基于视觉手势识别的研究—综述

任海兵,祝远新,徐光祐,林学訚,张晓平
(清华大学计算机科学与技术系,北京 100084)

摘 要: 手势是一种自然而直观的人际交流模式.基于视觉的手势识别是实现新一代人机交互所不可缺少的一项关键技术.然而,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.本文从手势建模、手势分析和手势识别等三个方面综述了基于视觉手势识别的研究现状及其应用.

关键词: 手势识别; 人机交互; 计算机视觉
中图分类号: TP242.6 **文献标识码**: A **文章编号**: 0372-2112 (2000) 02-0118-04

Vision-Based Recognition of Hand Gestures: A Survey

REN Hai-bing, ZHU Yuan-xin, XU Guang-you, LIN Xue-yin, ZHANG Xiao-ping
(*Dept. of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Hand gestures play a natural and intuitive communication mode for all human dialogs. The ability for computer to visually recognize hand gestures is essential for future human-computer interaction. However, vision-based recognition of hand gestures is an extremely challenging interdisciplinary project due to following three reasons: (1) hand gestures are rich in diversities, multi-meanings, and space-time varieties; (2) human hands are complex non-rigid objects; (3) computer vision itself is an ill-posed problem. This paper presents a survey on visual recognition of hand gestures from points of modeling, analysis and recognition techniques of vision-based hand gestures recognition.

Key words: hand gestures recognition; human-computer interaction; computer vision

1 引言

人与计算机的交互活动越来越成为人们日常生活的一个重要组成部分.特别是最近几年,随着计算机技术的迅猛发展,研究符合人际交流习惯的新颖人机交互技术变得异常活跃,也取得了可喜的进步.这些研究包括人脸识别、面部表情识别、唇读、头部运动跟踪、凝视跟踪、手势识别、以及体势识别等等.总的来说,人机交互技术已经从以计算机为中心逐步转移到以人为中心,是多种媒体、多种模式的交互技术.

基于视觉的手势识别研究正是顺应了这一潮流.然而,不同文化背景对手势的定义是有区别的.从手势识别的角度考虑,本文把手势定义为"手势是人手或者手和臂结合产生的各种姿势或动作,它包括静态手势(指姿态,单个手形)和动态手势(指动作,由一系列姿态组成)".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以及视觉本身的不适定性,因此基于视觉的手势识别是一个多学科交叉的、富有挑战性的研究课题.为了寻找突破口,必须研究人际交流中的手势用法,从而确定合理的研究范围.

一个基于视觉的手势识别系统的总体构成如图 1 所示.首先,通过一个或多个摄像机获取视频数据流.接着,系统根据手势输入的交互模型检测数据流里是否有手势出现.如果有,则把该手势从视频信号中切分出来.然后,选择手势模型进行手势分析,分析过程包括特征检测和模型参数估计.识别阶段,根据模型参数对手势进行分类并根据需要生成手势描述.最后,系统根据生成的描述去驱动具体应用.限于篇幅,本文仅从手势建模、手势分析和手势识别等三个方面介绍手势识别的研究及其应用.在文章的结尾给出一些基本结论.

