文章编号:1006-8309(2002)01-0027-03

手势识别技术及其在人机交互中的应用

李清水[1],方志刚[1],沈模卫[2],陈育伟[1]
(1.浙江大学电子工程系,浙江 杭州 310028;
2.浙江大学工业心理学国家实验室,浙江 杭州 310028)

摘要:手势是一种自然、直观、易于学习的人机交互手段,手势输入是实现自然、直接人机交互不可缺少的关键技术。目前的手势识别技术主要分为基于数据手套和基于视觉两种。这两种方法各有自己的长处,也都取得了一些研究成果,但都还不成熟。手势输入作为一种自然、丰富、直接的交互手段在人机交互技术中占有重要的地位。

关键词:手势识别;计算机视觉;时空表观模型;动态时空规整
中图分类号:TB 18;TP 391.4 **文献标识码**:A

当前,人机交互技术已经从以计算机为中心逐步转移到以用户为中心,是多种通道、多种媒体的交互技术。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设备,人机间的通讯将不再需要中间的媒体,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[1]。手势研究分为手势合成和手势识别。手势识别技术分为基于数据手套和基于计算机视觉两大类。本文将主要针对基于视觉的手势识别技术,从手势的定义、手势分割、手势建模、手势分析、手势识别等方面综述手势识别的研究现状,并讨论手势识别在人机交互技术中的应用。

1 手势的定义

由于手势(gesture)本身具有多样性和多义性,具有在时间空间上的差异性,加上不同文化背景的影响,对手势的定义是不同的。这里把手势定义为:手势是人手或者手和臂结合所产生的各种姿势和动作,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)。静态手势对应空间里的一个点,而动态手势对应着模型参数空间里的一条轨迹,需要使用随时间变化的空间特征来表述。手势和姿势(posture)的主要区别在于,姿势更为强调手和身体的形态和状态,而手势更为强调手的运动[2]。

2 手势分割

手势分割(Gesture Segmentation)是基于计算机视觉的,是指如何把手势从手图像中分离出来。在复杂背景情况下,手势分割困难重重,还没有成熟的理论作为指导,现有的算法计算度高,效果也不理想。主要有以下几种:①增加限制的方法,如使用黑色和白色的墙壁,深色的服装等简化背景,或者要求人手戴特殊的手套等强调前景,来简化手区域与背景区域的划分。②大容量手势形状数据库方法,如密西根州立大学计算机系的 Cui Yuntao 建立了一个数据库,其中有各种手势类在各个时刻不同位置不同比例的手型图像,作为基于模板匹配识别方法的模板。③立体视觉的方法,如纽约哥伦比亚大学计算机系的 Gluckman 利用两个不在同一平面镜子的反射图像,计算物体与摄像机之间的距离,根据距离信息分割出人手。

3 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,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,使得识别系统能够对用户所做的绝大多数手势做出正确的反应。目前,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。基于表观的手势建模又可分为基于 2D 灰度图象本身、基于手(臂)的可变形 2D 模板、基于图象属性和基于图象运动 4 种。基于 3D 模型的手势建模方法考虑了手势产生的中间

