“十一五”国家重点图书　　计算机科学与技术学科前沿丛书
计算机科学与技术学科研究生系列教材（中文版）

三维自然手势跟踪的理论与方法

冯志全　杨 波　著

清华大学出版社

“十一五”国家重点图书　　计算机科学与技术学科研究生系列教材（中文版）

三维自然手势跟踪的理论与方法

冯志全　杨 波　著

清华大学出版社
北京

内 容 简 介

本书旨在介绍三维手势跟踪的理论和方法，集中展现有关研究的国内外最新前沿进展，重点介绍近几年作者的最新研究成果。主要内容包括：①国内外最新研究状况，重点介绍研究背景、研究目标、研究方法、研究难点、关键科学问题、理论和应用成果。②对手势跟踪中的关键问题进行建模。③几种典型的三维手势跟踪理论和方法：基于分析-合成技术的手势跟踪方法、基于手势识别的手势跟踪方法、基于 Monte Carlo 理论的手势状态估计方法。④基于贝叶斯滤波理论的手势状态估计方法。⑤基于粒子滤波跟踪器的理论和方法。⑥手势特征提取的理论和方法。⑦探讨跟踪方法的评价问题。

本书可以作为信息科学技术领域高年级本科生或研究生教材，也可以供从事人机交互方向的科研和技术人员参考。

图书在版编目（CIP）数据

三维自然手势跟踪的理论与方法/冯志全，杨波著. —北京：清华大学出版社，2013.5
（计算机科学与技术学科前沿丛书）
计算机科学与技术学科研究生系列教材（中文版）
ISBN 978-7-302-31531-5

Ⅰ. ①三…　Ⅱ. ①冯…　②杨…　Ⅲ. ①数字图像处理－研究生－教材　Ⅳ. ①TP391.41

中国版本图书馆 CIP 数据核字（2013）第 029622 号

责任编辑：张瑞庆　顾　冰
封面设计：傅瑞学
责任校对：梁　毅
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954

印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：16.5　　**字　　数**：409 千字
版　　次：2013 年 5 月第 1 版　　**印　　次**：2013 年 5 月第 1 次印刷
印　　数：1～2000
定　　价：35.00 元

产品编号：050867-01

前言

鉴于新一代智能人机交互(HCI)系统和虚拟现实(VR)系统的应用需要,自然人手的三维(3D)跟踪理论和方法研究已经成为国内外广泛关注的热点问题之一,在三维互联网、手语识别、手指鼠标、虚拟物体控制、家电遥控、Windows 命令控制、手指绘画、机器人控制等领域得到初步应用。这种成功应用的主要原因在于 HCI 技术正在从以计算机为中心逐步转移到以人为中心。然而,目前绝大多数新型 HCI 仍然以数字手套为基本输入设备,其主要优点是:算法设计简单,精度高,速度快,无需摄像设备,数据采样结果不受光线等外界条件的影响,输入数据量小,可以直接获得手在空间的三维信息和手指的运动信息。其主要缺点是:影响操作者的沉浸感,且数字手套本身价格昂贵、容易损坏。为了克服这些弊端,研究者通过在手上作标记,进行带标记的手势跟踪研究。但该方法在带来方便和简单的同时,也带来了不便和麻烦。在基于人脸、头部、手臂、人手、人眼以及整个人体的输入方式中,由于在通信和操作中的灵巧性,人手是最有效、用途最多的输入工具。手势是一种自然、直观、易于学习的人机交互手段,以人手直接作为计算机的输入方式,人机间的通信将不再需要中间媒体,用户可以简单地定义一种适当的手势来对周围的机器进行控制;手势是人与人之间的一种非口头交流形式,它包括从用手指示方向和移动物体的简单动作到能够表达感情以及允许彼此交流的复杂手势。考虑到人们拥有做手势的大量经验知识,如果人们能够把这些技能从日常的经验中转换过来并用在人机交互方面上,就可以期盼直观的、操作简便的,并且功能强大的人机接口。