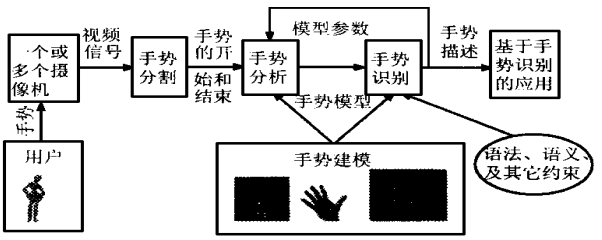

图 1 连续动态手势识别系统构成图

收稿日期:1999-06-01;修订日期:1999-09-10
基金项目:国家自然科学基金(No. 69873022)资助项目

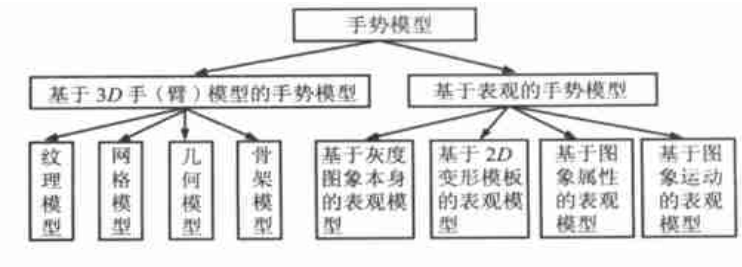

图 2　手势模型分类

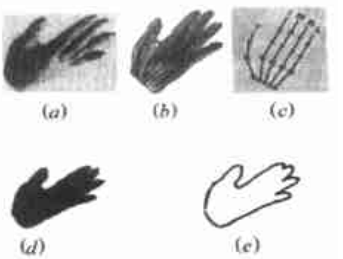

图 3　表示同一个手姿态的各种人手模型(a)有纹理的 3D 体模型(b)3D 网格模型(c)3D 骨架模型(d)二值影像(e)轮廓

2　手势建模

手势模型对于手势识别系统至关重要,特别是对确定识别范围起关键性作用,模型的选取根本上取决于具体应用. 对于某个给定的应用,一个非常简单并且粗糙的模型(例如,文献[5]使用图像梯度方向直方图去跟踪人手以及识别静态手势)可能就是充分的. 然而,如果要实现自然的人机交互,那么必须建立一个精细有效的手势模型(例如,文献[16]使用了 3D 人手模型),使得识别系统能够对用户所做的绝大多数(如果不是所有的)手势作出正确的反应(识别或拒识). 从目前的文献来看,几乎所有的手势建模方法都可以归结为两大类:基于表观的手势建模和基于 3D 模型的手势建模. 基于表观的手势模型是建立在手(臂)图像的表观之上,它通过分析手势在图像(序列)里的表观特征去给手势建模. 基于 3D 模型的手势建模方法考虑了手势产生的中间媒体(手和臂),一般遵循两步建模过程:首先给手和臂的运动以及姿态建模,然后从运动和姿态模型参数估计手势模型参数. 图 2 对这两类手势模型进行了进一步的分类. 图 3 给出表示同一种手姿态的几种模型,可以在这些人手模型的基础上进一步建立相应的手势模型.

2.1　基于 3D 模型的手势模型

基于 3D 手(臂)模型的手势模型又可以分为体模型、网格模型、几何模型以及骨架模型. 人体 3D 体模型主要用于跟踪和识别身体姿态[15],跟踪和识别的基本方法是基于合成的分析方法,简短地说就是首先合成人体的 3D 模型,然后改变模型的参数直到模型和真实人体映射出同样的视觉图像,从而分析身体姿态. 然而,即使这样的模型相当成熟,它们还是太复杂以致于不能实时地渲染. 更适合计算机实时处理的方法是使用简单的 3D 几何结构去给人体建模. 象广义锥和超二次曲面这样一类包含圆柱体、球体、椭圆体以及超矩形的结构常常用来近似身体某一部分,例如指节、前臂、或上臂[13,15]. 这些几何结构的参数相对简单一些,例如只用 3 个参数(高度、半径以及颜色)就可以完全地描述一个圆柱体. 然后,把简单的身体部分模型连结起来就可以得到象手、臂或腿等更复杂的身体部分 3D 模型. 使用手(臂)3D 模型时存在两个主要问题:其一,参数空间的维数高;其二,通过视觉技术获取这些模型的参数困难重重并且非常复杂.

最常使用的 3D 模型是 3D 骨架模型,其参数是经过简化的关节角度参数和指节长度. 人手的物理特性可以为 3D 骨架模型提供两组约束:静态约束(关节角度范围)和动态约束(运动依赖关系). 文献[16]使用了 26 个自由度的骨架模型并且利用了这样的约束条件. 文献[17]使用了带有类似约束条件的 27 个自由度的骨架模型.

2.2　基于表观的手势模型

第一类基于表观的手势模型使用 2D 灰度图像本身建立手势模型. 例如,文献[11]把人手的完整图像序列作为手势模板. 在手指跟踪应用里,仅仅手指的图像也可以用作模板[8]. 文献[4]提出运动历史图像作为手势模型. 运动历史图像是指在某个时间区间上累加图像序列里各单个像素点的运动位置而形成的 2D 图像.

第二类基于表观的手势模型建立在手(臂)的可变形 2D 模板的基础上. 可变形 2D 模板是物体轮廓上某些点的集合,一般把它用作插值节点去近似物体轮廓. 模板由平均点集合、点可变性参数,以及所谓的外部变形构成. 平均点集合描述了某一组形状的“平均”形状,点可变性参数描述了允许的形变. 通常称这两组参数为内部参数. 外部变形或者外部参数描述了一个可变形模板的全局运动,例如旋转、平移等. 基于可变形模板的人手模型通常被用于人手跟踪[7]. 文献[1]基于可变形模板跟踪人手以及进行手势识别. 最近,文[14]把 2D 可变形模板扩展成 3D 可变形模型(3D 点分布模型)用于手势跟踪.

