究中，我们同时说明了移动生物计量技术对于视觉缺陷用户的潜在服务价值。通过提供合适的反馈，获取的信号质量可以在很大程度上得到提高。

致谢

本研究得到了欧盟（EU）第 7 次框架研究课题的资助。MOBIO 项目资助编号为 2143124。所有作者在这里感谢欧盟对于本研究的财政支持，以及联盟合作伙伴的富有成效的合作。特别要感谢 Visidon 有限公司为开发手机用户界面所付出的辛勤努力。Rita Wong 感谢马来西亚槟榔城圣尼古拉斯视觉机构的所有志愿者参与到本次个案可用性研究中。

参 考 文 献

1. Ross, A., Poh, N. (2009). *Fusion in Biometrics: An Overview of Multibiometric Systems*, chapter 8, 273–292. Springer, London.
2. Bhavan. Y., Marg, S. (2009). *Biometrics design standards for UID applications*.
3. Johnson, P.A., Tan, B., Schuckers, S. (2010). *Multimodal fusion vulnerability to non-zero effort (spoof) imposters*.
4. Rodrigues, R.N., Ling, L.L., Govindaraju, V. (2009). Robustness of multimodal biometric fusion methods against spoof attacks. *Journal of Visual Languages Computing* **20**, 169–179.
5. Altinok, A., Turk, M. (2003). *Temporal integration for continuous multimodal biometrics Multimodal User Authentication*, pp. 131–137.
6. Azzini, A., Marrara, S., Sassi, R., Scotti, F. (2008). A fuzzy approach to multimodal biometric continuous authentication. *Fuzzy Optimization and Decision Making* **7**(3), 243–256.
7. Sim, T., Zhang, S., Janakiraman, R., Kumar, S. (2007). Continuous verification using multimodal biometrics. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(4), 687–700.
8. Niinuma, K., Jain, A.K. (2010). *Continuous user authentication using temporal information*.
9. Li, X., Chen, G., Ji, Q., Blasch, E. (2008). *A non-cooperative long-range biometric system for maritime surveillance Pattern Recognition*. 19th International Conference on ICPR, 1–4.
10. Nixon, M.S., Bouchrika, I., Arbab-Zavar, B., Carter, J.N. (2010). *On use of biometrics in forensics: gait and ear*, 1655–1659.
11. Tresadern, P., Cootes, T., Poh, N., Matejka, P., Hadid, A., Levy, C., McCool, C., Marcel, S. (2013). Mobile biometrics: Combined face and voice verification for a mobile platform. *Pervasive Computing, IEEE* **12**(1), 79–87.
12. Taigman, Y., Wolf, L. (2011). Leveraging billions of faces to overcome performance barriers in unconstrained face recognition. *arXiv preprint arXiv: 1108.1122*.
13. Cardinaux, F., Sanderson, C., Bengio, S. (2006). User Authentication via Adapted Statistical Models of Face Images. *IEEE Trans. on Signal Processing* **54**(1), 361–373.
14. Gorodnichy, D. (2005). *Video-based framework for face recognition in video*. Proceedings of The 2nd Canadian Conference on Computer and Robot Vision, 330–338
15. Poh, N., Chan, C.H., Kittler, J., Marcel, S., McCool, C., Rua, E., Castro, J., Villegas, M., Paredes, R., Struc, V., Pavesic, N., Salah, A., Fang, H., Costen, N. (2010a). An evaluation of video-to-video face verification. *IEEE Transactions on Information Forensics and Security* **5**(4), 781–801.
16. Wheeler, F.W., Liu, X., Tu, P.H. (2007). *Multi-frame super-resolution for face recognition*. First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007, 1–6.
17. Blanz, V., Vetter, T. (1999). *A morphable model for the synthesis of 3d faces*. Proceedings of the 26th annual conference on Computer graphics and interactive techniques, SIGGRAPH '99, 187–194. ACM Press/ Addison-Wesley Publishing Co., New York, NY, USA.
18. Schimbinschi, F., Wiering, M., Mohan, R., Sheba, J. (2012). 4D unconstrained real-time face recognition using a commodity depth camera. 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), 166–173.
19. Ricci, R., Chollet, G., Crispino, M., Jassim, S., Koreman, J., Olivar-Dimas, M., Garcia-Salicetti, S., Soria-Rodriguez, P. (2006). *Securephone: a mobile phone with biometric authentication and e-signature support for dealing secure transactions on the fly*. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 6250 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference.

20. Chetty, G., Wagner, M. (2006). Multi-level liveness verification for face-voice biometric authentication. *Biometric Symp.*
21. Teoh, A.B.J., Samad, S.A., Hussain, A. (2005). A face and speech biometric verification system using a simple Bayesian structure. *J. Inf. Sci. Eng.* **21**, 1121–1137.
22. Atanasoaei, C., McCool, C., Marcel, S. (2010). *A principled approach to remove false alarms by modelling the context of a face detector*, pp. 1–11, Aberystwyth, UK.
23. Tresadern, P.A., Ionita, M.C., Cootes, T.F. (2011). *Real-time facial feature tracking on a mobile device.*
24. Ahonen, T., Rahtu, E., Ojansivu, V., Heikkila, J. (2008). *Recognition of blurred faces using local phase quantization.* 19th International Conference on Pattern Recognition, ICPR 2008, 1–4.
25. Chan, C.H., Kittler, J. (2010). *Sparse representation of (multiscale) histograms for face recognition robust to registration and illumination problems.* 17th IEEE International Conference on Image Processing (ICIP), 2441–2444.
26. Glembek, O., Burget, L., Matějka, P., Karafiát, M., Kenny, P. (2011). *Simplification and optimization of i-vector extraction.*
27. Larcher, A., Lévy, C., Matrouf, D., Bonastre, J.F. (2010). *Decoupling session variability modelling and speaker characterisation.* Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH).
28. Poh, N., Kittler, J., Marcel, S., Matrouf, D., Bonastre, J.F. (2010c). *Model and score adaptation for biometric systems: Coping with device interoperability and changing acquisition conditions.*
29. Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters* **18**(9), 859–872. Audio- and Video-Based Person Authentication.
30. Perronnin, F., Dugelay, J.L. (2003). *A Model of Illumination Variation for Robust Face Recognition Workshop on Multimodal User Authentication (MMUA 2003)*, pp. 157–164, Santa Barbara, CA.
31. Zou, X., Kittler, J., Messer, K. (2007). *Illumination invariant face recognition: A survey. First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007*, 1–8.
32. Okada, K., Akamatsu, S., von der Malsburg, C. (2000). Analysis and synthesis of pose variations of human faces by a linear pcmap model and its application for pose-invariant face recognition system. Int'l Conf. on Automatic Face and Gesture Recognition, 142–149.
33. Salah, A.A., Nar, H., Akarun, L., Sankur, B. (2007). Robust facial landmarking for registration. *Annals of Telecommunications* **62**(1–2), 1608–1633.
34. Pietikainen, M., Hadid, A., Zhao, G., Ahonen, T. (2011). *Computer Vision Using Local Binary Patterns.* Springer.
35. Viola, P., Jones, M.J. (2004). Robust real-time face detection. *International Journal of Computer Vision* **57**(2), 137–154.
36. Roy, A., Magimai-Doss, M., Marcel, S. (2011b). Phoneme recognition using boosted binary features. Proc. IEEE Int. Conf. on Acoustics, Speech and Sig. Proc.
37. Roy, A., Magimai-Doss, M., Marcel, S. (2011a). *A fast parts-based approach to speaker verification using boosted slice classifiers.* DOI: 10.1109/TIFS.2011.2166387.
38. Poh, N., Kittler, J., Bourlai, T. (2010b). *Quality-based score normalization with device qualitative information for multimodal biometric fusion. IEEE Trans. on Systems, Man, Cybernatics Part B : Systems and Humans* **40**(3), 539–554.
39. Marcel, S., McCool, C., Matěcka, P., Ahonen, T., Černocky, J., Chakraborty, S., Balasubramanian, V., Panchanathan, S., Chan, C.H., Kittler, J., Poh, N., Fauve, B., Glembek, O., Plchot, O., Jančik, Z., Larcher, A., Lévy, C., Matrouf, D., Bonatre, J.F., Lee, P.H., Hung, J.Y., Wu, S.W., Hung, Y.P., Machlica, L.J.M., Mau, S., Sanderson, C., Monzo, D., Albiol, A., Albiol, A., Nguyen, H., Li, B., Wang, Y., Niskanen, M., Turtinen, M., Nolazco-Flores, J.A., Garcia-Perera, L.P., Aceves-Lopez, R., Villegas, M., Paredes, R. (2011). *On the results of the first mobile biometry (mobio) face and speaker verification evaluation ICPR (Contests).*
40. Tena, J. (2007). 3D *Face Modelling for 2D+3D Face Recognition.* PhD thesis, University of Surrey.
41. Hjelmas, E., Low, B. (2001). Face detection: A survey. *Computer Vision and Image Understanding* **83**, 236–274.
42. Kalal, Z., Matas, J., Mikolajczyk, K. (2008). *Weighted sampling for large-scale boosting.* Proc. BMVC.
43. Viola, P., Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features.* IEEE Conference on Computer Vision and Pattern Recognition.
44. Murphy-Chutorian, E., Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(4), 607–626.
45. Villegas, M., Paredes, R. (2008). *Simultaneous learning of a discriminative projection and prototypes for nearestneighbor classification.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
46. Givens, G.H., Beveridge, J.R., Draper, B.A., Phillips, P.J. (2005). *Repeated measures GLMM estimation of subject-related and false positive threshold effects on human face verification performance.* Proceedings of Computer Vision and Pattern Recognition (CVPR), 40.

第11章

迈向“真实的”3D交互显示器

Jim Larimer[1]，Philip J. Bos[2]，Achintya K. Bhowmik[3]
1. 加利福尼亚州，半月湾，图像矩阵公司
2. 俄亥俄州，肯特，肯特州立大学
3. 加利福尼亚州，圣克拉拉，英特尔公司

11.1 引言

我们用所有的感官来与环境交互，但毫无疑问，视觉给我们提供的周围环境印象是最直接、最重要的。光场能产生这些视觉体验的信号。当我们在有限的小范围内，从一个有利观察的角度看向任何方向，凝视经过的有限范围的所有信息都会包含在光场里。

现在的显示器能够重构2D和立体3D（s3D）图像，但在自然环境中的有些人类可以感知的视觉信号却在显示器重建图像时丢失了。人类从光场中提取信息的能力与捕捉、重构这些信号的类似能力之间的差距，正随着摄像头和显示器取代目前在计算和视频娱乐系统使用的技术而逐渐减小。丢失的各类信号正是本节的主题。

如果试图观察一个在现代显示器呈现的模糊散焦的图像，你是无法通过调焦使其清晰呈现的。这常常令人懊恼，因为使物体聚焦所需要的信息并未存储于显示屏上重构的信号中。2D、s3D和多视点s3D显示器并不能重构所有我们通常从现实世界的光场中所采集的信号。因此，当我们观察这些图像时需要操控开环回路。

光场的所有信息中仅有小部分能被人类视觉感测到。环顾四周的时候，我们能目测超出窄小的视觉波段以外的频率信息。此外，我们还可以看到未经过采样的视觉波段以内的信息。我们对光的偏振不敏感，在没有特殊视觉辅助的情况下，我们无法区分大强度或小强度的差异、细微空间内的信息或时间的快速变化，而且我们也不是很善于发现可视光的精确频谱特性。

视觉研究文献认为，我们无法感知到的光场信息是在可见度窗口以外的信息[1]。该窗口随观察条件的变化而变化，例如，离观察表面越近，就能接收到越详细的信息，并且光照

度的变化也会影响我们对短时间间隔发生的细节信息和变化的能见度。一部分信息始终超出可见度窗口。高效的工程设计需要清楚哪些信号可被视觉捕捉，哪些在能见度窗口之外。设计中若包含不可见信息会造成资源浪费。若显示器上未包含我们可见可用的信息，则意味着此时观看的图像与肉眼能观赏的自然风光将显著不同。

艺术家使用成像技术所有的功能和特点来创造独特的视觉体验。这方面的例子包括 Ansel Adams 的照片，他通过调整图像对比度使其更富有戏剧化，还有 Joel 和 Ethan Coen 导演，Roger Deakins 摄影的电影《逃狱三王》，其通过调整图像色彩将夏季青葱的密西西比打造成了燥热干旱的场景[2]。

不是每一次调整从光场捕获并显示于屏幕上的信号都能达到理想的目的。一旦忽略关闭控制回路所需要的系统信息就会引起视觉疲劳与不适[3]。显示器上观看图像所需的信息不足会导致一类问题，例如，如果观看对象出现的空间位置与它在 s3D 和多视点 s3D 显示器上聚焦的位置不匹配，就会引起视觉不适。

在理想情况下，成像技术、摄像头和显示器能够捕获或重建人类在与视觉环境进行交互时所使用的全部信号。对于某些任务，忠实地捕获和重建这些信号是理所当然的目标。然而，视觉媒体中的艺术表达表明控制和调整这些信号也同样重要。艺术家所重视的图像保真度是他们的艺术意图的再现。而对于诸如医学成像等其他应用来说，它可能用于检测疾病。在这种情况下，为提高检测能力而进行的信号调整就是工程设计的目标。不过这些不同的意图都依赖于技术。我们需要能够捕捉或重建一套完整的人类可感知的视觉信号的技术。从重构信号中删除信息或者将其变换的选择应该是能够设计的而不是随机的，或者说还存在可以克服的技术局限。