实际上,手势跟踪和交互技术已经引起国内外很多著名 IT 企业的高度关注。2008 年,苹果公司就把手势交互技术引入 MacBook 笔记本电脑;微软公司已经把手势识别功能引进到 Xbox 游戏机和 Windows;Softkinetic 的 CEO Michel Tombroff 认为:"与 3D 解决方案的引领者 Virtools 的合作,对我们来说开发无标识的 3D 手势识别跟踪是必然的选择。"2010 年 1 月 8 日,GestureTek 在 CES 2010 大会上宣布其最新专利成果,屡获殊荣的基于手势移动设备的交互应用软件现在已经支持 Android 操作系统。2010 年 9 月,Intel 在旧金山信息技术峰会上展示了 Intel 的 7 大研究方向,OASIS(Object-Aware Situated Interactive System)便是其中之一,它旨在研究如何在家庭环境中使用 3D 物体识别和基于手势的互动。手势跟踪和交互技术已经在手语识别、手指鼠标、虚拟物体控制、家电遥控、Windows 命令控制、手指绘画和机器人控制等领域得到初步应用。

然而,相比之下,目前在三维手势跟踪研究方面还比较薄弱,绝大部分研究和应用集中在二维手势。我们认为,随着计算机交互设备和交互技术的不断发展,二维图形用户界面的局限性越来越明显地体现出来。

(1) 从界面的信息表示能力来看,二维图形技术的一个重要的缺点是不能用一种自然

的方法表示复杂的多维关系。

(2) 从交互方式而言,在虚拟现实等环境下很难用传统的交互方式来进行自然、和谐的表达,因为用传统方法反而大大增加了用户的交互难度,同时也加重了交互任务的整合工作。

(3) 用户界面的发展历经了批处理、命令行和图形界面三个阶段,三维人机交互界面的研究已经成为一个紧迫的研究课题。

目前,HCI 技术正在从以计算机为中心逐步转移到以人为中心,以三维手势作为交互工具的必要性越来越突出:

(1) 交互方式逐渐演化为适应人类的行为习惯,而不是计算机程序,更加强调以人为本;

(2) 使用多种媒体、多种模式进行交互;

(3) 基于多通道和多媒体的自然、高效、智能化、无障碍的 HCI 将是新一代智能 HCI 的主要发展方向;

(4) 人类自然形成的与自然界沟通的认知习惯和形式必定是人机交互的发展方向,人机交互正朝着自然和谐的人机交互技术和用户界面的方向发展,而三维人机交互是其中一个重要的研究方向;

(5) 在 VR 等三维系统中,采用二维手势进行交互是不方便的;

(6) 网络从 2D 的时代也将逐渐转变为 3D 网络时代,这是信息时代的呼唤,也是人类和世界科技发展的必然趋势;

(7) 随着三维显示技术的迅速发展,光纤网络的逐渐普及,网络带宽的不断增加,3D 网络将成为未来发展的趋势,也具有非常好的市场前景。

在 3D 网络得以应用的条件下,3D 显示技术的进一步延伸,三维物体的数字全息显示将为人们提供巨大的视觉冲击,它能够在近似真实三维空间再现原物体的三维图形,为实现真 3D 立体显示提供可行的方法,但怎样与之进行交互,尤其以手势作为自然交互工具,怎样将交互与可视化融合到 3D 网络界面,显然是学者们面临的另一个极富挑战性课题之一。

手势跟踪往往是手势交互的前提和基础,其核心目标是在普通摄像头、自然光照条件、复杂背景和普通 PC 条件下,在线、实时、鲁棒、精准地逐帧获取用户手势的三维结构及运动参数。这种研究涉及计算机图形图像处理、人机交互理论、计算机视觉、射影几何学、预测估计方法学和软计算理论等众多交叉学科,对其进行深入研究,对于深化智能 HCI 的理论和应用,尤其是在虚拟现实中的应用,对于推进相关学科的应用研究都具有重要意义。

本书主要研究三维手势跟踪的理论和方法:一是梳理国内外前沿的研究进展和动态,二是揭示在通往研究目标之路上可能面临的关键科学问题和应用实践问题。

本书得到国家自然科学基金(No. 61173079, No. 60973093, No. 61173078, No. 60773109)、山东省自然科学基金重点项目(ZR2011FZ003)以及济南大学学科建设重点项目经费(YTD1103)的资助。本书也是团队成员集体劳动的结晶,尤其感谢团队的徐涛博士、郑艳伟老师、唐好魁老师和历届研究生为本书付出的艰辛努力。在此一并表示感谢。