第三类基于表观的手势模型建立在图像属性的基础上. 我们把从图像属性抽取的参数统称为图像属性参数,它们包括:轮廓、边界、图像矩、图像特征向量以及区域直方图特征等等. 例如,文献[7]使用了基于边界的轮廓特征. 由于图像矩计算简单,因此常被用作图像属性参数. 其它常被使用的属性参数还包括 Zernike 矩、方向直方图[12]、颜色直方图[5]等等.

第四类基于表观的手势模型通过计算图像运动参数,抽取手势模型参数. 这类表观模型主要用在动态手势识别里. 例如,文献[6]通过运动边界点以及方差约束计算光流,然后通过向量聚类以及运动平滑性约束抽取手势的运动轨迹,并根据轨迹坐标建立手势模型. 文献[10]通过区域相关性计算光流,然后进行光流聚类,把图像中运动区域分割成“运动块”,这些运动块分别对应于手、臂或身体其它部分等. 文献[19]提出的时空表观模型也是基于运动图像的. 跟上面这些模型不同,他们利用图像的变阶运动参数模型及鲁棒回归分析的方法去估计图像的运动参数,并同时分割出对应的运动区域. 然

后,基于图像运动参数的物理意义以及运动区域的形状特征构造帧间表观特征,最后由帧间表观特征构造手势的时空表观模型.

3 手势分析

3.1 特征检测

手势分析阶段的任务就是估计选定的手势模型的参数.分析阶段一般由特征检测和参数估计两个串行任务组成.在特征检测过程中,首先必须定位做手势的主体(人手).根据所用的线索不同,可以把定位技术分为基于颜色定位、基于运动定位、以及多模式定位等三种.绝大多数颜色定位技术依赖于直方图匹配[5]或者利用皮肤的训练数据建立查找表的方法[9].基于颜色定位技术的主要缺点是在不同的光照条件下皮肤颜色变化较大,这经常导致未被发现的皮肤区域或者误检测出非皮肤区域.利用限制性背景或者颜色手套[17],使得高效地、甚至实时地定位人手成为可能,然而对用户以及对接口设备施加了明显限制.

基于运动的定位 s 技术通常跟某些假设一起使用.例如,假设通常情况下只有一个人在做手势,并且手势者相对于背景(静止的)的运动量很小,因此,图像里的主要运动分量通常是手(臂)运动.文献[10,12]就使用了这种定位技术.为了克服利用单个线索定位的局限,基于多线索融合(即多模式)定位技术已经被提出来了.文献[2]中利用颜色、运动、和其它视觉线索的融合;另外,文献[19]基于运动和颜色信息的融合定位人手,达到了较好分割效果.

尽管不同手势模型的参数各不相同,但是用于计算模型参数的图像特征基元通常是非常相似的.常用的图像特征基元包括灰度图像[4,11]、二值影像[16,18]、区域[3,9,10,19]、边界及轮廓[6,13]或者指尖[8,17]等.

3.2 模型参数估计

3D 手模型通常涉及到两组参数:角度参数(关节角度等)和直线参数(指骨长度和手掌尺度等).从检测出的特征去估计这些运动学参数通常包括两个环节:初始参数估计和参数随时间的更新.到目前为止,所有 3D 人手模型都假设直线参数是预先已知的.这个假设把求解人手关节角度问题转化为逆运动学问题.给定 3D 终端效应器的 3D 位置和运动学链的基点,逆运动学的任务就是找出链里的指节之间的关节角度.逆运动学问题本质是病态的,允许有多个解,并且计算量大,因而不能用于实时问题.某些更简单的解决方法是让用户交互式地初始化模型参数[16].一旦估计出人手模型的初始参数,利用某种预测/平滑策略就可以更新参数估计.最常用的策略是卡尔曼滤波和预测.

如前所述,共有四类基于表观的手势模型.基于灰度图像本身的表观模型有许多不同的参数.在最简单的情况下,可以选择模型视图序列作为参数[11],也可以使用序列里各帧图像关于平均图像的特征分解表示.最近文献[4]累积图像序列里的时/空信息,从而形成单个 2D 图像,即所谓的运动历史图像.然后,基于 2D 图像描述技术(如几何矩描述或者特征分解)去参数化那些 2D 图像.基于可变形 2D 模板表观模型的典型参数是模板节点的均值 m 和它们的方差 v.通过在训练集上进行主成分分析(Primary Component Analysis,PCA)可得到模型参数.与可变形模板模型参数相联系的还有外部变形参数(指手或身体在工作区间里的旋转和平移运动).可以在类似于刚体运动估计的框架下估计模型参数的更新,所不同的是可变形模板需要估计由于模板可变性 dv 而引起的附加位移.基于图像属性表观模型的常用参数是手形几何矩、Zernike 矩、以及朝向直方图等等.这些图像特征参数易于估计,但是它们对图像中其它非手物体非常敏感.