图 11.1　视觉系统的三个方面。左上方是缪勒 - 莱尔错觉，右上方是马赫带，与空间频率有关的颜色外观现象以及色彩对比在下方。这些皆展示了视觉系统与视网膜捕获的光场信号如何相互作用。正文对这些现象有详细讨论

用肉眼从自然光场采集信息会产生扭曲、假象以及变形，尽管我们自以为在视觉上感受到了真实的外部环境。图 11.1 的左上方示出的缪勒 - 莱尔错觉（Müller - Lyer illusion），就是一个关于感知如何扭曲物理地面的真实例子。这些线具有完全相同的长度，然而看起来却长短不一。我们在照片、电影和显示器上所看到的颜色就表明，不同的物理输入信号能产生不同的感知体验。面对一个人的脸与看照片上的这张脸时，投影到视网膜上的光谱能量分布是完全不同的，尽管人们通常没有注意到这种区别。视觉研究文献中称这种现象为同色异谱；传播理论则称它为变形。那些看起来具有相同颜色但实际上有着两个不同的频谱能量分布的色彩，被称为条件等色，这常见于人为场景，自然场景中并不多见。

图 11.1 右上方的灰色条形在左边缘接壤较暗条形的地方更亮，而同时在右边缘临近浅

色条形处的颜色较深。尽管灰条的外观呈齿痕状，实际上每个灰条内的灰度是一样的。这种现象被称为马赫带，它用于增强边缘的可见度。它是视觉机制经过发展演变的一个例子，其中视觉对图像细节检测的感知准确度是增强的。这种演进与医学成像用于图像数据的转换以优化病症的检测是相同的。

图 11.1 的底部显示了人类视觉系统的另一特点。词语“Orange”和“Pink”皆以同色墨水印刷在此图的左侧和右侧，但出现在右边的两词颜色却不一样。正如光在整个图像内的任何特定位置中创建的信号会影响颜色感知一样，周围的图像也会影响颜色的感知。在这个例子中，字母的感知颜色取决于色条的间距以及黄色或蓝色条是否覆盖字母或放置在字母下方。

从表面反射的光的光谱特性取决于表面以及光源。光源会因自然光或人造光而发生变化，但大多数时候，我们都能够正确识别表面颜色。然而，在图 11.1 的下方的例子中，忽略光源的机制产生了不同于表面颜色的色觉认知。通过放大，右下方的图像就可以改变颜色的感知。随着放大倍数的增加，空间关系的尺度随之改变，字母形成不同颜色的错觉就会消失。

用于实现艺术的目标与为视觉系统添加变形、扭曲并创建和光场信号一致的感知假象的意象艺术手法，引出了视觉用途的几个问题：为什么我们从光场而不是其他地方提取信息？视觉感知有什么进化性目标？以及，我们对物质的感知与这些物质的物理基础事实是如何密切相关的？要了解当前的成像系统中光场的丢失信号如何影响我们的感知以及我们与机器的交互，并回答以上问题，可以了解一下生物视觉的进化背景。

11.2　生物视觉的起源

5 亿多年前寒武纪大爆炸后，捕食成为了生活的一部分，不久生物感觉系统便开始进化。随着视觉的进化，生物能找到食物且避免被捕食。视觉对认知起着重要作用，因为认知是一种语言，我们用它来表达思想以及我们对周围世界的理解。视觉提供我们用于周围直接环境导向的基本信息，它是我们行为过程依赖的主要输入数据。几乎所有的想法都有关联图像，椅子是一个可视化的图案，老虎是一只大猫。即使是“满意”这样的抽象概念，也可以被想象成脸上露出的笑容。视觉认知，即对图像的理解，不是照片的精神等价物；我们的视觉体验更类似柏拉图的理想和形式理念。我们看到的是人、物和行为，而不是他们投射到我们视网膜上的图像。眼见的过程是动态的且有意义的，它不是一个被动的机制。

人的视觉是以物质对象为中心的。我们使用来自于光场，通过学习、记忆，经神经信号处理的信息，来了解我们从视网膜上所感知的外部环境。视网膜上形成的图像是视觉的原始数据；这些原材料无法提供足够的信息来理解图像。要理解我们所看到的，我们要改变眼睛的位置，集中注意力将场景整理或者将其分段成为完整的物质对象。感知依赖于经验以及即时可用的数据。这从寻找和识别杂乱场景中的物体时所获的体验就可以明白。从这点看，一旦对象被认可和变得熟悉，就很容易看见它了。

图像的存在，是因为我们有一个类似于针孔摄像机的眼睛，如图11.2所示。要理解我们的眼睛如何从光场提取有用的信息以及光的物理性，首先要了解针孔摄像机。公元前几百年，墨子和亚里士多德描述了暗箱的连接方式，达·芬奇将其称为光成像[4]。几何光学的中心思想——任何表面上一点发出的光都被视为向表面之外各个方向发出的射线——就是源于针孔摄像机。寒武纪大爆发后不久便发现了针孔摄像机的原理，而像我们人类这样复杂的眼睛则进化于5亿年前[5]。

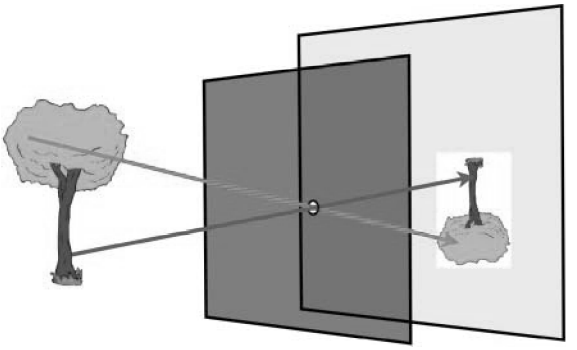


图11.2　暗箱（针孔摄像机）如该图所示。针孔摄像机的发现引发了几何光学和射线理论关于光沿直线传播的假设

Michael Faraday于1846年[6]第一次将光描述为光场，类似于他在电和磁领域提出的理论。近100年后，Gershun[7]将光场定义为3D空间中不可数的无穷点，其中每个点可被表征为一个辐射函数，这些函数取决于点在空间中的位置以及穿越每个方向的辐射（见图11.3）。

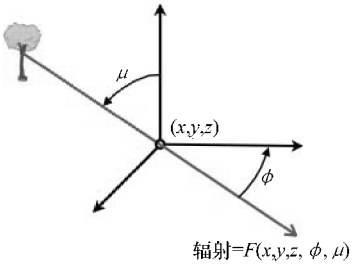


图11.3　空间中任一点通过射线显示，这些射线起源于树干，树干分别由相对于方位角ϕ和仰角μ的某一方向穿过整棵树。这个方向上，通过该点的辐射是$F(x, y, z, \phi, \mu)$。这是Gershun对光场中点的定义；Adelson和Bergen将此称为5D全光函数

光穿越光场中的点后，继续移动至点外，直到遇到阻碍，通过自然反射、折射或消失来改变它的运动轨迹。地面空间光场中穿越点的光线或者射线将在两个表面（光线或者射线的两端）终止。以这种方式定义的每条线都包含两种信息，且朝相反方向行进。如果光线不是很长，没受到阻碍，那么在沿射线的每一个点上这些信息几乎完全是多余的。射线携带的信息对于表面和产生数据包的光源来说是独一无二的，这也是生物视觉要采集和使用的信息。

Adelson和Bergen[8]将Gershun的辐射函数称为5D全光函数，表明一切从自由空间中的点可视的物质都包含在其中。他们描述了我们的视觉系统如何从光场中提取信息，以发现我们的视觉环境中物质和行为的特性。全光函数包含了有关空间中凝视的畅通无阻的表面信息，以及加入时间（此时为6D全光函数）后，这些表面如何随时间进行变化。

J. J. Gibson将我们从光场收集的信息称为动允性[9]，因为它能指导行为，例如，什么时候要躲开一个在不断靠近的物体。对于采取一个行动无用的信息就不是来自于光场信息。前面所示的马赫带就表明，我们的视觉系统有时可通过信号处理来增强信息并使某些特征（如边缘）更加突出和明显。忽略光源是动允性的一个例子。对于行为来说，更重要的是识

别表面的光反射性质，而不是感测光源如何改变那些光信号。例如，在火光和在阳光下看出食物是否可食或变质是同样重要的。图 11.1 中的虚幻色彩实际上正表明，视觉系统是如何试图阻止阴暗环境中的光源发生变化，或者阻止由照明度所带来的变化（例如，表面反射改变照射表面的光谱成分）。

从行为的效用性方面考虑，感知的物理基础事实可被看作是感知体验。进化过程使得认知得到优化，支持了生存。如果物质的真实物理属性对确保生物体和基因的存活而发生的行为至关重要，物质对象的感知就将对应于这些特性，从而使得基因能够传递给后代[10]。

针孔摄像机形成的图像取决于针孔位置处一半的全光函数。摄像机的指向决定选择哪一半。理想状态下，针孔摄像机有一个微乎其微的洞或孔，微弱的光通过这个洞或孔，形成图像。而在实践中，针孔摄像机的光圈对于光的波长来说总是足够大的，因此它包含许多理想针孔摄像机，每一个都略微偏移，又都在针孔摄像机的光圈平面内。和理想情况相比，真实的针孔摄像机能创建由许多理想针孔摄像机复合而成的图像，在这个图像里，每个投影图像相对于其他图像都略微偏移。增加光照下的偏移和组合图像如图 11.4 所示；它被表示为模糊。

在图 11.4a 中，展示了针孔摄像机领域中 3 个不同物体的表面的 3 点。箭头线表示源自这些点的光线，它们通过针孔并终止于摄像机的投影表面。在图 b 中，针孔略向右移动，投影点的相应位置，已经在向投影平面上转移。在图 c 中，展示了摄像机从其在图 a 的位置扩大到它在图 b 的位置时，投影平面上所有的理想化复合投影。现在每个表面上点的光线束都发送光线到投影表面上。关于这 3 个对象表面各点所得到的投影结果在图 c 中变得更大。这被称为模糊，它和针孔孔径的形状相同。此外，模糊重叠的这些区域，降低了重叠区域图像的对比度。由于孔径扩大，更多的光线能够进入，由此产生的针孔摄像机图像的清晰度会因模糊而逐渐降低。而模糊的量则取决于孔的大小和形状。

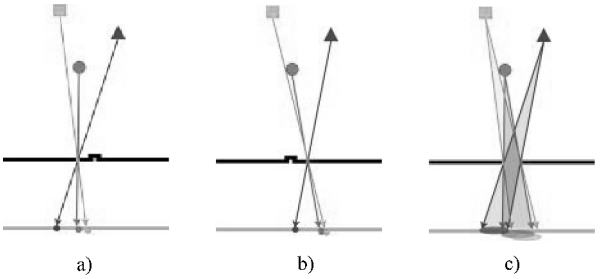


图 11.4　一个具有可变大小和位置的针孔摄像机。在图 a 上，这个小的光圈是 3 个物体被追踪到摄像机的投影面。小孔径向右移动到图 b 说明视差发生了变化，最后在图 c 上孔径由图 a 扩大到了图 b，说明模糊位置由所有的针孔图像叠加产生，这些图像由填充较大孔径的小孔填充

鹦鹉螺的针孔腔眼可以感知从图像上的圆形投影表面或视网膜形成的方向。实验室的测试已证明，鹦鹉螺的视觉系统能够感受物体在其视场的运动方向。这种对运动的行为反应证明了神经信号处理系统的演变过程，它能从一个小小的眼部光场提取动允性。该行为对鹦鹉螺逃离猎捕来说非常重要，这是理解鹦鹉螺的视觉系统图像的一个例子。

单独通过孔径的光线也可以根据它从点通过到达面所需要的时间长度来定义，光线从孔径进入到摄像机，最后在表面形成一个图像。这些距离都有些许不同，也就导致了时间上的

略微差异。该信息被称为相位。生物系统没有足够的反应速度来衡量时间的差异，但是相位对投射表面的空间影响是存在的。

随着针孔（瞳孔）孔径的增大，更多的光被成像到视网膜上，提高了灵敏度，但由于模糊降低了空间分辨率。其结果便是一种更为复杂的腔式眼部结构在寒武纪爆炸时期得以演变，该结构在孔径入口处有一个透镜，类似于今天的摄像机。一种名为“大蜗牛”（Helix）的常见蜗牛就有该结构的眼睛。具有晶状体的腔式双眼在遥远的古代就有记录，最早可以追溯到几百万年前寒武纪爆炸时期。与人眼相似的复杂的腔式眼部在寒武纪时期的前 5000 万年间进行了演变[5]。

光经过孔径进入，在孔径中放置的透镜吸收了所有从位于焦距内的物体表面各点发射的光，并把这些光聚焦于投影上的一个点。图像的细节是动允行为的关键（如觅食），所以，在生物系统中，眼睛已经形成能够改变焦点的透镜系统，目的是为了聚焦物体表面的细节。这个透镜可以有效地解决由于模糊造成的空间分辨率降低的问题。

透镜的解决方案需要一些控制透镜焦距的方法。在焦距之外的表面仍然是模糊的。焦点之外的点的聚焦位置是在投影面的前面或后面。这些光束通过摄像机或眼睛后在投影面形成的投影被分散得比较模糊。由于摄像机和眼睛上的光圈是圆的，所以在投影面形成的模糊也是模糊圈。当在这个系统中加入一个透镜后，模糊的大小取决于物体相对于投影面的距离，无论它是在镜头的焦距前面还是后面。就像模糊的针孔镜头不仅限制了空间分辨率，还因为重叠的模糊圈而导致图像对比度降低。

有一种关于生物视觉比较荒谬的说法认为，腔式眼睛对于相位是不敏感的。模糊信息就是相位差异的产物，因为光在任意时间从点到表面的折射进入孔径口的时间是不同的。生物视觉中的焦点机制利用模糊来关闭控制回路，使镜头聚焦在由注意机制决定的距离。