作者简介:李清水(1973-),男,山西洪洞人,浙江大学信息科学与工程学院教师,硕士,主要研究领域为计算机人机交互合多通道用户界面设计。

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[3]。图 1 是同一种手姿态的几种模型。

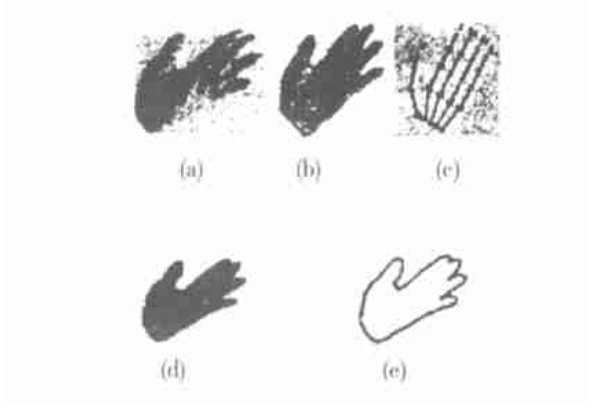

图 1　表示同一个手姿态的各种人手模型

(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

4　手势分析

手势分析阶段的任务就是估计选定的手势模型的参数。一般由特征检测和参数估计组成。在特征检测过程中,首先必须定位做手势的主体(人手)。定位技术有:①基于颜色定位:利用限制性背景或者颜色手套。②基于运动的定位:这种定位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的运动量很小[4]。③基于多模式定位:例如利用运动和颜色信息的融合定位人手[5],优点是能克服单个线索定位的局限。

不同建模方式参数估计方法不同:基于灰度图像本身的表观模型在最简单的情况下,可以选择模型视图序列作为参数[4],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;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和它们的方差。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;基于图像属性表观模型的常用参数是手形几何矩、Zernike 矩、朝向直方图[4]等。这些图像特征参数易于估计,但是它们对图像中其他非手物体非常敏感;基于运动图像表观模型的参数包括平移运动参数[6],旋转运动参数,以及图像变形参数等。例如 Becker 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[7]。

5　基于数据手套的手势识别

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。基于数据手套手势输入(Glove-based Gesture Input)的优点是输入数据量小,速度高,能直接获得手在空间的三维信息和手指的运动信息可识别的手势种类多,且能够进行实时地识别。

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。

6　基于计算机视觉的手势识别

基于计算机视觉的手势输入特点是对用户的运动限制少,需要处理的数据量大,处理方法复杂,不适合实时地识别。对静态手势的识别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。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、熟练程序会在轨迹的时间轴上引起非线性波动。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[3]。

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,Zernike 矩,特征图像系数向量,或者 3D 空间的运动速度[7]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归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。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首先利用某种特定属性的函数,例如在时间方向求和,把模型参数空间的一条轨迹压缩为单个点,然后利用静态手势识别算法完成动态手势的识别。例如 Bobick 等人提出的基于运动历史图像的实时动态手势识别就是利用这种方法[8]。相对前两种方法来说,这种做法更为简单和粗略,会丢失部分或者全部运动信息。

7　手势识别在人机交互中的应用

作为一种三维交互设备。

用于虚拟环境的交互。如虚拟制造和虚拟装配、产品设计等。虚拟装配通过手的运动直接进行零件的装配,同时通过手势与语音的合成来灵活的定义零件之间的装配关系。还可以将手势识别用于复杂设计信息的输入。

用于手语识别。手语是聋哑人使用的语言,是由手型动作辅之以表情姿势由符号构成的比较稳定的表达系统,是一种靠动作/视觉交际的语言。手语识别的研究目标是让机器"看懂"聋人的语言。手语识别和手语合成相结合,构成一个"人-机手语翻译系统",便于聋人与周围环境的交流。手语识别同样分为基于数据手套的和基于视觉的手语识别两种。基于 DGMM 的中国手语识别系统选取 Cyberglove 型号数据手套作为手语输入设备,采用了动态高斯混合模型 DGMM (Dynamic Gaussian Mixture Model) 作为系统的识别技术,可识别中国手语字典中的 274 个词条,识别率为 98.2 %[9]。

