基于RGB-D摄像头的实时手指跟踪与手势识别

刘鑫辰 傅慧源 马华东
(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要 近些年,基于视觉的手部跟踪与手势识别一直是人机交互和计算机视觉等领域的研究热点。传统方法主要是使用单目或多目RGB摄像头等设备获得手部位置、方向等信息,但RGB摄像头易受到复杂背景、光照变化、纹理的限制,导致其准确性、实时性和鲁棒性都较差。随着可获得场景深度信息的家用RGB-Depth(RGB-D)摄像头的发展和上市,可以利用深度信息较好地克服上述环境问题。首先定义了一个基于RGB-D摄像头的3D交互空间,根据深度信息将手部区域从复杂背景、多变的光照条件下进行分割;然后提出了一种基于深度摄像头的手指识别和跟踪方法,该方法基于手部轮廓对人手及手指进行识别和跟踪;最后通过对手指位置和轨迹的跟踪进行手势识别,从而实现人机交互。对提出的方法进行的实验验证了它的准确性、实时性和鲁棒性。

关键词 手指跟踪,手势识别,计算机视觉,人机交互,RGB-D摄像头
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.10.011

Real-time Fingertip Tracking and Gesture Recognition Using RGB-D Camera

LIU Xin-chen FU Hui-yuan MA Hua-dong
(Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract In recent decades, visual interpretation of finger and hand gestures has been an attractive direction in both computer vision and human-computer interaction areas. Traditional methods use a monocular RGB camera or multiple RGB cameras to get hand information. But it is limited by clustered backgrounds, lighting conditions, textures and other environment factors, which makes the accuracy, robustness and efficiency cannot satisfy real-time interactions. With the coming of consumer-level RGB-D camera, above limitations can be overcame with depth data got from a RGB-D camera. We first defineed a 3D interaction space with the RGB-D camera and segmented the hand region from backgrounds with the help of depth information. Then we proposed a real-time finger recognition and tracking approach using a depth camera which mainly use the contour of hands. At last, the human-computer interaction was achieved with the position and trajectory of fingers from the above method. Based on the proposed method, we designed several experiments. The results validate the accuracy, effectiveness and robustness of our approach.

Keywords Finger tracking, Gesture recognition, Human-computer interaction, Computer vision, RGB-D camera

1 引言

随着计算机软硬件、人工智能、多媒体、机器人技术的快速发展,传统基于键盘、鼠标和平面显示器的人机交互技术已无法满足当今人们的生产生活需求,以人为本的科技观使更加方便、自然、快速的人机交互技术成为近年研究的热点[1,3]。在人与人的日常交流中,视觉是信息的主要输入源,手势是信息的重要输出源,因此基于视觉的手势跟踪与识别是新一代人机交互技术的关键[4,14]。

基于视觉的手势跟踪与识别一般分为两个过程,即手部分割及手势识别。手部分割的目的是得到手的视觉信息,将其作为识别过程的输入。该过程需要将手部的视觉图像从复杂的背景图像、多变的光照条件下分割出来,得到手的位置、方向、形状等信息。本文研究的重点是在室内复杂环境条件下手部及手指的跟踪和手势识别。传统的手部分割和跟踪方法主要使用RGB摄像头、红外热像仪及可穿戴式传感设备(如数据手套)等获得以上的手部信息。在Lars[5]等的工作中,采用了形状与颜色特征相结合的方法,从单一RGB图像提取手部信息,再通过级联模型和粒子滤波进行手势识别。在Yikai[6]等的工作中,使用了基于形状特征的Adaboots方法进行手部识别,再用基于肤色的方法对RGB图像中的手部区域进行分割。在Kenji[7]等人设计的交互桌面系统中,采用了红外热像仪,通过感应人手的温度进行分割和跟踪,但对于温度的敏感性也是其局限性之一。Robert[8]等人设计了一种色块手套,通过对手套上的不同色块的识别进行分割和跟踪,得到手指和手掌的位置及方向信息。

随着微软的Kinect和华硕的Xtion等消费级RGB-D摄像头的发展和普及,交互场景的深度信息也更加容易获

到稿日期:2013-06-30 返修日期:2013-08-12 本文受国家自然科学基金创新研究群体项目(61121001),高等学校博士学科点专项科研基金(20120005130002),国家杰出青年科学基金项目:网络环境中多媒体计算(60925010)资助。