透镜系统中的模糊量取决于相机视野中的孔径大小和与焦点表面的相对距离。实验已证明，模糊是在视觉系统中估算已有图像不同深度的唯一线索[11]。视觉系统能够提取投影到视网膜上的由相位差生成的有用的信息，在这个有限的模糊视野状态下，该系统能够用的就是从光场中采集的相位相关信息。

随着摄像机孔径内理想针孔的位置发生变化，包含在全光函数内的空间信息也随之改变。这些变化的产生是由于视差。移动摄像机或眼睛向左或向右看往往就能绕过一个遮蔽物，因为这样做会从光场采集到一个略有不同的全光函数。然而，其中一些遮挡信息可以在镜头或眼睛的任意位置获取，因为摄像机的入口孔径或眼睛不是空间里单一的一个点。只有在聚焦面上包含视差信息的光线丢失，因为透镜将所有这些来自遮挡物和观察物的光线混合叠加一起，聚集到位于摄像机或眼睛内的投影面上。遮挡表面上的信息在投影面上依然可以获取。

图 11.5a 展示了将球形聚焦到一个单透镜摄像机的投影面上，投影面显示为图中的垂直线。射线束显示，所有的光线的集体路径，它们从球形上的一个点射出，经过摄像机的入口孔径，并集中成投影面上的一个点。在球形后的三角形被部分遮挡了。但是其表面射出的位于摄像机光学轴上的光依旧被投射到投射表面上。这些光从摄像机孔径入口的边缘进入并模

糊地显示在投射表面。

将摄像机置于绿色三角形的成焦距离点上会使来自三角形的光线聚焦。该说明对应的是图 11.5b 所示。摄像机通过改变与投影面的距离来改变焦点，如图 11.5 所示。腔式眼睛会改变透镜的焦距，使其视野内的不同焦距聚焦于眼睛的视网膜表面。这些方法基本都可以实现同一目标。现在三角形的点是在锐聚焦点，尽管它被沿着摄像机或眼睛主光轴方向的球形遮挡着。当存在这种情况时，改变焦距可以使遮挡表面可见。遮挡的表面的可见性取决于封堵器大小、遮挡物和被遮挡物表面的间隔距离、图像的对比度和入口孔径或瞳孔大小。通过改变焦点是可能看到遮挡表面后的物体的。

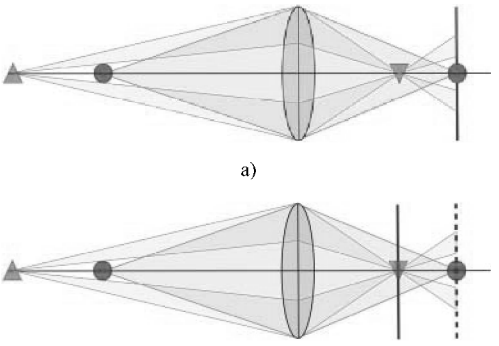


图 11.5　图 a 内，一个球形与其后的三角形通过透镜成像显示在投射表面。在图 b 内，通过改变透镜的焦距，三角形被置于焦点位置。虽然三角形被球形遮挡，它仍然能够成为焦点，因为该图中透镜入口边缘处有足够的光可以被采集。三角形的对比度因为遮挡的球形的模糊图像将降级（即衰减）。受限的改变焦点位置将能“看穿”该球形

普通摄像机无法捕获包含这些射线的角度的信息，一旦图像被拍摄，该信息将会丢失。在生物视觉系统，该信息被感测为模糊，并且可以通过改变焦点，或者通过稍微转换眼睛的方位来获取。由于旋转中心和眼的光学节点是相对彼此移动的，一个微小的移动就会产生一个平移的变化和一个新的前进方向。生物视觉已经发展到使用所有这些技术来获取光场的视差信息。

运动视差是由物体运动产生的，包括物体在摄像机或眼睛的视野内的运动，或者摄像机和眼睛自身的运动。运动视差是由视频录制下来的，在观看时能呈现非常有深度感的图像，其中的物体在视野中或摄像机中移动。前面提到的鹦鹉螺的例子表明，生物视觉系统的早期进化创造了能够从光场中提取视差信息的机制。双目或多目的视觉系统与重叠的视野也从静态图像中提取了视差信息[5]。

我们很难觉察到可以通过改变眼睛的焦距来跨越遮挡物，但这却被认为是小物体在近视野内的普遍现象。研究证明图像内的模糊信息能在感知深度时发挥作用[11]。当使用普通的静止视频摄像机的时候，光场内能够引起这类感知的视差信息就会丢失。如果没有这些信息，这些动作是无法实现的。未来的成像系统将能增加还原此类信息的能力，也因如此，有助于改进目前的视频成像技术。

如果没有一个活动进程来分析图像数据，相机是无法理解图像的。唯一可以被普通摄像机捕捉的数据可以描述为一个 2D 数组，该数组的信号值根据其在投射平面的位置进行索引。至于眼睛，图像形成所在的视网膜上遍布了光感受器，它把图像作为点状神经信号进行

编码，这些信号由大脑视觉通路内的神经元网络分析。

摄像机视野中到物体的距离可以从图像中的物体大小、受其他物体遮挡的物体，模糊、视角，或由于光线在大气的散射而导致的对比度衰减等信息获取。投射模糊的视差信息即使在静态图像中也是一个从光场中获取物体距离的有用信号。为了从这些信号中提取距离信息，图像处理过程必须能够将图像分割成独立的对象。这解释了在图像分析中，生物视觉和成像处理两者在根本上均为物体导向的过程。

我们的视觉系统已经形成了能够利用视差来估计物体距离和物体接近速率的神经机制。这些估测是基于一段只有几毫秒的从光场中采样的数据得出的。目标识别和鉴定、线性的和空中的视角、遮挡、熟悉性和其他的视网膜成像的特点，都被用于视觉系统以增强我们的图像理解力，包括对视觉领域的物体位置、大小、范围和地势等。

11.3　光场成像

光场信息透过瞳孔会取决于眼睛的焦点而或多或少被视觉感知。如上所述，聚焦将所有通过瞳孔的光线聚集到视网膜上一点，这些光线是透镜焦距内物体表面的各个点反出的。失焦的点反射的光线分散在视网膜上呈现一圈圈模糊图像，这些模糊图像的大小取决于焦点及瞳孔直径。图 11.6 说明了从两个箭头的四个端点发射出的四束光线在模型眼中的聚焦。最左边的箭头是在焦点上的，而另一个则不是。

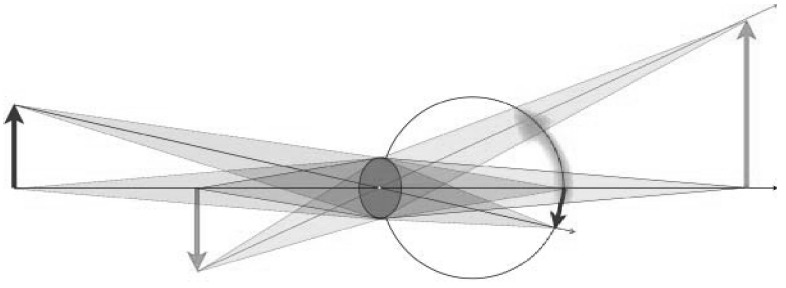


图 11.6　该图展示了模型眼聚焦蓝色箭头。透明的光束来自两个箭头的四个端点，它们形成了视网膜上的成像的端点。最左边的箭头图像在视网膜上清晰呈现，且各光束聚焦于视网膜上的一点。另一个箭头的物象则聚焦于视网膜后面的投射面上。从该箭头端点射出的光线模糊地分散在视网膜上的一大片区域。当失焦物体遮挡住聚焦的物体时，这些模糊降低了聚焦物体的图像对比度

聚焦投影表面的图像可以集合通过瞳孔的光线所传递的信息。当光线传递的信息有相关性时，（即从透镜焦距内的表面上同一点发出的光线）集合信号就得到加强。当投射到视网膜上的光线源于不同表面上的点时，信息便不具相关性，而且混杂在一起，难以辨别清楚，进而降低信号强度和图像对比度。我们从光场中捕捉的信息并未丢失，但若想获取就得聚焦于不同表面。

传统 2D 和 3D 立体图像对成像系统中图像的捕捉和重建不支持重新聚焦，并且 3D 立体

图像对系统中仅有的视差信息又受两台摄像机的位置限定。当欣赏自然风景时，我们移动双眼，传递信息至大脑，大脑快速运转从光场中获得更多的视差信息。双眼可以重新聚焦以分配并整理从光场中所有点反射进我们瞳孔内的光线。

传统图像技术无法保存聚焦信号和大量的视差信息，但我们的视觉系统却已经进化并可以利用这类信息。其结果是让人烦恼不快的。比如在观看大屏视频时，观众可能试图观察焦点之外的物体或转头环视屏幕。不管如何努力，他们都无法将一个离焦的物体置于显示的焦点，他们也无法越过遮挡物看到后面的图像。

在 2D 视频序列中移动摄像机或物体在场景中移动时，可以唤起人们对景深的感知。但是运动一旦停止，这些运动视差驱动的视觉线索就会伴随着层次感一起消失。在标准的 2D 和 s3D 图像中，无法穿越遮挡物体以及固定的图像焦点是无法获得光场数据导致的。

具有保留光场中感知信息功能的摄像机已经问世。很多人为研发这种摄像机献计献策，如 Lippmann 和 Ives [12] 以及最近的 Adelson、Wang [13] 和 Ng 等人[14]。Lytro 近期推出一款光场摄像机，Raytrix 研发的商务光场摄像机也蓄势待发[15]。理解这些摄像机捕捉光场中信息的工作原理后，就会明白建立一个真实光场显示的要求，它需要能够重塑任何头部的视差信息以及支持关注驱动的聚焦。

Lytro 和 Raytrix 的全光摄像机设计与普通摄像机设计或者眼睛的基本结构相似。为阐述全光摄像机的工作原理，这里会用到结构与眼睛相似的球形摄像机。我们将全光摄像机的光圈看作瞳孔，形成图像的投射面看作视网膜。眼睛的晶状体，称为主镜头，位于非常靠近眼睛瞳孔的地方。我们假设这与全光摄像机的结构相同，尽管这并非必要条件[16]。

全光摄像机内有一组微型针孔摄像机，这些摄像机的位置就如同视网膜在眼睛中的位置。这些针孔摄像机也可配置镜头，但不是必需。这些位于投射面的微小摄像机上能够捕捉透过瞳孔的光线投射物象。每个针孔摄像机都有唯一的光圈，位于这组摄像机内统一间隔的位置。光圈位置统一有利于确保光场信息采样的一致性，但这也不是强制要求。例如，人眼并没有一系列相同的感光器，因为我们感知的物象是视觉系统构建的，并非仅仅是视网膜上形成的短暂成像。图 11.7 阐述了两个被多倍放大的此类针孔摄像机；而该图中的其他针孔摄像机则太小而无法辨认。

每个微型摄像机都捕捉着差别细微的物象，这取决于摄像机在列阵中的位置，光线从 3D 物体空间的各个点反射进入全光摄像机的方向取决于这些点在视野中的位置。光线通过全光摄像机的光圈被投射到微型摄像机针孔列阵中的位置则取决于全光摄像机的主镜头焦距。全光摄像机并不需要可以调焦的镜头来让视野内的点清晰成像。利用全光摄像机数据重建 2D 图像，且该图像聚焦的距离与主镜头不同，这是可以通过重组微型针孔摄像机捕捉到的数据而实现的。

3D 物体中分离的点射出的两束光线进入全光摄像机光圈（与瞳孔相同的位置），一般会在不同的针孔摄像机中成像。3D 物体空间中同一点反射出的两束光线进入全光摄像机后只会被投射到微型摄像机列阵的相同针孔中。关于这点内容体现在图 11.7 中，该图追溯四束通过瞳孔的光线被投射到一系列针孔摄像机内。一道从这两个箭头顶端反射的光线穿过瞳孔

中心，这些光线在该图中标为虚线。另一道从两个箭头的末端反射的光线穿过瞳孔外围的同一位置，这些光线在此图中标为实线。

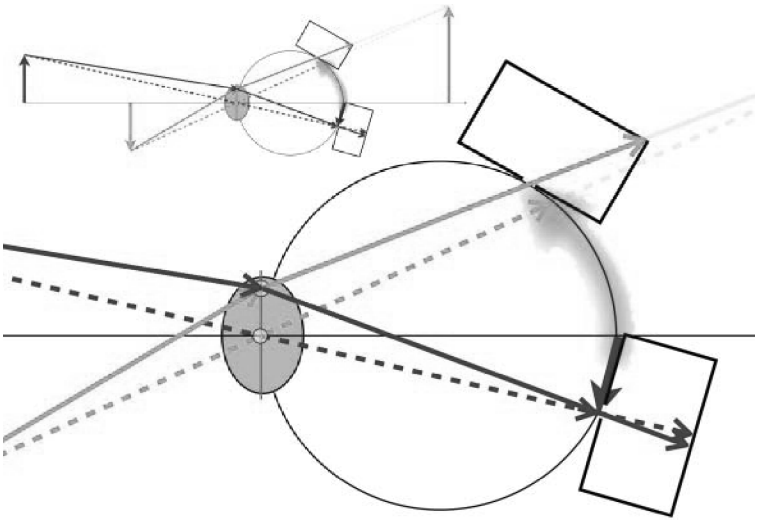


图 11.7　对左边箭头顶端反射的两条光线和右边箭头顶端反射的两条光线进行追踪，从上面的小插图发现它们最终聚焦于不同的深度。左边的箭头反射的光线聚焦于视网膜上，而右边的箭头反射的光线聚焦于视网膜的后面。放大的插图呈现了这些光线的轨迹，它们或被光圈捕捉或流失，这点可以参考位于视网膜上的两个针孔摄像机间的绿色虚线位置