由于作者研究水平有限,书中错误在所难免,欢迎读者批评指正。

冯志全　杨波

2013 年 1 月

目录

第1章 绪 论

1.1 人手跟踪的意义

人机交互和信息处理是21世纪公认的4项重大技术之一，“和谐人机交互理论和智能信息处理基础研究”是国家重点基础研究发展计划(973计划)中“十五”后三年重要支持方向，在《中华人民共和国国务院国家中长期科学和技术发展规划纲要(2006—2020年)》中，把“智能感知技术”以及“虚拟现实技术”作为重点发展领域及优先发展主题。

目前，鉴于新一代智能人机交互(HCI)系统和虚拟现实(VR)系统的应用需要，自然人手的三维(3D)跟踪理论和方法研究已经成为国内外广泛关注的热点问题之一[1]。

VBI(Vision-Based Interface)是HCI中的一个核心课题，而基于计算机视觉的自然手(Natural Hand or Naked Hand)的3D跟踪又是VBI中的一个重要内容，利用人体姿势尤其是手势作为输入设备已经成为人机智能交互(Human Computer Intelligent Interaction，HCII)或感知用户界面(Perceptual User Interface，PUI)的重要组成部分。

众所周知，HCI技术正在从以计算机为中心逐步转移到以人为中心，具体表现为：

(1) 交互方式逐渐演化为适应人类的行为习惯，而不是计算机程序，更加强调以人为本；

(2) 使用多种媒体、多种模式进行交互；

(3) 基于多通道和多媒体的自然、高效、智能化、无障碍的HCI将是新一代智能HCI的主要发展方向。

然而，目前绝大多数新型HCI仍然以数字手套为基本输入设备，其主要优点是算法设计简单、精度高、速度快、无需摄像设备、数据采样结果不受光线等外界条件的影响、输入数据量小、可以直接获得手在空间的三维信息和手指的运动信息。其主要缺点是影响操作者的沉浸感，且数字手套本身价格昂贵、容易损坏。为了克服这些弊端，研究者通过在手上作标记，进行带标记的手势跟踪研究。但该方法在带来方便和简单的同时，也带来了不便和麻烦。

在基于人脸、头部、手臂、人手、人眼以及整个人体的输入方式中，由于在通信和操作中的灵巧性，人手是最有效、用途最多的输入工具[2,3]。手势是一种自然、直观、易于学习的人机交互手段，以人手直接作为计算机的输入方式，人机间的通信将不再需要中间媒体，用户可以简单地定义一种适当的手势来对周围的机器进行控制；手势是人与人之间的一种非口头交流形式，它包括从用手指示方向和移动物体的简单动作到能够表达感情以及允许彼此交流的复杂手势。考虑到人们拥有做手势的大量经验知识，如果人们能够把这些技能从日常的经验中转换过来并用在人机交互方面，就可以期盼直观的、操作简便的，并且功能强大的人机接口。随着计算机视觉技术的发展，利用普通摄像头便可以实现对手势运动信息的

非接触性捕获。对使用者来说，这是一种更加自然并符合人们自身行为习惯的交互方式。近年来，基于机器视觉的自然人手运动的跟踪已经逐步发展为人机交互领域中的热点问题，尤其是虚拟现实技术的广泛应用，推动了这一研究的迅速发展。与自然环境中的手势不尽相同，在人机交互中，操作性手势和交流性手势都可以用作操作物体和交流信息的方式。很多研究中都以手势运动作为输入来控制虚拟三维物体，这些虚拟物体可以利用计算机图形学方法生成，例如各种游戏效果和虚拟装配设备等。另外，很多粗略的人手跟踪技术中大都把静态的手势识别技术结合进来，以静态手势的识别结果作为输入命令的一部分，直接对虚拟物体进行操作。基于计算机视觉的自然人手 3D 跟踪将为现实或虚拟环境里的 HCI 提供新的模式，从而实现更直接、更自然、更和谐的人机交互，目前已经引起了国际上的高度重视。在过去二十多年中，研究者围绕着关节式物体的视觉运动分析这一课题做出了大量的研究，覆盖了运动检测、三维建模、运动估计、跟踪与识别、行为理解与语义分析等内容，涉及图像处理与分析、计算机视觉、计算机图形学、人工智能、认知心理学、模式识别和统计数学等多个研究领域。目前的研究成果还很难满足实际应用所提出的实时性、鲁棒性和准确性的要求，对相关技术的研究带来了一定的挑战，因此关节式物体的运动分析方法研究正得到越来越多研究者的关注。南加州大学、牛津大学、剑桥大学、波士顿大学、大阪大学等众多大学和研究机构已经在自然人手跟踪与交互领域取得了很多研究成果。