用于多通道、多媒体用户界面。正如鼠标没有取代键盘,手势输入也不能取代键盘、鼠标等传统交互设备,这一方面由于手势识别的设备和技术问题,另一方面也由于手势固有的多义性、多样性、差异性、不精确性等特点。手势识别要想取得比较高的识别率,仍有很长的路要走。手势输入在人机交互中应用的精髓不在于用来独立地用作空间指点,而是为语言、视线、唇语等交互手段通道提供空间的或其他的约束信息,以消除在单通道输入时存在的歧义。这种做法是试图以充分性取代精确性[2]。比如上面提到的虚拟装配中就需要语音的配合来定义零件之间的装配关系。使用口语和手势接口进行分子结构设计的研究也是一个成功的例子。手势非常适合于指点、表达形状、几何变换和装配等任务。语音对于表达抽象概念及离散属性(或命令)是具有绝对优势的,而且可以涉及视觉不及的对象。视线应用于人机交互在目标选择等方面具有直接性、自然性和双向性等特点。将手势输入和这些交互通道结合,将增强现有的人机交互模式,从而实现更为直接、自然、和谐的人机接口,如图 2。这种多模式的人机交互技术已经成为当前研究的热点,多通道人机界面将在可预见的将来占主导地位,并进一步促进虚拟现实技术的发展。我们在多通道用户界面的研究中,初步实现了将基于数据手套的手势输入整合到用户界面中的基本方法和机制,并通过可用性测试和评价的实验,得到一些有意义的结果,相信随着研究工作的深入,将取得更大收获。

用于机器人机械手的抓取[10]。机器人机械手的自然抓取一直是机器人研究领域的难点。手势识别,尤其是基于数据手套的手势识别的研究对克服这个问题有重要的意义,是手势识别的重要应用领域之一。

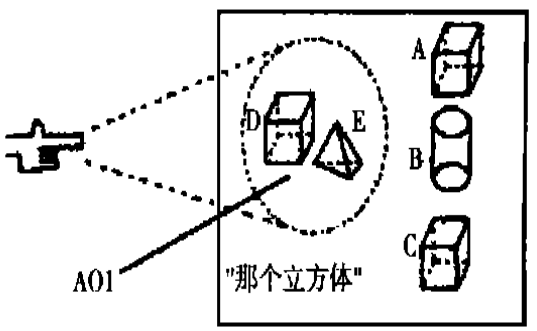

图 2　手势和语言结合确定目标

8　结束语

在当前的手势识别技术条件下,不论是基于数据手套的识别系统,还是基于计算机视觉的识别系统,都存在很多问题。比如,基于数据手套的手势识别系统,手套的大小会限制手套的戴用者范围;手套同手之间会发生滑移,影响精度等。基于视觉的手势识别系统中,假设景物中运动且具有人体皮肤色度特征的物体就是做手势的人手,这个假设在景物中出现大面积人脸时就不成立了[11]。期待尽快在一般手势的识别问题上取得突破是不现实的。但同时应该看到,即使在现有的技术条件下,手势输入仍然展现出广阔而美好的应用前景。这也将进一步促进手势识别技术的研究。

(下转第 33 页)

$$z(i) = \frac{1}{3}x_{i1} + \frac{1}{3}x_{i2} + \frac{1}{3}x_{i3}$$

$$= \frac{2}{3}x_{i\,\mathrm{I}} + \frac{1}{3}x_{i\,\mathrm{II}}$$

虽然,x_{I}的权重被人为加大了。

若用主成份变量 y_1,y_2 建立新的评价函数,有

$$z(i) = \left(\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}\right)y_1(i) + \left(\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}\right)y_2(i)$$

$$= \frac{2}{3}\left(\frac{\sqrt{2}}{2}x_{i1} + \frac{\sqrt{2}}{2}x_{i2}\right) + \frac{1}{3}x_{i3}$$

$$= \frac{2}{3}\left(\sqrt{2}\,x_{i\,\mathrm{I}}\right) + \frac{1}{3}x_{i\,\mathrm{II}}$$

x_{I}的权重不仅没有减小反而更加大了。

另外,主成份分析之前,任意两个样本点的距离,(以欧氏距离为例)为

$$d^2(j,k) = (x_{j1} - x_{k1})^2 + (x_{j2} - x_{k2})^2 + (x_{j3} - x_{k3})^2$$

$$= 2(x_{j\,\mathrm{I}} - x_{k\,\mathrm{I}})^2 + (x_{j\,\mathrm{II}} - x_{k\,\mathrm{II}})^2$$