最左边的箭头是聚焦的，它在主镜头的焦距内，这就使镜头焦距内摄像机视野中的点在视网膜上形成清晰物象。实线和虚线都被镜头投射到微型摄像机中针孔摄像机的同一位置。这两条光线从这一位置穿过针孔后被投射到针孔摄像机的后面。光线被投射到针孔摄像机后面的位置与光线进入光圈时的位置相关。这些位置与光线从箭头顶端传播过程中的相位差相关。传统摄像机则丢失了这些方向信息。大部分甚至全部被投射到这个针孔中的光线都是从最左端箭头的同一点发出的。投射在针孔摄像机后面的物象记录了每条光线的位置和瞳面相位。

追踪右边箭头顶端反射光线轨迹的虚线和绿色实线被投射到针孔列阵的不同位置，原因是这个箭头不在焦距内。不同的针孔摄像机会记录这两条光线的方向信息。全光摄像机捕捉到这些光线中的所有方向信息，因此不会有方向信息丢失。瞳面或光圈面早先就集合了一系列全光函数，各个函数在光圈面内彼此间有微小的位移关联。微型摄像机阵列就是模拟这些全光函数。摄像机阵列的投射面上任意位置都与一条光线或者某个全光函数的指向关联。全光摄像机可同时高效地模拟全部函数。

重新排列全光摄像机收集的数据可以重建一张不同于全光摄像机主镜头焦距的 2D 图像，该图可在焦距内任意聚焦。重建后的图片分辨率受到微型摄像机阵列和镜头分辨率的限制。能在任何深度重建 2D 图像的全光成像系统不可以是被动系统；它要求图像处理过程。这种摄像机最基本的工作原理与 Lippman 描述的系统相似[12]。

总之，场景中焦距内的点发射的光线会被阵列中特定的针孔摄像机捕捉到，焦距外的点发射的光线会被不同的针孔摄像机捕捉到。捕捉方向信息是捕捉自由空间中小范围光场的关键，比如全光摄像机入场光圈的定位。

图 11.8 中阐释了 3D 物体空间中的位于相对眼睛同一方向上一远一近两个点在全光摄像机中的表现。在图 11.7 所阐述的两条箭头的基础上，此处利用多个点进行阐述。这些图片呈现了摄像机的顶视角，最左边箭头指向摄像机视角右方，而右边箭头则指向左边。右边箭头失焦，其成像基部遮挡了最左边箭头聚焦成像的一部分基部光线。

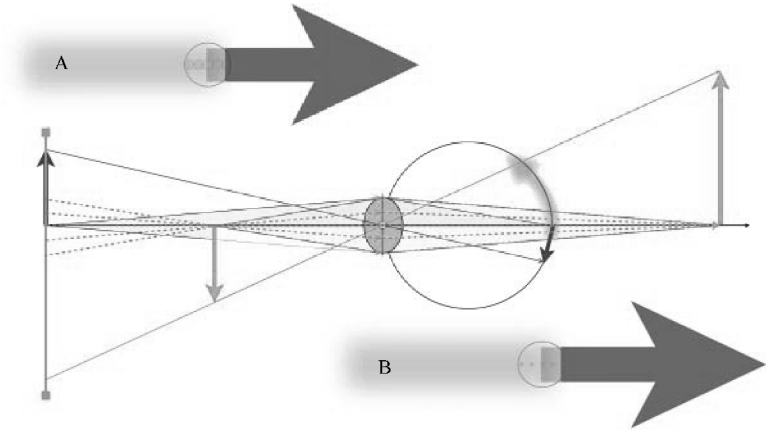


图 11.8　该图中间是一个眼睛的顶视图，眼睛聚焦于指向观察者右方的绿色箭头。这箭头部分被失焦的绿色箭头遮挡。图中左侧紫线代表着光场显示的水平横截面。绿色虚线表示显示屏发出的离散地重构包含在光场中的模糊信号的光线。假设重构模糊的光线数多到足够促使眼睛聚焦，那就意味着这个显示屏的观察者可以在重构的观察空间内以任意景深聚焦。尽管如此，重构的模糊物象看似示意图右下角 B 插图中呈现的视网膜上一系列密集的绿点，或如左边 A 插图中逐渐融为模糊物象一系列交叠的小圆圈。这里的关键在于驱动人类视线焦距控制所需的分辨率和在光场显示器中产生高图像品质所需的光线充裕度可能会有巨大不同

图 11.8 中虚线代表右边箭头基部反射的光线穿过摄像机光圈（暂时忽略从右边箭头基部延续直到竖线的那部分虚线），被投射到摄像机投射面或视网膜上微型摄像机阵列相邻的五个微型摄像机光圈中。这些微型摄像机虽未在该示意图中呈现，但却将被置于虚线横穿视网膜的中心位置。

微型摄像机的分布决定被捕捉图像的空间分辨率。该图左上角 A 插图是全光摄像机投射面上的物体成像插图。投射面上右边箭头的失焦物象基部的大圈表示该箭头基部的一个点的模糊物象。大圈内的五个小圆圈表示相邻微型摄像机阵列内的像素空间，表明了该点反射光线的方向信息。

主镜头视轴上的微型摄像机将从聚焦于摄像机投射面上的左边箭头基部反射的光线中取样 3 条光线，而且 3 条光线未在示意图中标出。左边的箭头遮蔽了右边箭头基部的反射光线，而被遮蔽的光线会穿过主镜头的左半部分。通过图像处理重新排列数据可以重建 2D 图

像，重建的图像将摄像机的焦点从左边箭头转移到右边箭头。如此，右边箭头基部的五个圆圈将组合在一起构成变为聚焦箭头的重建物象的一个像素。置于全光摄像机视轴上微型摄像机捕捉的左边箭头基部反射的 3 条取样光线将在重新聚焦和重构的 2D 图像中被分散变得模糊不清。

现在换种方式分析图 11.8 中的示意图。假设最左边的垂线是从光场中显示屏上方观察到的线。这样分析示意图，形似眼睛的摄像机不再是摄像机，而是真正的眼睛。垂线表示的显示屏由许多投影仪组成，每个显示屏上的点都在其位置上重构全光函数。在观察者可以窥见的重塑的光场内，观察空间位于显示屏的右边，即示意图中眼睛所在的位置。显示屏必须可以重建存在于观察空间及屏幕后面的有限空间内的全光函数，所以当眼睛位于观察空间内任意一点时，显示屏都能模拟重建那些类似虚拟的全光函数。

如今对于分辨率至少有两种要求：第一，为了优质的图像，显示元间距要很合理才能重建可接受的空间频率范围。与空间细节有关的眼睛能见度窗口受到如下限制：明亮度、对比度和空间频率[17]。如今对于高质量的显示屏，其设计在近观察点可生成近 30 线对空间信息。对于人们拿着的手机，其显示屏可以距眼睛很近，这就意味着每英寸超过 200 像素的点距。

在多数观察条件下，图片重组信号中几乎没有足够的对比度要求重建图像中更多的空间信息。如果能见度窗口的空间分辨率被突破，超过此限制的额外空间分辨率便可提高图片质量。增加的细节可以产生改善色阶的空间抖动效果，通过一位或多位的灰度信息重建图片。通过至少 3 个量级范围调节图片元素强度的能力和构建纯正黑色的能力都是主观图像质量的决定因素。

光场中含有使神经系统完成聚焦回路的信号，因此这些信号一定也会被重组。我们并没有充足的理由做出如下假设：提供充足的聚焦信号是为实现令人满意的图片质量提供充足信号。

这里对图 11.8 做另一种分析，显示屏前面有一个向右的箭头，屏幕上有一个向左的箭头。虚线表示光场内显示屏发出后进入观看显示屏的眼睛中的光线。仅针对此例中显示屏的水平方向而言，5 条光场图片元素投射光线经过代表右边箭头基部的一个点，之后穿过眼睛瞳孔在视网膜上成像。如果 5 条光线必须穿过特定大小直径的瞳孔以完成聚焦控制回路，那么这就要确定从观看者观看到的显示屏屏幕到观察空间内焦点间的最远距离。在屏幕后方同样存在一个相似的距离限制。

当眼睛聚焦于图 11.8 光场中屏幕表面上左边箭头时，重构右边箭头基部的 5 条光线一定会在观察者的视网膜上形成模糊的光圈。理论上讲，在观察者视网膜上形成的模糊光圈会有重叠并被放大，就像该图中的 A 插图。然而，如果这些光线向 B 插图中那样呈现明亮的点状，没有重叠，那么这些光线便足以驱动聚焦但却不足以确保足够好的图片质量。这些光线的光学要求和驱动聚焦所需的光线数量至今仍是未知数，因为到现在还没有进行能确定这些要求的合理实验。

图 11.8 中显示技术的另一方面是通过显示屏让人感知观察者与屏幕之间的距离。在此

例中，右边箭头遮挡了左边箭头，所以左边箭头基部投射的光线并没有全部展现出来。但假设观察者要逐渐靠近显示屏。这种情况下，将模型眼摆放在光场中显示屏前面重构的虚拟箭头和屏幕之间。在虚拟重建的光场中，人体就可以穿过显示屏前重建的虚拟物体。一旦发生这种情况，虚拟物体造成的遮挡就会消失。光场显示必定可以感知观察者相对于显示屏的位置，并据此调整重建的物象。

如今 3D 立体图像对成像显示中有一种剪辑伪影技术，在娱乐影视产业中被称为边缘伪影。在日常视觉体验中，人们能感受到物体被遮挡，比如看着房内的人走过一扇窗户。此人走到墙后面就会消失，这是最平常不过的想法。在图像对显示重构方面，尤其在娱乐影视和电影中经常要在显示屏幕上重构各种物体和任务。当这些物体和人物移动超出屏幕上下左右重建的图像对边界时，物体或人物的影像就会被不明显的遮挡物遮挡，这就会破坏画面感，让人看着很奇怪。

娱乐界综合采用 4 种方法避免这些伪影。屏幕前面和边缘的物体会被设为失焦状态以降低它们在屏幕中的显著度和屏幕上的色彩对比度。在屏幕前方，物体接近屏幕边缘时就会被晕影变迹或者空间变迹。这同样使靠近屏幕边缘的物体色彩对比度降低。在图像对中一个或两个图像中有浮动窗口或高对比度的遮挡边缘（如果剧院有的话）就能形成一种遮挡屏幕前面物体的表面。第四种方法是通过改变图像对视差来调节屏幕不同部分物体的平整度——一般将屏幕前面或边缘部位的物体显示得更平整以降低对这些伪影的显著感知。

图 11.9 展现了屏幕光场显示重构的等效剪辑伪影。在图 11.9 上图中，上指和下指的箭头被光场中显示面产生并进入观察者瞳孔的光线全部体现出来。图 11.9 下图则表明如果将图像和观察者转化到左边（请记住这是一幅顶视图）就会发生光线剪辑。如今显示屏上没有重构下指箭头顶端的空间，它就这样从视野中消失了。这基本上就是在当今的 3D 图像对重构中，边缘伪影的大体情况。当今娱乐视频行业是否使用相似的方案来减轻这种因素是未来研究与开发的方向。

图像是在视频通信系统中被计算机图像系统捕捉或重建的，在该系统中人们有可能控制物体在显示屏上的尺寸大小和视角。他们只能揣测这些参数会对光场中显示屏上重建的物象外观产生怎样的影响。如见的 3D 图像对成像技术的另一种独特现象是纸板效应。从体育馆顶部利用广角远摄镜头观看，从 3D 图像对显示屏上观看足球比赛使得观看者犹如扁平的微型人。造成这种感知现象的原因现在已成为一个研究话题。或许是因为不同宽景深和强聚焦的场景之间夸张的差异所造成的缩放比例问题，但是这种感知伪影产生的确切原因还待定。不管怎样，随着光场显示和光场捕捉技术成为视频通信系统工具集的一部分，类似的问题还会发生。

适用于 2D 图片展示的同一时空分辨率要求同样适用于光场中的显示屏。从相距半米的距离观看，100dpi 的显示屏上每度可视角产生约 15 线对，这足以完成对于此观看距离的展示任务。对于可以近距离观看到的手持式显示器，200dpi 或者更高的清晰度才是合理的。这与现在显示屏的时间分辨率要求一样。避免或控制如闪烁、抖动运动模糊等瞬间伪影和最近记录的 3D 图像对成像中的时间伪影[18]，都得考虑确定任何具体任务要求的帧频，尤其当