基于运动图像表观模型的参数包括平移运动参数,旋转运动参数,以及图像变形参数等等.文献[6]通过对图像的平移运动参数进行聚类,抽取人手在图像平面的运动轨迹.文献[3]基于宽基线立体视觉跟踪人手及头部运动,然后把人手在 3D 空间的平移运动速度作为模型参数.文献[10]中使用的手势模型参数,包括图像块的平移运动以及旋转运动参数.而在文献[19]中提出的时空表观手势模型参数则更丰富,包括平移运动参数、旋转运动参数,膨胀参数、变形参数、以及方位参数等等.

4 手势识别

手势识别就是把模型参数空间里的轨迹(或点)分类到该空间里某个子集的过程.静态手势对应着模型参数空间里一个点,而动态手势则对应着模型参数空间里的一条轨迹,因此它们的识别方法有所不同.静态手势识别算法包括基于经典参数聚类技术的识别和基于非线性聚类技术的识别.

与静态手势不同,动态手势涉及时间及空间上下文.绝大多数动态手势被建模为参数空间里的一条轨迹.不同用户做手势时存在的速率差异、熟练程度会在轨迹的时间轴上引起非线性波动,如何消除这些非线性波动是动态手势识别技术必须克服的一个重要问题.考虑到对时间轴的不同处理,现有的动态手势识别技术可以分归三类:基于隐马尔可夫模型(Hidden Markov Models, HMM)的识别,基于动态时间规整(Dynamic Time Warping, DTW)的识别,基于压缩时间轴的识别.

在基于 HMM 的识别算法里,每种手势有一个 HMM.可观察符号对应着模型参数空间里的向量(点),例如几何矩向量[9,18],Zernike 矩,特征图像系数向量,或者 3D 空间的运动速度[3]等等.基于 HMM 识别技术的优点包括提供了时间尺度不变性,保持了概率框架,以及具有自动分割和分类能力.

DTW 方法是具有非线性时间规一化效果的模式匹配算法,使用某种指定属性的非线性规整函数对时间轴上的波动近似建模,通过弯曲其中一个模式的时间轴使之跟另一个模式达到最大程度的重叠(此时的残差距离最小)从而消除两个时空表示模式之间的时间差别.实际上,它是 HMM 的简化,对于比较简单的时间序列,它们二者是等价的.文献[13]基于 DTW 匹配两个运动模式.文献[11]假设两个序列的时间终点是一样的,然后利用经过修改的在时间上向后匹配的 DTW 方法进行弹性匹配.利用 DTW 算法从时间上对准了两个模式之后,利用规一化的相关运算来寻找两个模式之间的相似性.文献[19]采用的最优动态规划匹配算法也属于基于 DTW 的识别算法.DTW 方法的优点是概念上简单,也比较有效,在测试模式和参考模式之间允许充分的弹性,从而实现正确的分类.

基于压缩时间轴的识别就是首先利用某种特定属性的函数,把模型参数空间的一条轨迹压缩为单个点(例如在时间方向求和),然后利用静态手势识别算法完成动态手势的识别.文献[4]提出基于运动历史图像的动态手势识别就利用了基于压缩时间轴的识别方法.

5　结论与展望

考虑到手势本身的多样性、多义性、差异性等特点以及技术的局限,在可预见的将来要想从整体上解决一般手势的识别问题是不现实的.为了寻找突破口,必须研究手势分类和人际交流中的手势用法,从而为确定合理的识别范围及建立符合人类行为习惯的交互模型提供指导.

总体上说,现有的手势模型可以归为两大类:基于 3D 手(臂)模型的手势模型和基于表观的手势模型.原理上,基于 3D 手(臂)模型的手势模型适合于给所有手势建模,而基于表观的手势模型通常只适用于给交流性手势.然而,一方面,基于 3D 手(臂)模型的手势模型不仅参数多,计算复杂性高,而且为抽取模型参数而使用的许多近似过程导致模型参数的估计很不可靠.另一方面,基于表观的手势模型的计算复杂性低,易于达到实时.因此,目前绝大部分手势识别系统都采用了基于表观的手势模型.不过,最近随着计算机性能的提高,已经有人开始把经过简化的 3D 几何模型用于识别一定数量的交流性手势.