从应用层面上看，准确跟踪人手各关节的运动，在人手的配置空间中精确地恢复人手各个关节的运动参数是动态手势语义理解的前提，是手势识别研究的基础，也是进行基于自然手势的人机交互理论研究和应用研究的关键。

1.2 人手跟踪的研究目标

本书采用计算机视觉技术，其有效地把跟踪结果的鲁棒性和实时性较好地融合或统一起来，为基于三维自然人手的识别和交互打下基础。

图 1.1 中给出了手势跟踪的一般体系结构，它主要由图像分割、特征提取以及确定手势模型参数等几个部分所构成。图像分割的任务是把人手从背景图像中分离出来，以便于进行特征提取，这里的特征主要包括输入图像的边界、指尖、轮廓以及关节位置等。这些特征对跟踪结果的影响方式也是不同的，例如，如果把指尖作为特征进行跟踪，则跟踪算法往往在遮挡情况下具有较差的鲁棒性。系统进一步根据 2D 图像特征、跟踪的历史信息以及人手 3D 约束条件确定人手 3D 模型参数，这些参数可以是关节的 3D 位置，也可以是关节之间的夹角，这个过程的核心目标是得到人手的局部运动和全局运动。得到帧图像所对应的 3D 模型之后，就可以进行手势识别、绘制和输出，最后根据手势所代表的语义实现人机交互任务。

如果用 $C^{(k)}$ 表示第 k 帧图像中各关节的观测值，$H^{(k)}$ 表示第 k 帧手势，则人手跟踪问题可以描述为

$$H^{(k)} = \varphi(H^{(k-1)}, C^{(k)})$$

$H^{(k)}$ 可以表示为

$$H^{(k)} = (J_0^0, J_0^1, J_0^2, J_0^3, J_1^0, J_1^1, J_1^2, J_1^3, \cdots, J_4^0, J_4^1, J_4^2, J_4^3)$$