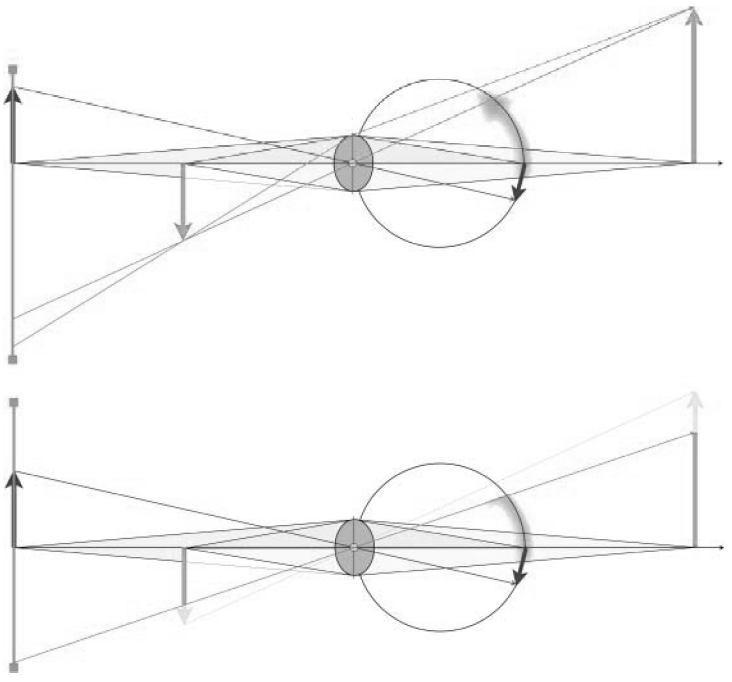


图 11.9　该图上图表明光场中显示屏（最左边的垂线）的侧视图，重构了聚焦的上指箭头和遮蔽该箭头一部分的失焦的下指箭头。屏幕下部发出的光线对于呈现失焦的下指箭头模糊的顶端是必要的。如果该图像和观看者一起向该屏幕下方移动，那么在光场中就会发生更加常见的边缘伪影，正如在 3D 图像对展示中，那个本该遮挡下指箭头顶端的物体，即那个理应在观察者与箭头之间的物体正逐渐消失，这对于观察者来说很奇怪

交错显示成为这些设备重构结构体系的一部分时。

11.4　迈向“真实的”3D 视觉显示

综上所述，把显示屏看作全光函数生成器面板，就可以理解真实的 3D 立体显示的光学成像系统的必要条件了。形象点说，就是把显示系统看作一扇窗户。想象把一扇窗户分割成很多小块——小到当我们堵住窗户只留其中一小块时，我们只能透过这一小块空窗看到色彩和亮度，空窗太小根本无法窥见图像细节。综合前面各节的内容描述，这里的窗户可以视为针孔阵列，从每个针孔中射入的光线是位于针孔上的点的全光函数的一半。在我们描述显示系统时，每块小窗户就对应一个像素。例如，图 11.10 显示光线从小块窗户的近中心位置穿过。

穿过这小块窗户的光线色彩及亮度取决于光线的角度。从图 11.10 光线进入观看者眼睛的角度分析，观察者看到墙壁的颜色，但是换个角度他就可以看到屋顶或者一块窗户的颜色。因此，透过整个窗户的每一小块的是很多光线，这些光线因角度、色彩和亮度的不同而

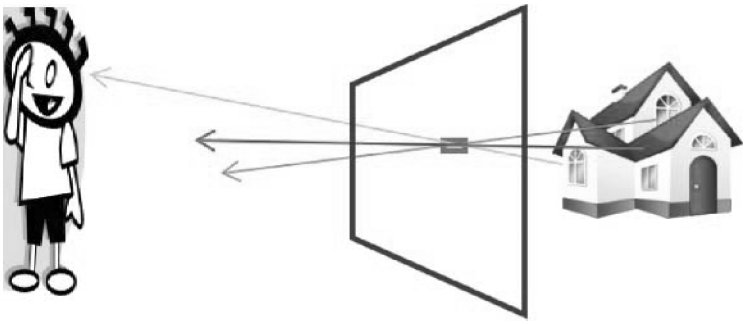


图 11.10　从窗户后面玩具房子的不同点发射的光线穿过窗户上某一“像素”的过程。对于站在如图所示位置的观察者来说，这个像素就是墙壁的颜色

各具特色。

把显示屏看作窗户，我们只考虑物体在显示屏后面成像的情况。如前面所示，物体在显示屏前面也可以成像。在这种情况下，显示屏就不再是呈现全光函数定义图像的面板了。如上所述，显示屏对其前面的物体成像时，要调节显示屏输出，同时要考虑观察者与显示屏的相对位置。

这个窗户类比同样也表明，在某些情况中，显示屏要够大才可以显示逼真的、栩栩如生的 3D 图像。例如，图 11.10 中玩具房子比显示屏小得多。如果屏幕后面有座真房子，屏幕大小大概与房子窗户相等，那么很明显观察者只能看到这座房子的很小一部分。

从屏幕与窗户之间的类比中可以看出，3D 立体显示和 2D 显示的区别在于像素（全光函数）信息的角度依赖性。

这种角度依赖信息揭示了 3D 场景的 3 个方面：

1）物体之间的相对运动（物体运动时眼睛会看到不同的内容）。

2）立体视觉（每个眼睛看到的内容不同）。

3）焦点（被瞳孔获取的从拦截的场景中某一点投射的光线，其角度扩散是由该点到观察者的距离决定）。

除了这些区别于 2D 显示的特点外，同样重要的另一个特点是 3D 立体显示的分辨率保持很高——接近人眼的极限分辨率——因为纹理线索在感知景深和图像的逼真度上是很重要的。假如我们将“极限”3D 立体显示屏类比成上面描述的窗户，我们需要高像素密度的显示屏以便调整从不同视角方向传来的高角度分辨率光线的颜色和亮度。如果我们眼睛接收一束旋转了零点几度的射线锥，我们就有充足的角度分辨率获得合适的焦点，很有可能我们只需要 0.1°的角度分辨率。如果想从 100°的视野里观看窗户，我们需要每个像素产生 100 万道光线，每道光线的颜色和亮度各不相同。配备这些规格的 3D 立体显示器其带宽会增加到具有同等尺寸和分辨率的 2D 显示器的 100 万倍，这远远超过现在的液晶显示技术。

由于这种信息内容问题，许多 3D 立体显示系统仅提供立体视觉线索。通过“裸眼立体显示”系统就可以达到这个效果，显示器上每个像素的颜色和亮度会随着观察角度的变化

而变化[19]。在这些系统中，信息的角度依赖性不会太高，只要能让不同的视觉内容进入人的两只眼睛即可。但是这些系统中显示器的空间分辨率就会降低。

佩戴偏光眼镜也可以提供立体视觉效果。采用该方法的系统被分类为使用“主动”偏光眼镜系统和使用“被动”偏光眼镜系统[20]。然而，“被动”偏光眼镜具有质量轻和使用主动快门以改善光线明亮度（见图11.11）的优点。

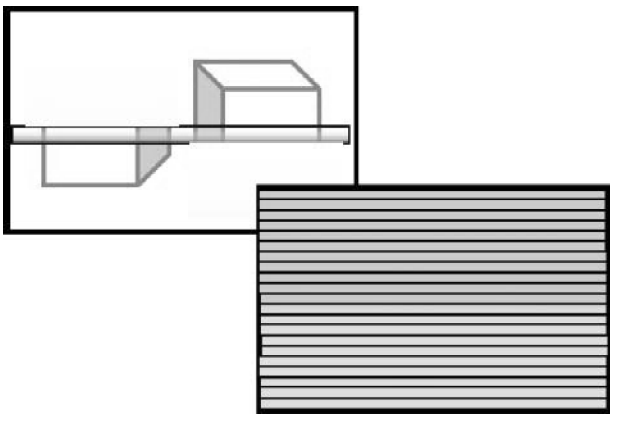


图11.11　显示屏在背景中，偏振旋转器在前景中。图像被从上到下光栅扫描，从左眼视野转到右眼视野中。偏振旋转器被设计成能够输出一种偏振态光线，可以从观看者佩戴眼镜的左镜片传输，以及另一种可以通过右镜片传输的偏振态

图11.11展示了在光场中显示屏连续显示的左眼和右眼视野图像，同时显示屏实时扫描图像落在该屏幕的下半部分，抹去了先前的右眼视野图像，并呈现了左眼视野图像。与此同时，控制光线偏振态的分档主动快门改变了从右眼传输的偏振态转至从左眼传输的偏振态的不同分档[21]。

如果显示屏距离观看者较远，那么上述的第一条和第三条线索就变得可有可无（比如在电影院中），只要有3D影像就可以显示3D图像。但是这种简便做法却不适用于距离观看者较近的场景，比如桌面显示屏或者手机显示屏上的画面。这样的话，相对运动和焦点线索都很重要，尤其是相对运动线索。相对运动效应的重要性很容易被理解，这要通过阐述如果观看者正在移动，场景中3D信息是如何看似从场景中跳出并映入观看者眼帘的。一段很棒的视频可以阐释这种效应，已经由Lee制作完成[22]。

考虑相对运动线索却不留意焦距线索，这会降低形成立体影像和平缓运动所要求的角度分辨率的要求。这样角度分辨率可能会满足一个角度要求，也可以进一步理解为仅限于满足水平方向的要求。这里3D立体显示屏的带宽仅仅为2D显示屏带宽的100倍。然而只考虑单个观看者并使用头像追踪技术的话，只要能显示两个画面的系统就可以。

zSpace已经展示过使用被动偏光眼镜和头像追踪技术的系统[23]，另外SuperD研发了自动立体显示系统[24]。这些系统能够提供相对运动和立体显示线索的效率很高。但是“真正的”3D立体显示却必须还要有焦点线索。众所周知，立体显示线索和相对运动线索对于真正的3D立体显示意义非凡，但是与之相比焦点线索的重要性却没那么明显。

早期研究焦点线索重要性的文章是由Inous[25]完成的，他指出当呈现3D图像时，眼睛的调节反应倾向于与瞳孔反应一致并聚焦于目标深度。但事实上眼睛为了看清图像会聚焦于图像源，这就与调节反应相矛盾。近期Shibata[26]发表了一篇全面论述关于这种矛盾产生的

不利之处的文章。

图 11.12 呈现了 Shibata 的“舒适区”图示。该坐标轴呈现的是屈光度，即以 m 为单位测量的焦距的倒数。我们假设显示屏距离观看者半米（聚散度距离是 2*D*），那么立体图像的舒适观看距离范围为67 ~ 40cm（1.5*D* ~ 2.5*D*），或者屏幕后面约 17cm 和屏幕前面 10cm 处。这种深度范围很有限，对于真正的互动式、浸入式 3D 立体显示的深入研发仍是重大问题。

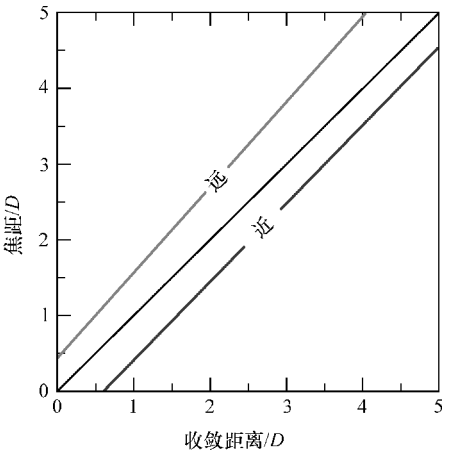


图 11.12　舒适区位于标识“远”和“近”的两条线之间，其中的聚散度距离与焦距相似。改编自 T. Shibata，J. Kim，D. Hoffman，M. Banks，2011

为解决这个问题，Kajiki[27] 和 Takaki[28] 已经研发出具有高角度分辨率足以向瞳孔呈现一些不同图像的自动立体显示系统，该系统可以产生正确的聚焦反应。图 11.13 显示 3D 物体反射出分散的光线通过眼睛合理聚焦反应聚焦于视网膜上。

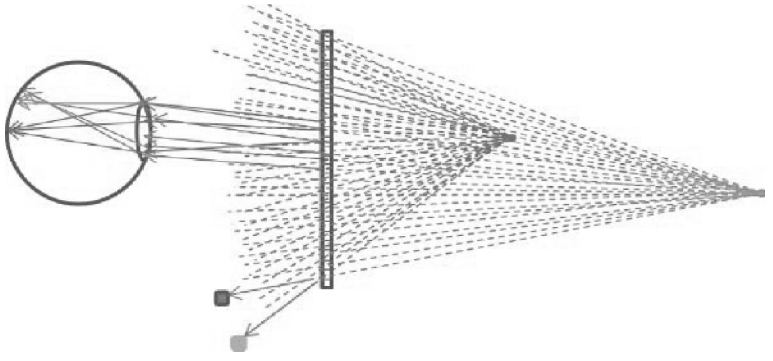


图 11.13　对高密度水平视差图的调节。该图中显示屏上呈现给观看者两个物体（两点）并获得该显示屏的顶视图，大约有 25 像素。每条虚线都对应了其中一个点发出的光线穿过显示屏的每个像素。对于最上面的模型眼，屏幕底端的像素会显示相距最远的物体颜色，然而将模型眼位置放低一点，就可以观察到相距较近的物体颜色。对于其他的模型眼摆放位置，这两点发射的光线就无法被观看底端像素的模型眼看到。距离显示屏较近的点发射出的光线与距离屏幕较远的点发射的光线相比，由于两点在显示屏后面的相对运动而更加分散。这只模型眼将距离较近的点发射的光线聚焦在视网膜上。在这种情况下，距离较远的点发射的光线就无法被聚焦了

Takaki 进一步说明瞳孔截获的清晰光线的角分辨率与眼睛正确的调节反应深度范围之间的关系[29]。上述方法有一个问题在于，光线的角度扩散限于水平方向，因而容易造成散光问题。Kim 使用光的倾斜光线提出了解决该问题的一个方法[30]。图 11.14 展示 2 条光线穿