主成份分析之后,以 y_1,y_2 为新的变量系统,样本点间的距离为

$$d^2(j,k) = (y_{j1} - y_{k1})^2 + (y_{j2} - y_{k2})^2$$

$$= \left(\frac{\sqrt{2}}{2}x_{j1} + \frac{\sqrt{2}}{2}x_{j2} - \frac{\sqrt{2}}{2}x_{k1} - \frac{\sqrt{2}}{2}x_{k2}\right)^2 + (x_{j3} - x_{k3})^2$$

$$= \frac{1}{2}\times 4(x_{j\,\mathrm{I}} - x_{k\,\mathrm{I}})^2 + (x_{j\,\mathrm{II}} - x_{k\,\mathrm{II}})^2$$

$$= 2(x_{j\,\mathrm{I}} - x_{k\,\mathrm{I}})^2 + (x_{j\,\mathrm{II}} - x_{k\,\mathrm{II}})^2$$

可见,主成份分析对数据的重复信息没有任何改善。因此,那种在选取指标变量时认为多多益善,不怕重复的做法是十分错误的。

3.4 主分量 y_2,y_3,…,y_m 用于评价需谨慎

由于主成份分析的特殊性,除了第一主分量 y_1 具有综合原数据变量的能力外,其他主分量 y_2,y_3,…,y_m 的物理意义是不明确的。只有对具体问题应用相关的专业知识才能较好地解释。因为绝大多数情形下,协方差矩阵 $V>0$,只能保证第一特征向量 $L_1>0$,而 L_2,L_3,…,L_m 中的各分量一般都不会保证同号,也就是说,y_2,y_3,…,y_m 若作为评价指数,需视具体问题谨慎对待。

4 结言

主成份分析法作为多元统计分析的一种常用方法,在处理多变量问题时具有其一定的优越性。通过主成份分析,往往能发现原问题中蕴含的某些综合性、深层次特征。由于主分量具有良好的特性,因此在评价问题中经常采用。

参 考 文 献

[1] 张尧庭,方开泰. 多元统计分析引论[M]. 北京:科学出版社,1982.

[2] 王学仁,王松桂. 实用多元统计分析[M]. 上海:上海科技出版社,1990.

[3] 杨维权,刘兰亭. 多元统计分析[M]. 北京:高等教育出版社,1989.

[收稿日期]2001-06-19
[修回日期]2001-08-07

(上接第 29 页)

参 考 文 献

[1] 邹晨,张树有,谭建荣,等. VR 环境中产品设计手势的定义与合成[J]. 工程图学学报,2000,21(2):107-110.

[2] 方志刚. 计算机手势输入及其在人机交互中的作用[J]. 小型微型计算机系统,1999,20(6):418-421.

[3] 任海兵,祝远新,徐光祐,等. 基于视觉手势识别的研究—综述[J]. 电子学,2000,28(2):118-121.

[4] Freeman WT, Tanaka K, Ohta J, et al. Computer vision for computer games[A]. Proc. Int'l Conf. Automatic Face and Gesture Recognition[C], Killington, 1996. 100-105.

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

[6] Quek F. Unencumbered gestural interaction[J]. IEEE. Multimedia, 1996, 36-47.

[7] Becker DA. A Real-time Recognition, Feedback and Training System for T'ai Chi Gestures[D]. (Master thesis) MIT Media Lab. 1997.

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

[9] 吴江琴,高文. 基于 DGMM 的中国手语识别系统[J]. 计算机研究与发展,2000,7(5):551-558.

[10] 魏军,王家顺,王田苗,等. 面向虚拟制造和装配的新型数据手套设计[J]. 机械工程学报,2000,26(2):91-94,98.

[11] 任海兵,祝远新,徐光祐,等. 连续动态手势的时空表现建模及识别[J]. 计算机学报,2000,23(8):824-828.

[收稿日期]2001-05-17
[修回日期]2001-08-01