其中，$J_f^s(f=0\sim4, s=0\sim3)$ 表示当前帧中第 f 个手指的第 s 个关节的 3D 位置参数。

图 1.2 中进一步诠释了运动人手跟踪的基本问题：由时刻 $k-1$ 的 3D 手势和时刻 k 的

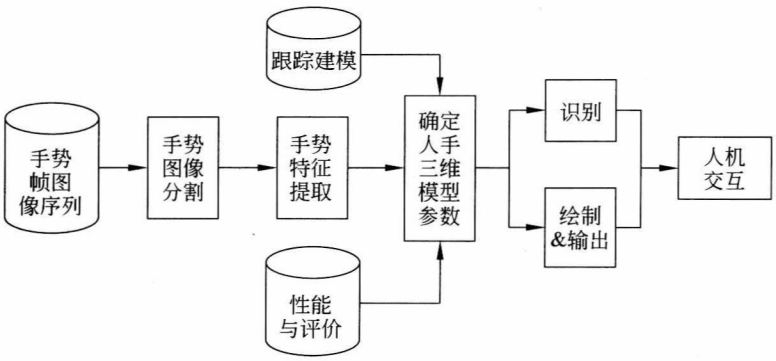

图 1.1　基于模型的手势跟踪与识别系统的一般体系结构

帧图像预测和跟踪时刻 k 的 3D 手势。

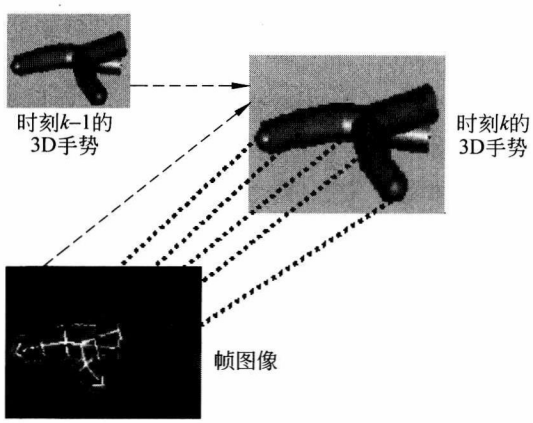

图 1.2　运动人手跟踪问题描述

从跟踪方法上看，基本上可以把运动人手跟踪方法分为两个大类：基于表观(Appearance-Based)的跟踪方法和基于模型(Model-Based)的跟踪方法。基于表观的方法也称为基于视图(View-Based)的方法，其特征在于，在学习从图像特征空间到手形状空间的映射关系后，直接根据图像得到手势，其本质是在图像特征空间和手势空间之间建立一种映射关系。具体地说，要求预先定义一组手势集，从每个手势中提取唯一的手势描述子，通过从图像特征空间到手势空间的映射直接得到手势的估计，也就是通过图像特征(点、线、角或纹理区域等)来识别手势或手势状态。该方法无须计算 3D 手势，一般需要在大量描述各种可能手势的数据集的基础上进行学习和训练。该方法主要适用于跟踪通信性手势。本书探讨基于模型的三维手势跟踪理论和方法。

1.3　人手跟踪的研究现状

1.3.1　可穿戴 HCI 系统

手势是人手的一种运动，以表达一种情感或信息[4]，早在 20 世纪 70 年代人们就开始对

其进行分析和研究。把手势作为输入设备的最初研究需要专用硬件设备的辅助，数据手套[5]就是这样一种典型的输入设备。数据手套利用机械或光学传感器把手指的运动转化为电信号，计算机通过这种电信号可以获取手的位置、手指的伸展状况等丰富信息。数据手套采用光纤技术获取人手的伸展等信息，用磁传感器技术实现位置跟踪。如 1993 年 B. Thamas 等做的自由手遥控目标的系统是凭借数据手套作为输入的媒介，但这需要实验者戴上一个专用设备。

文献[6]中给出了一种基于数字手套的以虚拟现实和增强虚拟现实为应用平台的人机交互界面，该系统提供两种类型的交互：一是菜单系统，给每个手指分配一个菜单条；二是通过跟踪具有 6 自由度的双手上的标记来实现对虚拟 3D 物体的操作引擎。文献[7]中还介绍了一种重量比较轻的输入设备，该设备仅仅由戴在食指上的一个可弯曲的传感器、一个戴在手上的加速传感器以及一个用于激活作用的微型开关组成。在文献[8]的研究中，采用了一种无线手指跟踪器。在食指上戴上超声波发射器，把接收器安装在 HMD 上，该接收器可以跟踪发生器的 3D 位置，在 400mm 的范围内分辨率可以达到 0.5mm(如图 1.3 所示)。文献[9]把电容传感器放在袖口(Wristband)上(如图 1.4 所示)，以确定手指的姿态。图 1.4 中的图(a)为传感器的外形，图(b)为传感器的原理图。当发射器被信号波(一般几百千赫)激发时，接收器可以接收这种波信号。接收信号的大小正比于发射信号的频率和电压。由于对位于手腕上的传感器的位置非常敏感，因此仅仅用于区分两种不同的手势(拳头和指向)。在 Norimichi Ukita 等的研究过程中，在 HMD 上安装有红外线装置，在红外相机上加上一个彩色相机，由分路器在两个相机中产生两个完全相同的图像[10]，因此，使用红外图像的深度信息可以从彩色图像中识别出物体。该装置可以用指尖画物体，并识别物体。