过一个倾斜面进入瞳孔的设想。图 11.15 展示 2 条或 4 条光线进入瞳孔的效果，而焦点落在 3 个不同物体的其中之一。

使用集成成像是一个更加全面的方法，光线沿各个角度发散。近期 Xiao[31] 就对该方法进行了综述。图 11.16 显示了该系统中的聚焦效果。

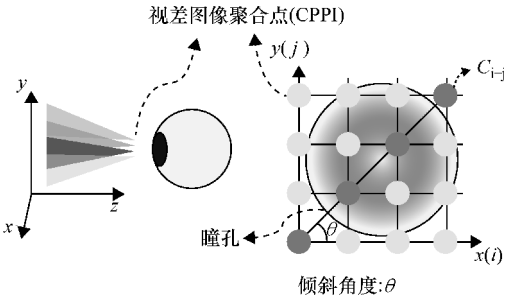


图 11.14　从同一个点发散出来的光线汇集进入瞳孔。来源：S－K Kim，S－HKim，D－W Kim 2011。经 SPIE 允许转载

我们看到以上方法包括焦点线索要求有很多光线进入每只眼睛瞳孔中，这些增加的信息使得显示屏的带宽变得非常高，正如前面所述。

因此若要限制显示屏的带宽要求，我们的窗口要更加智能化，以便将光线只传输至观看者眼中而不将光线四处发散。眼睛追踪系统可以确定观看者眼睛的位置，显示屏的每个像素到时只需要将特定颜色和亮度的光线按照一定的角度穿过像素进入观看者眼睛里（见图 11.13）。例如，如果需要 3 条光线进入双眼瞳孔中才能合理成像，那么 6 个视域系统才足够。

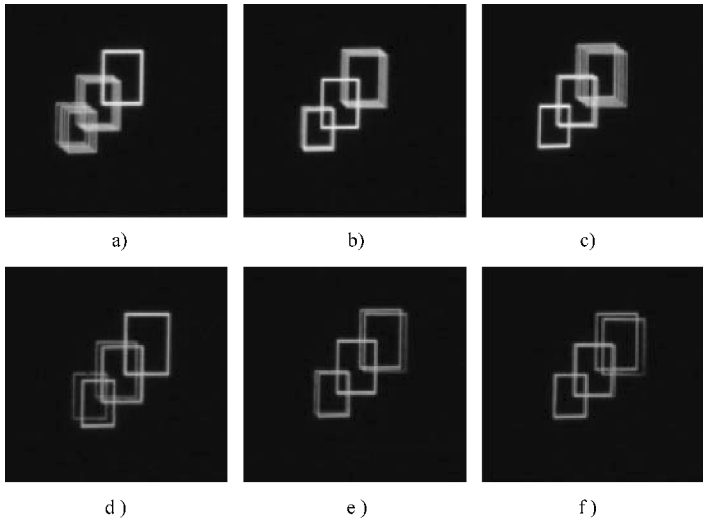


图 11.15　对比两种情况下的散焦效果，使用从相同的几个点（上一排）发散的 4 条光线和从相同的几个点（下一排）发散的 2 条光线分别对物体在 0.25m 处聚焦呈现 a）和 d），在 0.6m 处聚焦呈现 b）和 e），在 1.8m 处聚焦呈现 c）和 f）。来源：S－K Kim，S－HKim，D－W Kim 2011。经 SPIE 允许转载

使用头像追踪技术以保证进入观看者眼睛中的光线呈现高角度分辨率，同时保持相对较高的图像分辨率，这种自动立体显示系统已经由 Nakamura 等人研发出来[32]。另外有人提出了将镜头阵列摆放在显示屏前面的方法[33]。与将光线发散至每只眼睛瞳孔中的想法相关的

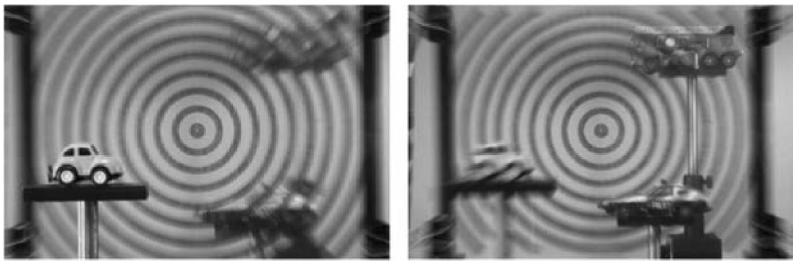


图 11.16　计算机重建集成成像系统的聚焦效果。左图中较近的小汽车是焦点，右图中较远的卡车是焦点。来源：X. Xiao，B. Javidi，M. Martinex – Corral，A. Stern 2013。经美国光学学会（OSA）允许转载

是像全息图一样提供非平面的波阵面。然而实现这种多观看者显示中存在的带宽和技术问题令使用这种显示技术的高分辨率视频设备无法运行。Reichelt 等人的著作中有一章清楚地阐明了这些问题，并提出了将全息信息只呈现给观看者观看范围内的解决方法，如图 11.17 所示[34]。

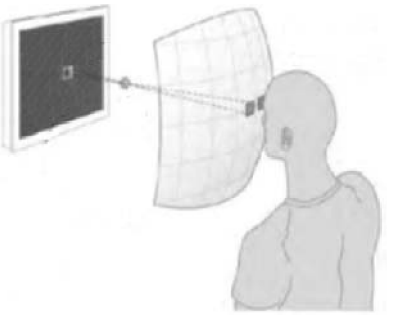


图 11.17　用于降低信息和衍射角度要求的方法，其中全息图像信息仅仅发送至观看者眼睛里。来源：S. Reichelt，R. Haussler，N. Leister，G. Futterer，H. Stolle，A. Schwerdtner 2010。经 SeeReal 技术公司允许转载

尽管上述方法从概念上讲极具吸引力，但实现满意的图像分辨率和进入眼睛中充足光线以产生焦点线索却不容易。解决调节辐辏问题的另一种方法是在显示屏与观看者之间增加一个镜头。Yanagisawa 基于具有可调节焦距镜头的显示器开发并分析出一种原型系统[35]。这些概念已经被 Shibata 进行了详细研究[36]。

这些思考集合了立体显示和分光显示的优点以解决焦点问题。立体显示的优点在于自然地刻画焦点，缺点是无法合理地处理隐藏的图像，这种系统的带宽与深度平面的数量成正比。Love 提出了一种系统，系统中平面分光显示屏前面放置了一片镜片，以便为显示的立体图像提供光场顺序聚焦平面，如图 11.18[37] 所示。光场顺序方法要求刷新率与焦点深度平面的数量相乘。但是，对于典型立体显示来说这并不是大问题，因为对于可接受的显示来说焦点平面的数量或许不会太多。由于处在真正的立体显示中眼睛能够聚焦的深度平面数量不会受限。

Bos 提出了解决调节问题的另一种方法，即使用者佩戴可固定的对焦距镜片（如双倍焦距或渐进镜片）[38]。使用多焦距镜片可以使佩戴者的眼睛焦距与 3D 物体的目标位置相协调，同时佩戴矫正镜片的使用者眼睛焦距可让图像在视网膜上聚焦。该方法实际应用于利用眼睛追踪技术测量使用者眼睛的内束以决定观看者目视深度，同时根据观看者与显示屏之间的距离来调节观看者佩戴的电子镜片的功率。

第 8 章中 Drewes 讲过眼睛凝视追踪技术的不断发展引入了低成本的眼睛追踪系统，可

以运用于这种应用中。该方法胜于其他方法之处在于不管是显示系统，还是传统的立体显示器都不需要额外带宽或者牺牲图像分辨率。具体来讲，我们可以想一下具有如上描述的头像追踪技术的“被动偏光镜”分光系统。图 11.19 体现了该系统原理[39]。

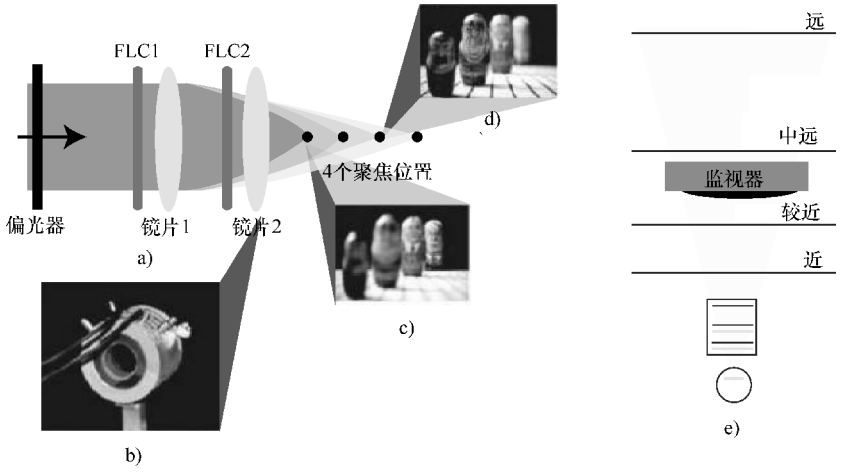


图 11.18　左边的多焦距镜片，以及右边的模型眼、镜片和显示屏概况。左边的 4 个聚焦位置对应右边的 4 个深度平面[37]。来源：G. Love, D. Hoffman, P. Hands, J. Gao, A. Kirby, M. Banks 2009。经美国光学学会允许转载

通过对矫正镜片合理调焦，我们可以使眼睛的焦度与会聚平面聚焦一致，同时可以使显示平面上的图像聚焦于视网膜上。

如上所述，观察观看者瞳孔的“内束”能使计算机找到眼睛的会聚点，并因此得出观看者与所观看的 3D 场景中物体之间的距离。结合观看者与显示屏之间的距离信息，我们就可以确定矫正镜片所需的焦度。假设眼睛镜片与平面之间的距离为“dp”，与会聚点之间的距离为“dc”，与视网膜之间的距离为“dr”，那么眼睛镜片的角度为

$$Pe = 1/dr + 1/dc \tag{11.1}$$

这样眼睛镜片就会聚焦于会聚平面。但是对于实际聚焦于视网膜上的图像，我们还需要电子透镜的焦度“Pl”，计算如下：

$$Pe + Pl = 1/dr + 1/dp \tag{11.2}$$

这就得出

$$Pl = 1/dp - 1/dc \tag{11.3}$$

图 11.19 中焦度为 - 0.5 屈光度。适用于该系统的电控透镜已出现[40]。

图 11.20 是一个直径约 1cm 的透镜图。该电控透镜调节物体焦点的功能可以从图 11.21 中看出，如图 11.21 所示，假设我们的眼睛可以调节至 50cm 的距离，那么实物与透镜之间的距离可以为 40 ~ 60cm。这就表明如果显示屏与实物之间的距离固定为 50cm，那么使用电子透镜可以使眼睛重新聚焦达到显示屏距离眼睛 40 ~ 60cm 的聚焦效果。

这种简易系统对于实现 3D 分光显示系统非常实用，可以降低调节与会聚之间的矛盾，

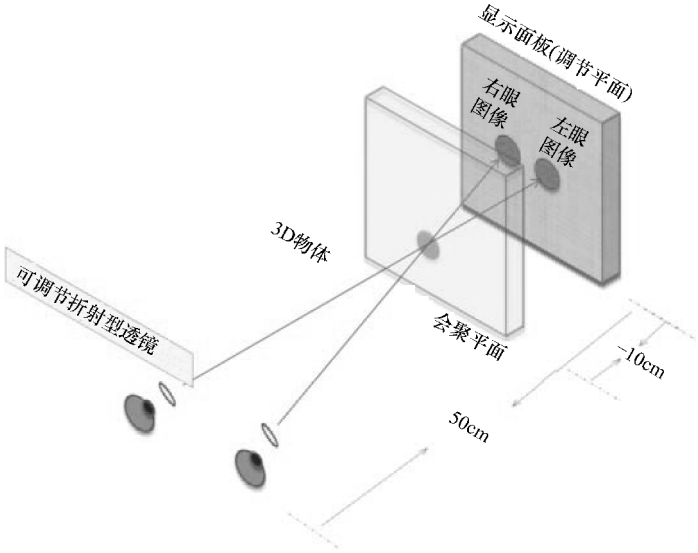


图 11.19　利用靠近眼睛的可微调透镜可以将凝视调节至会聚点，同时眼睛聚焦功能结合可微调透镜就可以使显示屏上的信息聚焦于视网膜上。来源：Source：P. J. Bos and A. K. Bhowmik 2010。经 SID/Wiley 允许转载

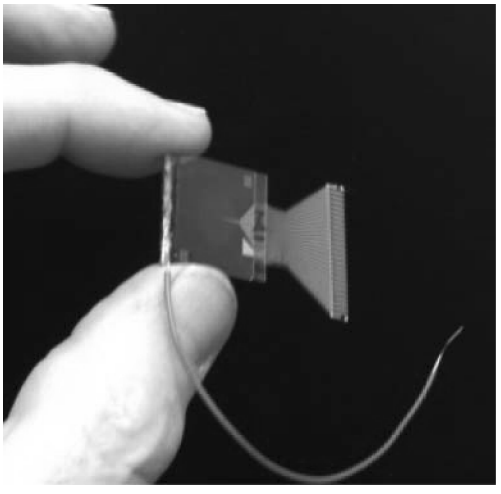


图 11.20　可微调液晶透镜与图 11.19 中使用的透镜相似

因而缓解眼疲劳。但我们要清楚这种基于矫正镜的方法不过是一种“临时办法”，这与高分辨率超级多视角、集成成像设备、See – Real 提出的全息方法或者 Love 提出的立体显示概念等更加自然的方法不同。比如，这种方法虽然解决了调节辐辏比例失调问题，但其效果是使整个显示屏聚焦在被观察图像某一特定方面的深度。人为地模糊那些没有聚焦在被观察物深度的图像可以缓和这种问题。