成功的手势识别策略应该考虑手势的时间 - 空间上下文,即考虑手势的语法规则.语法规则既要反映手势的语言学特征,又要反映手势的空间特征.然而,到目前为止只有很少数量的系统使用语法规则.基于单摄像机在复杂背景下实时识别多种手势是手势识别的发展方向.正如鼠标并没有取代键盘一样,手势识别系统并不是为了取代键盘,相反,手势识别将增强现有的人机交互模式,从而实现更直接、更自然、更和谐的人机接口,并且促成一些新的应用,例如手语识别等.研究多模态交互技术已逐渐成为研究人员的共识.英语的语音识别与唇语理解结合起来已成为不少人研究的课题.使用口语与手势接口进行分子结构设计的研究是虚拟现实应用的一个较为成功的例子.

参考文献

[1] T. Ahmad, C. J. Taylor, A. Lanitis, T. F. Cootes. Tracking and recognising hand gestures, using statistical shape models. Image and Vision Computing, 1997, 15: 345～352

[2] Y. Azoz, L. Devi, and R. Sharma. Vision-Based Human Arm Tracking for Gesture Analysis Using Multimodal Constraint Fusion. Proc. 1997 Advanced Display Federated Laboratory Symp., Adelphi, Md., 1997

[3] David Alan Becker, Sensi. A Real-Time Recognition, Feedback and Training System for T'ai Chi Gestures. (David Alan Becker, Master thesis), MIT Media Lab, May, 1997

[4] A. Bobick, J. Davis. Real-time recognition of activity using temporal templates. Proc. of Third IEEE Workshop on applications of computer vision, Florida, 1996, 39～42

[5] G. Bradski, Boon-Lock Yeo, Minerva M. Yeung. Gesture for video content navigation. SPIE 3656 (Proc. of the IS&T/SPIE Conf. on Storage and Retrieval for Image and Video Database VII), San Jose, California, 1999, 230～242

[6] Quek F. Unencumbered gestural interaction. IEEE Multimedia, 1996: 36～47

[7] R. Cipolla and N. J. Hollinghurst. Human-robot interface by pointing with uncalibrated stereo vision. image and vision computing, Mar. 1996, 14: 171～178

[8] J. L. Crowley, F. Berard, J. Coutaz. Finger tracking as a input device for augmented reality. Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, 1995, 195～200

[9] T. Starner, J. Weaver, et al. Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans. PMAI, 1998, 20(12): 1371～1375

[10] R. Culter and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. Proc. of 3rd Int'l Conf. Automatic Face and Gesture Recognition, Japan, 1998

[11] Trevor J. Darrell, Irfan A. Essa, Alex P. Pentland. Task-specific gesture analysis in real-time using interpolated views. IEEE Trans. PAMI, Dec. 1996, 18(12): 1236～1242

[12] W. T. Freeman, K. Tanaka, J. Ohta, and k. Kyuma. Computer vision for computer games. Proc. Int'l Conf. Automatic Face and Gesture Recognition, Killington, Oct. 1996: 100～105

[13] D. M. Gavrila, L. S. Davis. Towards 3D model-based tracking and recognition of human movement: a multi-view approach. Proc. Int'l Workshop on Automatic Face and Gesture Recognition, Switzerland, 1995: 272～277

[14] Tony Heap, David Hogg. Towards 3D hand tracking using a deformable model. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Vermont, 1996: 140～145

[15] C. Wren, A. Azarbayejani, et al. Pfinder: real-time tracking of the human body. Proc. Int'l Conf. Automatic Face and Gesture Recognition, Killington, 1996, 51～56

[16] J. J. Kuch, Visionr-based hand modeling and tracking for virtual teleconferencing and telecollaboration. Proc. IEEE Int'l Conf. Computer Vision, Cambridge, Mass., 1995

[17] J. Lee and T. L. Kunii. Model-based analysis of hand posture. IEEE Computer Graphics and Applications, Sept. 1995: 77～86

[18] V. I. Pavlovic, R. Sharma et al. Gestural interface to a visual computing environment for molecular biologists. Proc. of the 2nd Int'l Conf. on Automatic Face and Gesture Recognition, Vermont, 1996: 30～35

[19] G. Xu, Y. Zhu, Y. Huang et al. Automatic visual recognition of isolated hard gestures with computing spatio-temporal representations. Proc. of the 1998 Symp. on Image, Speech, Signal Processing and Robotics (ISSPR '98), 1998, Hong Kong, I: 49～54

任海兵　1998 年毕业于清华大学计算机科学与技术系,获工学学士学位.同年,被保送到清华大学计算机应用专业直读博士学位,师从徐光祐教授.

祝远新　1995 年毕业于清华大学计算机科学与技术系,获工学学士学位.同年,被保送到清华大学计算机应用专业直读博士学位,师从徐光祐教授.1999 年,获得清华大学计算机应用专业博士学位.同年,进入 University of Missouri-Columbia 读博士后.