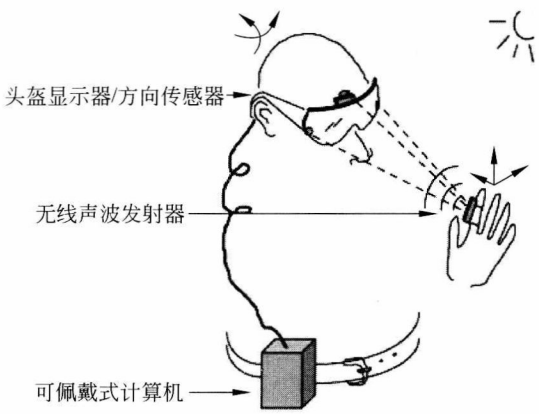

图 1.3 一种无线手指跟踪器

在文献[11]中，操作者头上安装两个摄像机，一个朝前，一个朝下指向人手，采用 HMM 对操作者的位置和行为进行估计。该系统可以感知周围环境，具有一定的智能性。Sturman 开发出用于 VR 系统的图形人机交互界面，可以用手按钮、定位器或选择设备，其速度可以达到 3～6fps[12]。Codella 实现了一个多用户 Virtual World 系统，该系统采用手势、声音、立体图像以及头部运动等多通道技术，仿真包含可形变物体的房间，这些物体可以同时由两人实时地进行创建、抓取以及抛接[13]。文献[14]中，设计了一种 Glove Talk 系统，该系统使用 5 个神经网络，定义 203 个手势，每个手势对应一个词。文献[15]中提出的

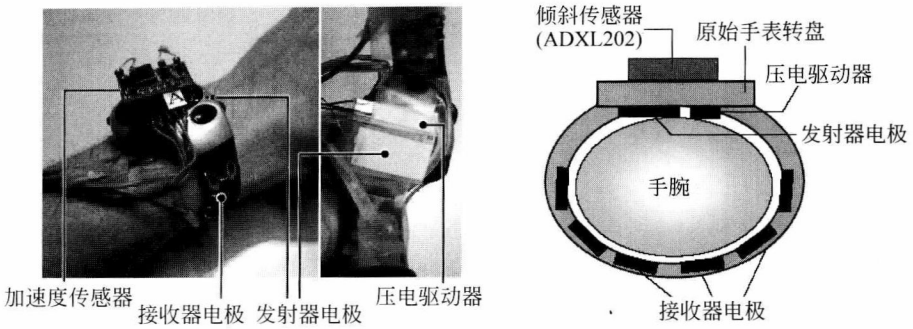

图 1.4　腕式电容传感跟踪器

Charade 实时系统，可以用手势浏览超文本演示系统，可以识别 16 个手势命令，这些命令均由三部分构成：起始手势、动态阶段和结束手势。不同手势之间的区别主要根据开始手势和动态阶段来进行。在文献[16]提出一种 VirtualPaneArchitecture(VPA)体系，它由虚拟控制盘和现实控制盘所组成，操作者可以戴上数据手套转动虚拟按钮，移动虚拟滑条，在虚拟屏幕上指示方向。Figueiredo 设计出用于虚拟现实系统中的 GIVEN 系统(Gesture-based Interaction in Virtual ENvironments)[17]，操作者可以抓取虚拟物体，围绕虚拟物体转动。GIVEN 系统还使用接触式传感器，以增强在虚拟现实环境中的真实感。事实上，可穿戴 HCI 系统已经成为多通道 HCI 研究的一个重要领域。

在国内，不少大学和科研机构很早就开始研究基于数字手套的应用。丁国富、李琪等根据人手的生理解剖模型建立人手的标准物理模型，根据数字手套返回的数据将这些数据施加于虚拟手模型，以实现人手的实时运动跟踪的目的[18,19]。

可以看到，数据手套是虚拟现实技术中广泛使用的交互设备。基于数据手套的手势识别严格来说其实不能算作一种真正的手势识别。传统的交互设备，如鼠标(笔)等其实也可以认为是一些手势输入设备。基于数据手套的手势识别目前较多采用神经网络等方法。由于神经网络可以用静态的和动态的输入，很适合用快速、交互的方式进行训练，而不必用一种解析的方式定义传递特征。还可以根据用户个人情况调整网络的连接权值，使手势识别程序能适应不同的用户。存在的不足是手势识别网络依赖于设备。当使用不同的手套设备时，要改变网络的拓扑结构，并重新训练网络得到新的连接权值。

在这种背景下，基于计算机视觉的人手跟踪和识别研究应运而生。最初，研究者通过在手上作标记，例如在手腕和手指处贴上或画上特殊颜色的圆点来识别手势。这虽然给识别带来了方便，但同样给实验者带来麻烦。最后人们终于把注意力集中到自然手上，通过专用加速硬件和脱机训练，一些研究者成功地研制了手势系统，但其识别的手势仅限几种。例如 Freeman 和 Roth 等提出的基于方向直方图的手势识别系统。图 1.5 和图 1.6 分别给出了基于数字手套和基于计算机视觉系统的硬件环境。