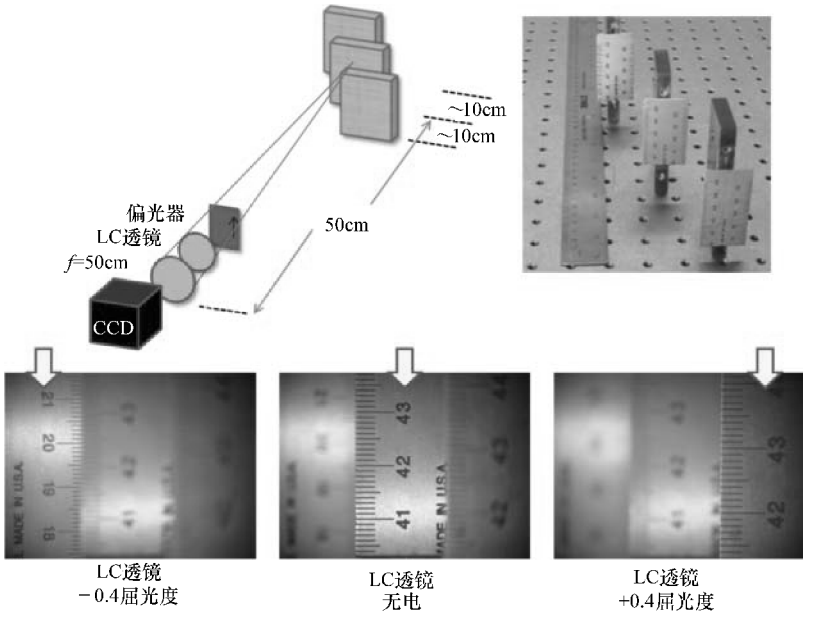


图 11.21　下图的 3 张照片是通过拍摄上图中所设的几组场景得到的，通过对可微调透镜施以 3 种不同的电压以提供透镜如图所示的不同焦度

11.5　与 3D 显示屏上的视觉内容交互

在前面各章中，我们讨论了3D 视觉原理、3D 视觉信息捕捉和向用户传达逼真视觉体验的“真实的”3D 显示器的要求。之后我们回顾了为达到这些目标而做出的技术进步。本节内容会讨论人机信息输入以及与系统中的显示内容进行的交互活动。

从 3D 显示系统的发展历史来看，这些系统都集中于一种基本应用——向用户呈现 3D 图像或者视频以产生深度感知。近年来，显示器中迅速增加了触觉感知功能，尤其是在移动通信和娱乐设备上，因此这些显示屏就成为了主要的人机交互界面。此外，实时 3D 成像技术和计算机视觉技术的进步逐渐实现了显示屏前的 3D 空间中的用户交互[41]。这些技术发展使得人们在 3D 环境中直接并直观地操控实物成为可能，引发了人机交互中革命性的变化。一系列实证研究表明，置身于 3D 空间的用户面对 3D 图像呈现时，他们的自然反应是伸出手指与图像交互[42]。由于我们在日常生活中习惯了与真实世界的触碰交互，这种反应是意料之中的。

正如在本章中前几节的讨论，“真实的”3D 显示不能仅局限于呈现所展示场景中图像的立体像对（立体显像线索），它还要提供与用户的头部和眼睛运动相一致的连续变化的图像（运动视差线索）。此外，聚焦于观察对象的双眼的会聚必须要与双眼中晶状体的焦点一致（焦点线索）。这些要求对实现真实的 3D 视觉体验很重要，但是对于交互应用程序来说

更加关键，它们促成了用户利用双手或手指等真实的身体部位去触碰 3D 空间的虚拟物象。

与真实的身体部位进行交互的 3D 空间虚拟物象——如用手指“触碰”或用双手“抓握”，仅限于对负视差虚拟出的物象，因此它们是浮现在用户和显示屏表面之间的。但是正视差虚拟出的物象由于浮现在显示板后面的虚拟空间，与其进行相似的直接 3D 交互是无法实现的，因为用户无法穿过实体显示屏表面，无法触碰受到阻挡的物象。所以在这些虚拟环境中与物象交互就要求使用虚拟出的肢体部位，比如在虚拟环境中虚拟出手，并利用对用户真实的手进行动态捕捉得到的运动对其驱动。

我们分析一下虚拟物象浮现在用户和显示屏表面之间的负视差案例。尽管负视差可以通过提供立体显像线索实现与虚拟物体的真实交互，但是由于用户与显示内容之间的距离很近，交互中愈加重要的因素是运动视差和焦点线索。当用户伸出手去抓取该空间中浮现的虚体物体时，他们的虚拟系统就得同时观察虚拟物体和真实的手。此时用户头部和眼睛是运动的，但是视网膜上对于所展示物体的成像却是固定，在这些情况中缺少运动视差线索就会引发混乱和不适，因为真实的手很明显会在视网膜上形成与用户手的运动一致的连续变化的画面。同理，缺少焦点线索也会产生非常明显的辐辏调节紊乱问题，所以不管是物体还是手指都会显示很模糊，所以两者都无法同时聚焦成像。既然我们的视觉系统会使用这些视觉线索来理解现实世界并引导我们的交互活动，那么当我们面对着 3D 显示屏上显示的内容并与之进行交互时，缺少这些线索引起的视觉冲突就会影响我们的行为。若要解决该问题就得避免将真实物体和虚拟物体同时展现在视野中，并使用虚拟出来的相同的手与虚体物体进行交互。

视觉冲突对用户行为的影响以及上述方法的有效性已经由 Bruder 等人根据费茨法则（Fitt’s Law）实验进行了评估[43]。图 11.22 对这些实验以及 Bruder 实验得出的重要结果进行了阐述。有趣的是，根据人们的直觉显示，与 3D 显示虚拟的物体使用真实的双手进行交互时是非常有效的，与之相比，使用虚拟的双手与相同环境中的虚拟物体交互时则会犯很多错误。

图 11.22　在立体显示物体的 3D 选择过程中出现的视觉冲突：a）用户聚焦于她的手指上，虚拟物体显得很模糊。b）在与真实的指尖相距固定距离的地方放置一个虚拟偏移光标（白色标识处）就可以减缓虚拟冲突。c）虚拟偏移手标识为用户选择图像提供她熟悉的和附加的尺寸和距离线索。来源：Bruder，Steinicke，Stuerzlinger 2013。转载经 IEEE 许可

除了上述的 3D 视觉线索外，利用 3D 虚拟显示技术实现实时交互应用还有其他对于人

为因素方面的要求。这些要求包括用户的快速响应，这样，用户行为和引发的系统响应之间就不会有可感知的延迟或滞后，后者将影响视觉内容上的变化。

La Viola 已经研究并论述了由于系统滞后造成的用户行为和视觉反应之间的不一致反应[44]。Stuerzlinger 等人提出限制用户在虚拟环境中的交互自由度以减少失败或不协调的体验[45]。另外，虚拟显示器和人机信息输入的实时图像捕捉设备的视野也是确保良好用户体验的关键因素。例如，用户与配备了 3D 姿势输入功能的 3D 显示内容进行交互，如果因为 3D 成像设备的视角有限而将用户的交互活动限制在小范围内，这将会是一次令人失望的体验。

针对系统和设备的交互功能而设计的用户界面融合了 3D 虚拟显示和 3D 用户信息输入技术，除了要考虑上述技术或系统的局限性，还要仔细思考我们与现实世界的交互中人为因素的影响。例如，不同于传统的具有触屏感知叠加功能的 2D 显示，当触摸显示屏前面的虚拟物体时，这类交互不会产生触觉反馈。因此，用户界面设计需采用其他方法来向用户提供实时反馈，比如在具体的应用场景中合理设计的视听线索和交互的程度、范围。比如，挤压一个虚拟气球就会导致气球产生合理的变形，这是通过气球形状和颜色的变化传达出的，该气球变化应该与用手指运动施加在气球上的力是一致的。

此外，若能利用好听觉线索也可以产生逼真的交互效果，比如在挤压变形的气球时发出的吱吱声是与挤压动作力度成正比的。与之相似，当交互活动范围达到视角上限时，精心设计的视听线索可以为用户提供指导。

除了基于视觉的 3D 姿势交互之外，像第 3 章 Breen 等人将声音理解为一种输入信息以及第 8 章 Drewes 描述的眼动追踪技术和交互等多种形式的用户界面，可以使用户与 3D 显示内容之间的交互活动更具浸透性，带来更具参与性的交互体验。关于如何使用户的人机交互活动变得简单、直观的讨论详见 LaViola 等人撰写的多模态界面章节（第 9 章）。

11.6　结语

当我们使用全部感觉与认知去体验现实世界时，视觉扮演着最重要的角色。虚拟显示器成为电子设备运行中不可或缺的元素，从手腕上佩戴的嵌入小型显示器的手表，到智能手机、平板电脑或手提电脑等中型显示器，再到电视机或信息咨询台等大型显示器。虚拟显示器近年来有了很大进展，其亮度、对比度、速度和色彩性能等视觉质量得到了明显改善。此外，能够向观看者播放具有立体显示线索的虚拟内容的立体显示器已经进入主流市场。但是，能够提供逼真的浸入式虚拟体验的“真实的”3D 显示技术还需提供运动视差和焦点线索。在本章中，我们讨论了人类视觉和 3D 立体视觉的基本原理表达“真实的”3D 立体显示要求，同时回顾了实现这些系统和发展状态的技术要求。

近年来，显示屏变得极具交互性。通过增加触屏感知层和相关的用户界面，手机显示屏已经成为移动设备中普遍应用的人机界面系统。正如触控技术章节的描述，基于触碰的用户界面继续被快速应用于广泛的设备和系统中。接下来的章节中论述了基于视听的人机界面的

发展。语音识别算法、3D 立体成像和交互以及眼动追踪技术的进步为实现用户与虚拟显示内容的交互发生革命性变化奠定了基础。将这些新的人机界面和交互技术添加到可以呈现“真实的”3D 立体虚拟内容的显示器中，并配以设计合理的多模态用户界面体系，这样就有望在显示屏前的 3D 立体空间中为用户带来逼真的交互体验。

很明显“交互式显示”时代已经到来。在接下来的几年至几十年，我们期望看到交互式显示进一步发展并应用于更广泛的设备和系统中。这的确是需要深入探究发展的新领域，但该领域却拥有带来令人振奋的新型交互应用和用户体验的巨大潜力。

参 考 文 献

1. Watson, A.B., Ahumada, A.J., Farrell, J.E. (1986). Window of visibility: psychophysical theory of fidelity in time-sampled visual motion displays. *J. Opt. Soc. Am* **3**, 300–307.
2. Adams, A. (1948). *Camera and Lens: The Creative Approach*. ISBN 0-8212-0716-4.
   Adams, A. (1950). *The Print: Contact Printing and Enlarging*. ISBN 0-8212-0718-0. http://en.wikipedia.org/wiki/O_Brother,_Where_Art_Thou%3F#cite_note-CGS-7.
3. Shibata, T., Kim, J., Hoffman, D.M., Banks, M.S. (2011). The zone of comfort: Predicting visual discomfort with stereo displays. *J. Vis*. **11**(8), 11, 1–29.
   Banks, M.S., Akeley, K., Hoffman, D.M., Girshick, A.R. (2008). Consequences of incorrect focus cues in stereo displays. *Information Display* **7**/8, 10–14.
4. Needham, J. (1986). *Science and Civilization in China: Volume 4, Physics and Physical Technology, Part 1, Physics*. Caves Books, Ltd, Taipei.
   Richter, J.P. (ed.) (1970). *Aristotle, Problems, Book XV*. The notebooks of Leonardo da Vinci. Dover, New York.
5. Land, M.F., Nilsson, D.-E. (2001). *Animal Eyes*. Oxford University Press. ISBN 0-19-850968-5.
6. Faraday, M. (1846). Thoughts on Ray Vibrations. *Philosophical Magazine* S.3, Vol **XXVIII**, N188.
7. Gershun, A. (1936). *The Light Field*. Moscow. Translated by Moon, P and Timoshenko G. in *Journal of Mathematics and Physics* 1939 **XVIII**, 51–151.
8. Adelson, E.H., Bergen, J.R. (1991). The plenoptic function and the elements of early vision, In Landy, M., Movshon, J.A. (eds.) *Computation Models of Visual Processing*, 3–20. MIT Press, Cambridge.
9. Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston. ISBN 0-313-23961-4.
   Gibson, J.J. (1977). The Theory of Affordances (pp. 67-82). In Shaw, R., Bransford, J. (Eds.). *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Lawrence Erlbaum, Hillsdale, NJ.
10. Dawkins, R. (2006). *The Selfish Gene: 30th Anniversary Edition*. Oxford University Press. ISBN 0-19-929114-4.
11. Held, R.T., Cooper, E.A., Banks, M.S. (2012). Blur and Disparity Are Complementary Cues to Depth. *Current Biology* **22**, 1–6.
    Held, R.T., Cooper, E.A., O'Brien, J.F., Banks, M.S. (2010). Using blur to affect perceived distance and size. *ACM Transactions on Graphics* **29**, 1–16.
12. Lippmann, G. (1908). Epreuves reversible donnant la sensation du relief. *J. de Physique* **7**, 821–825.
    Ives, H.E. (1930). Parallax panoramagrams made possible with a large diameter lens. *JOSA* **20**, 332–342.
13. Adelson, T., Wang, J.Y.A. (1992). Single lens stereo with a plenoptic camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(2), 99–106.
14. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P. (2005). *Light Field Photography with a Hand-held Plenoptic Camera*. Stanford Tech Report CTSR 2005-02.
15. www.lytro.com, www.raytrix.de
16. Xiao, X., Javidi, B., Martinez-Corral, M., Stern, A. (2013). Advances in three-dimensional integral imaging: sensing, display, and applications. *Appl. Optics* **52**(4), 546–560.
17. van Nes, F.L., Bouman, M.A. (1967). Spatial Modulation Transfer in the Human Eye. *JOSA* **57**(3), 401–406.
18. Hoffman, D.M., Darasev, V.I., Banks, M.S. (2011). Temporal presentation protocols in stereoscopic displays: Flicker visibility, perceived motion, and perceived depth. *JSID* **19**(3), 255–281.