刘鑫辰(1988-),男,硕士生,主要研究方向为计算机视觉,E-mail:Liuxc.bupt@gmail.com;傅慧源(1986-),男,博士生,主要研究方向为计算机视觉;马华东(1964-),男,教授,博士生导师,主要研究方向为多媒体系统与网络、传感器网络、网格计算、形式化技术等。

取[13-16]。然而，目前它们大多被应用于人体骨骼、肢体运动等较大尺度对象的数据捕捉[16]，对于手及手指等小尺度对象的识别与跟踪才刚刚起步[9]。Iason[10]等人使用 Kinect 分割手部区域，然后采用基于模型的方法将获取到的手部 3D 信息转化为 3D 模型，再与模型库中的模型进行匹配从而识别不同手势。然而，基于模型的识别方法通常需要建立庞大复杂的模型库，模型匹配的过程也需要大量的计算，虽然识别准确率很高但效率较低。在 Cem[11]等人的工作中，使用随机决策树(RDF)对 Kinect 深度图像进行逐像素分类，再通过 Mean-shift 算法估计手指关节点的位置，从而跟踪手指。但基于分类和学习的方法通常需要建立巨大的数据集，并需要大量时间对分类器进行学习。Zhou[12]等人使用 Kinect 对手部进行分割，用特殊的腕带提高分割精度，提出了基于手部轮廓的 FEMD 算法，其结合模型匹配方法对手指进行识别和跟踪，能够识别 10 种不同的手势。该方法虽识别精度很高，但仅适用于静态图像中的手势识别，并不能达到实时的人机交互的目的。

本文提出了一种基于 RGB-D 摄像头的实时手指跟踪与手势识别方法，整体流程如图 1 所示。首先，我们设计了一个 3D 交互空间，在交互空间内通过深度信息分割手部区域。然后，通过手部区域得到手的轮廓信息，在轮廓的基础上提取手指指尖位置。最后，通过对指尖位置及轨迹的分析，识别出若干种手势。该方法在实验中体现了较好的实时性、精确性和鲁棒性，将来可以应用于如模拟多点触摸、虚拟现实、增强现实、游戏娱乐等方面，实现新一代人机交互。

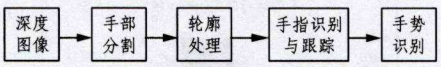

图 1　基于 RGB-D 摄像头的实时手指识别与跟踪流程

2　3D 交互空间

本文首先根据 RGB-D 摄像头的基本参数设计了一个 3D 交互空间，只有在该空间内的手部及手指才能够被分割和跟踪，因此交互过程被限定在交互空间内。

深度摄像头一般通过发射红外线或激光的方式来探测场景，再通过红外线传感器或激光传感器得到场景反射回来的射线，最后通过计算发射射线与反射射线的关系得到场景内各点的深度信息，常用的技术有结构光技术和 ToF(Time-of-Flight)技术等等[8]。我们以微软的 Kinect 摄像头作为实验设备，Kinect 的视野范围为：水平视角 57°，垂直视角 43°，深度范围为 800 到 4000 毫米，平均精度可达 1 毫米，采集速率为 30 Fps。由于距离传感器越近则深度的精确度越高，因此我们定义了距离传感器 800 到 1000 毫米之间的范围作为 3D 交互空间，在此深度范围内的手部将被分割出来。3D 交互空间如图 2 所示，左右两图分别表示俯视图和侧视图。

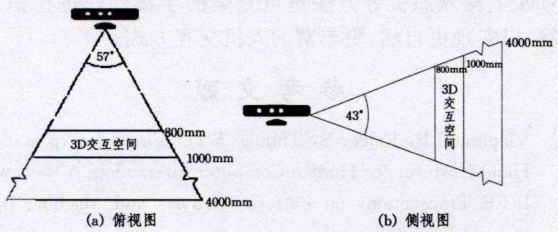

图 2　3D 交互空间示意图

在我们的方法中，人需要将手置于 3D 交互空间中，并确保手掌正对深度摄像头，手指方向向上。我们定义了如图 3 所示的 6 种不同手势，手在做出这些手势后可以在交互空间中移动位置。