基于计算机视觉的运动人手跟踪系统旨在根据由摄像机获取的帧图像序列恢复人手的运动参数，例如关节角度。由于手势本身具有的多样性、多义性，以及时间和空间上的差异性等特点，加之人手是复杂变形体及视觉本身的不适定性，因此基于视觉的手势识别是一个极富挑战性的多学科交叉研究课题。更由于人手具有高自由度、指骨的尺寸比较小以及遮

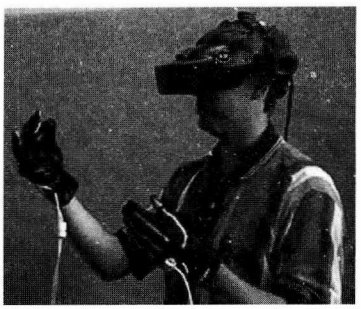

图 1.5 基于可穿戴 HCI 的人手跟踪与识别系统

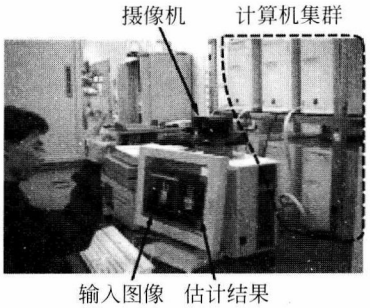

图 1.6 基于计算机视觉的人手跟踪与识别系统

挡现象，因此在研究中所面临的问题错综复杂，引起研究者的浓厚兴趣，从不同角度提出了许多技术解决手段，取得了累累硕果，比较成熟的技术手段至少包括 HMM 技术[20]、参数化 HMM 技术[21]、基于 HMM 的阈值模型[22]、基于位置分类[23]、Bayesian 网络[24]、链物体(Articulated Objects)跟踪[25]、遮挡存在条件下的多链跟踪技术[26]、人体运动的表示和识别技术[27]、人手行为理解技术[28]、基于 2D 表观技术的 3D 跟踪技术[29]、基于语言学的框架[30]、滤波技术[31-34]以及其他技术。

1.3.2 人手运动跟踪与识别

该方法也称为基于视图(View-Based)的方法，其特征在于在学习从图像特征空间到手形状空间的映射关系后，直接根据图像得到手势，其本质是在图像特征空间和手势空间之间建立一种映射关系。具体地说，要求预先定义一组手势集，从每个手势中提取唯一的手势描述子，通过从图像特征空间到手势空间的映射直接得到手势的估计，也就是通过图像特征(点、线、角或纹理区域等)来识别手势或手势状态。该方法无须计算 3D 手势，一般需要在大量描述各种可能手势的数据集的基础上进行学习和训练。该方法主要适用于跟踪通信性手势。

可视跟踪(Visual Tracking)是基于表观方法中的一种重要方法，它又可以分为基于分割的跟踪和基于视窗的跟踪(Blob Tracking)两个类。基于分割的跟踪要求跟踪目标图像的整个边界，大多采用基于图像分割、光流等基本技术设计算法。这种方法存在的主要问题是难以满足实时性要求，应用领域狭窄。基于视窗的可视跟踪则是通过矩形或者椭圆形跟踪窗体包围、锁定所要跟踪的目标图像，它又可以分为两类，即基于目标建模和定位的跟踪以及基于滤波和数据关联的跟踪。基于目标建模和定位的跟踪算法由下至上处理目标的外观变化，它一般由三个部分组成：目标的建模、相似度度量、匹配算法[35]。为了高效率地在当前帧中寻找对应于目标图像的模式，往往需要采用基于梯度的匹配算法。这类方法的关键就是获得相似度函数在当前帧中的概率密度分布。一旦得到了此概率密度分布，就可以根据其梯度的变化方向求得匹配搜索的最佳路径，从而提高匹配的效率，满足实时性要求。Mean-Shift 理论成为目标跟踪算法中的一种有力工具[36-38]，它具有输入参数少、鲁棒性强、实时性高等特点。