19. Dodgson, N. (2005). Autostereoscopic 3D displays. *Computer* **31**(August).
20. Kim, J.H. (2010). Evolving Technologies for LCD Based 3-D Entertainment. *Information Display* **9**, 8.
21. Bos, P.J. (1993). *Stereo Computer Graphics and Other True 3D Technologies*, Chapter 6. McAllister D. (Ed.). Princeton University Press.
22. Lee, J. (2007). *Head Tracking for Desktop VR Displays using the Wii Remote*. http://www.youtube.com /watch?v=Jd3-eiid-Uw
23. http://zspace.com/
24. http://www.superd3d.com/
25. Inoue, T., Ohzu, H. (1997). Accomodative responses to stereoscopic three-dimentional display. *Applied Optics* **36**, 4509.
26. Shibata, T., Kim, J., Hoffman, D., Banks, M. (2011). The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision* **11**, 1.
27. Kajiki, T., Yoshikawa, H., Honda, T. (1996). Ocular Accommodation by Super Multi-View stereogram and 45-view Stereoscopic Display. *Proceedings for the third international display workshops (IDW'96)*, **2**, 489.
28. Takaki, Y. (2002). Universal Stereoscopic Display using 64 LCDs. *Proc. 2nd International Meeting of Information Display*, 289, Daegu, Korea.
29. Takaki, Y., Kikuta, K. (2006). 3D Images with Enhanced DOF produced by 128-Directional Display. *Proc. IDW '06*, 1909.
30. Kim, S.-K., Kim, S.-H., Kim, D.-W. (2011). Full parallax multifocus three-dimensional display using a slanted light source array. *Optical Engineering* **50**, 114001.
31. Xiao, X., Javidi, B., Martinex-Corral, M., Stern, A. (2013). Advances in three dimensional integral imaging: sensing, display, and applications. *Applied Optics* **52**, 546.
32. Nakamura, J., Takahashi, T., Chen, C.-W., Huang, Y.-P., Takaki, Y. (2012). Analysis of longitudinal viewing freedom of reduced-view super multi-view display and increased longitudinal viewing freedom using eye-tracking technique. *Journal of the SID* **20**, 228.
33. Hong, Q., Wu, T., Lu, R., Wu, S.-T. (2007). Reduced Aberration Tunable Focus Liquid Crystal Lenses for 3D displays. *SID Symposium Digest* **38**, 496.
34. Reichelt, S., Haussler, R., Leister, N., Futterer, G., Stolle, H., Schwerdtner, A. (2010). Holographic 3D displays – Electro-holography with the Grasp of Commercialization. In Costa, N., Cartaxo, A. (eds). *Advances in Lasers and Electro Optics*, Chapter 29. INTECH.
35. Yanagisawa, N. *et al.* (1995). A focus distance controlled 3D television. *The journal of three dimensional images* **9**, 14.
36. Shibata, T., Kawai, T., Ohta, K., Otsuki, M., Miyake, N., Yoshihara, Y., Iwasaki, T. (2005). Stereoscopic 3D display with optical correction for the reduction of the discrepancy between accommodation and convergence. *JSID* **13**, 665.
37. Love, G., Hoffman, D., Hands, P., Gao, J., Kirby, A., Banks, M. (2009). High speed switchable lens enables the development of a volumetric stereoscopic display. *Optics Express* **17**, 15716.
38. Bos, K. (1998). Reducing the accommodation and convergence difference in stereoscopic three-dimensional displays by using correction lenses. *Optical Engineering* **37**, 1078.
39. Bos, P.J., Bhowmik, A.K. (2011). Liquid-Crystal Technology Advances toward Future True 3-D Flat-Panel Displays. *Inf. Display* **27**, 6.
40. Li, L., Bryant, D., van Heugten, T., Duston, D., Bos, P. (2013). Near-diffraction limited tunable liquid crystal lens with simplified design. *Optical Engineering* **52**, 035007-1.
    Li, L., Bryant, D., van Heugten, T., Bos, P. (2013). Physical limitations and fundamental factors affecting performance of liquid crystal tunable lenses with concentric electrode rings. *Applied Optics* **52**, 1978.
    Li, L., Bryant, D., van Heugten, T., Bos, P. (2013). Near Diffraction limited and low haze electrooptical tunable liquid crystal lens with floating electrodes. *Optics Express* **21**, 8371.
41. Bhowmik, A.K. (2013). Natural and Intuitive User Interfaces with Perceptual Computing Technologies. *Inf. Display* **29**, 6.
42. Grossman, T., Wigdor, D., Balakrishnan, R. (2004). Multi-finger gestural interaction with 3D volumetric displays. *Proceedings of the 17th annual ACM symposium on User interface software and technology*, 61–70.
43. Bruder, G., Steinicke, F., Stuerzlinger, W. (2013). Effects of Visual Conflicts on 3D Selection Task Performance in Stereoscopic Display Environments. *Proceedings of IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE Press.
44. La Viola, J. (2000). A discussion of cybersickness in virtual environments. *SIGCHI Bulletin* **32**, 47–56.
45. Stuerzlinger, W., Wingrave, C.A. (2011). The Value of Constraints for 3D User Interfaces. In Brunnett, G., Coquillart, S., Welch, G. (eds). *Virtual Realities*, 203–223.

附录

缩略语

缩写	含义	缩写	含义
2D	二维	DET	检测误差权衡
3D	三维	DFP	数字条纹投射
AAM	主动外观模型	DLP	数字光处理
AD	绝对差异	DMD	数字微镜装置
ADC	模 – 数转换器	DMR	数字多点电阻
AEC	自动回音消除	DNN	深度神经网络
AFE	模拟前端	DP	动态规划
AG	防炫光	DRS	语篇表述法结构
AI	人工智能	DRT	语篇表述理论
AiO	一体式	DSI	视差空间图像
AM	声学模型	DSP	数字信号处理器
AMOLED	主动矩阵有机发光二极管	DST	色散信号技术
AMR	模拟多点触控电阻	DTW	动态时间规整
AMN	人工神经网络	DWT	数码波导触控
APR	声学脉冲识别	ECS	眼接触传感器
ASIC	专用集成电路	EEG	脑电图
ASR	自动语音识别	EER	等错误率
ASTM	美国材料试验学会	EM	电磁
ASW	自适应支持权重	EMG	肌电图
ATM	自动柜员机	EMI	电磁干扰
ATO	氧化锡锑	EMMA	可拓展多模态注释标记语言
BCI	（人脑和计算机）脑机接口	EOG	眼电图
BE	后端	EPD	电子纸显示屏
BOM	材料清单	ETRA	眼动跟踪研究和应用
CAD	计算机辅助设计	FE	前端
CAT	集群适应性训练	FOV	视场
CCD	电荷耦合器件	FPC	挠性印制电路
CERN	欧洲核子研究组织	FPGA	现场可编程门阵列
CMOS	互补金属氧化物半导体	FSM	有限状态机
COGAIN	视线交互通信	FST	有限状态转换器
CPU	中央处理器	FTIR	受抑全内反射
CRF	条件随机场	G2P	字素到音素
CRT	阴极射线管	GLMM	广义线性混合模型
CT	计算机 X 射线断层扫描	GMM	高斯混合模型

（续）

缩写	含义	缩写	含义
GPS	全球定位系统	NL	自然语言
GPU	图形处理器	NLG	自然语言生成
GSI	手势和语音基础结构	NLMS	归一化最小均方差
GUI	图形用户界面	NLU	自然语言理解
HCI	（人与电脑）人脑交互	NRE	一次性工程费用
HLDA	异方差线性判别分析	OCA	光学透明黏合剂
HMI	（人与机器）人机界面	ODM	原始设计制造商
HMM	隐马尔可夫模型	OEM	原始设备制造商
HRI	（自然人与机器人）人机交互	OGS	单镜片方案
HSL	色度－饱和度－亮度	OLED	有机发光二极管
HTER	半总错误率	OPWM	最优脉宽调制
IEEE	美国电气和电子工程师学会	OS	操作系统
IOB	由内向外开始	OWL	网络本体语言
IP	知识产权	PC	个人电脑
IPS	共面转换	PCA	主成分分析
IR	红外线	p－Cap	投射电容
ITO	铟锡氧化物	PCB	印制电路板
iVSM	插入电压传感矩阵	PDA	个人数字助理
JFA	联合因素分析	PDF	概率密度函数
LBP	局部二值模式	PET	聚对苯二甲酸
LCD	液晶显示（器）	PET	正电子发射计算机断层显像
LCDM	亮度补偿式差异性测量法	PIN	光电二极管
LDA	线性判别分析	PLP	感知线性预测分析
LDPP	学习判别投射和原型	POI	信息点终端
LED	发光二极管	POMDP	部分可观察马尔可夫决策过程
LM	语言模型	POS	销售点
LoG	高斯－拉普拉斯算子	PPI	每英寸像素
LVCSR	大型词汇连续语音识别	PSD	平面散射检测
MAGIC	鼠标和凝视输入级联	PSOLA	基音同步叠加法
MAP	最大后验概率	PWM	脉宽调制
MARS	多点触控模拟电阻感应器	QA	问题解答
MCE	最小分类错误	QDA	二次判别分析
ME	调制效率	RAM	随机存取存储器
MFCC	梅尔频率倒谱系数	RASTA	相对光谱分析
MLIR	最大似然线性回归	RBF	径向基函数
MMIE	最大交互信息估计法	RDF	资源描述框架
MMSE	最小均方误差	RDFS	资源描述框架图示
MOBIO	移动生物计量	RDP	基于可靠性的动态程序设计
MPE	最小音素错误	RFI	射频干扰
MRI	磁共振成像	RGB	红－绿－蓝
MTC	小组委员会	RL	强化学习
NAP	冗余属性投影	ROC	受试者工作特征（曲线）
NCC	归一化互相关	RRFC	反向斜铺场电容
NER	命名实体识别	RRS	丰富站点摘要
NFI	近场成像	s3D	立体像对三维
NIR	近红外线	SAD	绝对误差和
NIRS	近红外光谱学	SAW	表面声波
NIST	美国国家标准与技术研究院	SAYS	边说边滑

（续）

缩写	含义	缩写	含义
SBM	平方二进制法	TIR	全内反射
SD	平方差	TOF	飞行时间
SDK	软件开发包	TTS	文本语音合成
SDRT	分段语篇表达式理论	UI	用户界面
SID	国际信息显示学会	UID	唯一标识
SLM	统计语言模型	US – VISIT	美国访客和移民身份指示技术
SMS	短信服务	VA	虚拟助理
SNR	信噪比	VR	语音识别
SPIE	国际光学工程学会	VTLN	声道长度均值化
SPWN	正弦脉宽调制	WFST	加权有限状态传感器
SRGS	语音识别语法规范	WFT	加窗傅里叶变换
STFT	短时傅里叶变换	WIMP	人机交互的简约风格（窗口、图标、菜单、指示）
TCON	定时控制器		
TFA	全要素分析	WTA	胜者全得
TFT	薄膜晶体管		

作者简介

Achintya K.Bhowmik 博士曾是英特尔公司最大商业部门——PC部分的负责人，现为英特尔公司感知运算部门技术总监，他负责领导研发基于自然人机交互技术和视觉计算技术的下一代解决方案。这些技术包括视觉感知、语音识别、生物传感、沉浸式显示、多模态用户界面与应用等。

Achintya K.Bhowmik博士出版和发表过多本著作与多篇论文，获得过27项授权专利，他是IEEE的高级成员，同时也在加利福尼亚大学圣克鲁兹分校教授移动传感和计算机视觉课程。

如果你想写作、翻译，或者推荐优秀外版图书，都请随时联系我。
策划编辑：林桢
邮箱：linzhen_dgdz@163.com
QQ：61909973
电话：010-88379212
微信：alexlinzhen

关于本书

过往的科幻现已成真，在人工智能时代现在我们与计算机、手机和娱乐设备的互动正在经历革命性的变化，基于触摸、手势、语音和视觉的自然人机交互正在逐渐替代使用键盘、鼠标和游戏手柄等的交互。显示设备也从单纯的显示设备转变为提供更具吸引力和沉浸式体验的双向交互设备。本书将深入讲解基于触摸、手势、语音和视觉等自然人机交互领域的技术、应用和未来趋势。

本书适合从事人机交互领域工作的研究、设计、开发人员，相关专业师生，以及人工智能时代下对人机交互未来发展趋势有浓厚兴趣的人士阅读。

本书特色

- 提供了有关触控技术的明确指导，包括优点、局限性和未来的趋势。
- 涵盖了基于语音交互的语音输入、处理和识别技术的原理讲解和应用案例解读。
- 提供了新兴的基于视觉感知技术和手势、身体、面部、眼球追踪交互的详解说明。
- 讨论了多模式自然用户交互方案，直观地将触摸、语音和视觉结合在一起，实现真实感互动。
- 审视实现真正3D沉浸式显示和交互的要求和技术现状。

地址：北京市百万庄大街22号
邮政编码：100037
电话服务
服务咨询热线：010-88361066
读者购书热线：010-68326294
010-88379203
网络服务
机工官网：www.cmpbook.com
机工官博：weibo.com/cmp1952
金书网：www.golden-book.com
教育服务网：www.cmpedu.com
封面无防伪标均为盗版

机械工业出版社微信公众号　传播电类内容提升专业知识　关注电类行业动向　聚焦前沿科技

上架指导　人工智能/人机交互
ISBN 978-7-111-59782-7
策划编辑◎林 桢 / 封面设计◎ 子时文化 Zishi Culture

ISBN 978-7-111-59782-7
9 787111 597827
定价：99.00元