由于预测-刷新(Prediction-Update)方法需要初始化过程，因此 Quan Yuan 提出了一种针对 2D 人手的时变滤波(Temporal Filtering)跟踪方法[39]，它可以确定录像序列每帧中双

手的包围盒(Bounding Box)。在该方法中,首先使用基于颜色和运动残数(residue)特征的方法确定少数几个候选位置,然后用时变滤波器确定最佳位置。使用肤色后验概率在候选区域上的均值作为观测概率的估计值;用人手位置、速度等来估计转移概率。同时,提出了一种概率方法,确定人手轨迹的开始位置和结束位置。确定这样的轨迹,在发生遮挡、重叠等情况下,跟踪器可以停止继续跟踪,然后重新置位。用手工方法在每帧中标出方框以表示真实数据(Ground Truth),如果跟踪结果手势的中心位置落在该方框内,表示成功。实验结果表明,该方法比 Condensation 方法可以取得更高的成功率。

文献[40]利用双手的同步性,提出一种基于卡尔曼滤波遮挡检测和手势跟踪算法。一个所谓的画圆实验表明,在双手运动过程中,它们的速度是高度同步的,而不存在相位差别,该同步原理成为跟踪的基础。为了进行遮挡检测,用矩形包围盒将手包围起来,如果这两个矩形相交,则发生遮挡;为了防止错误检测,采用 Kalman 滤波进行预测;为了防止把左右手弄混,定义左右手的中心,并在连续图像序列中对它们进行比较。

事先准备好参考手势的图像,再用机器学习确定手势是基于表观的运动人手跟踪中的典型技术[41-45],有的用图像检索手段[45,46]求出与输入图像最相似的参考手势。例如,在文献[47]中提出基于 SMA(Specialized Mappings Architecture)的机器学习找出二维输入图像与手势之间的对应关系,他们利用 Specialized Forward Mapping Function 及一个 Feedback Matching Function 来估计输入图像的对应手势。Athitsos 和 Sclaroff 利用 chamfer distance、edge orientation histogram 以及 Hu moment[48]等几种方式,在分割出手势图像中进行层次化检索[45]。在 2003 年,他们提出以 Euclidean embedding 和 probabilistic line matching 方法从复杂背景图像中进行手势估计[46]。他们的基本思想是:首先构造一个三维的虚拟手,从包含该虚拟手的球面的离散位置进行投影,得到数据库,作为先验模型数据。当处理一个图像帧时,提取该帧的特征,再把该特征与数据库中的特征进行比较,根据相似度得到相应的三维虚拟手。他们使用 26 个手势,4128 个视点,库中共有 107 328 个图像,使用不同的相似性测量算法,例如 including chamfer distance、edge orientation histograms(边界方向柱状图)、shape moments(形状矩)和 detected finger positions(被测手指位置),把它们的带权和作为总的匹配代价。在 1.2GHz PC 上,检索每帧需要 3～4 秒(在非彩色背景下)。他们又把背景变成混乱背景,时间变为 15 秒/帧。该方法的主要优点是避免了获取数据尤其是获取对应点的困难,且总可以显示一个完整的三维手势。基于先验模型方法的主要特点是:利用 3D 模型估计手的外形;因为手的外形与视点无关,利用人手的关节模型估计运动参数,所以这些算法可以得到直接的与视点无关的识别。除此之外,Stenger 用树状结构表示库中的手势图像,以 Bayesian filtering 的方式搜索最佳解,从而加速手势估计,在 Celeron 433MHz 机器上获得的跟踪速度为 3 帧/秒(背景为黑色)[47-50]。

Shimada 提取出形状描述子,并与提前计算好了的大量模型进行匹配,这些模型是由一个 3D 手模产生的,共有 125 个手势,为每个手势存储 128 个视点,共 16 000 个形状。为避免对每帧进行穷举搜索,建立了模型形状之间的转换网络,在每个时间步保持一套假设前提,在每个假设前提的附近搜索[29]。在 6 台个人计算机组成的计算机群上实现了实时性。

在文献[51]所提出的方法中,通过颜色分割得到外形轮廓,把形状运动作为特征,把轮廓与大量的模型进行比较,这些模型是 3D 模型的投影,他们设计了 2400 个 3D 手势,86 个