图 3　检测 6 种手势

3　基于 RGB-D 摄像头的实时手指跟踪与手势识别

3.1　手部分割与跟踪

使用 Kinect 深度摄像头可以得到其视野范围内的一幅分辨率为 640×480 的深度矩阵，矩阵每一点的值表示该点到摄像头的距离，深度值以毫米计。根据 3D 交互空间的深度范围，将根据式(1)对原始深度图进行处理，以分割出规定范围内的深度像素。

$$D_{range}(x,y)=\begin{cases}D_{raw}(x,y), & \text{if } D(x,y)\in[minDepth,\\ & maxDepth]\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

其中，整数 $x\in[0,640]$，$y\in[0,480]$，$D_{raw}(x,y)$ 为在 (x,y) 点处的原始深度值(以毫米计，下同)，即分割出的图像中 (x,y) 点处的深度值，根据我们设计的交互空间，$minDepth$ 为 800，$maxDepth$ 为 1000。

按照上述方法提取出深度范围内的深度图像后，需要按照以下步骤对其进行处理。

1. 对深度图像二值化，得到手部区域的二值化图，即 3D 交互空间内部的像素设为 1，外部像素为 0。

2. 采用 Suzuki85 算法对二值化图进行处理，得到手部区域的轮廓，并以序列形式存储，得到轮廓序列 Seq。

3. 采用 Distance Transform 算法计算手部区域内部所有像素点到轮廓的最小欧氏距离，我们将最小距离最大的内部点作为手心点。

4. 根据式(2)，从轮廓底部开始，逆时针遍历步骤 2 得到的轮廓序列，将轮廓序列变换为序列点的顺序与其到手心点距离的曲线称为轮廓序列-距离函数曲线，该方法与 Zhou[12]采用的方法类似。

$$L(i)=\sqrt{(Seq(i)\cdot x-x_{center})^2+(Seq(i)\cdot y-y_{center})^2} \tag{2}$$

其中，$i\in[0,lengthof(Seq)]$，$L(i)$ 为序列 Seq 中第 i 个元素到手心点的欧氏距离，$Seq(i)$ 为序列 Seq 中第 i 个点。

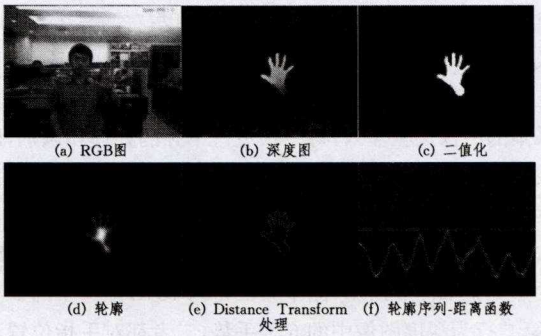

图 4

以上处理的结果如图 4 所示。通过上述 4 步处理，我们得到了手心点的位置以及轮廓序列-距离曲线，通过前者可以跟踪手部整体的位置，而后者则作为下面方法的输入，用于手指的识别与跟踪。

3.2 手指跟踪与手势识别

在我们提出的方法中，主要通过对手部的轮廓序列-距离函数进行处理，来识别手指尖的位置并进行跟踪。从轮廓序列-距离函数的曲线中可以看出，指尖的特点是位于轮廓序列曲线的局部极大值点处。因此，我们通过式(3)遍历轮廓序列-距离曲线，计算轮廓序列-距离函数的一阶差分，进而求差分值稳定地从正值跳变为负值的序列点，并将其判断为指尖点。

$$diff(i)=\begin{cases}0, & \text{if } i=0\\ L(i)-L(i-1), & \text{otherwise}\end{cases} \tag{3}$$

为得到稳定的差分跳变点，需要从前往后顺序遍历上面得到的差分曲线。以序点 i 为中心，我们设定了 2 个宽度为 C 的滑动窗口，对前后窗口按式(4)、式(5)计算窗口累计量。

$$w_{left}(i)=\sum_{k=1}^{C} diff(i-k) \tag{4}$$

$$w_{right}(i)=\sum_{k=1}^{C} diff(i+k) \tag{5}$$

其中，$w_{left}(i)$为序列点 i 的左窗口累计量，$w_{right}(i)$为其右窗口累计量，$i\in[k, lengthof(Seq)-k]$，C 为滑动窗口的宽度，我们取经验值 10。当 $w_{left}(i)$大于我们设定的经验值 α，且 $w_{right}(i)$小于经验值 β 时，序列点 i 被认为是一个指尖点，我们取 α 为序列中最大差分值的 0.8 倍，$\beta=-\alpha$。识别到的指尖点将被标记在 RGB 图和深度图中，由于 Kinect 深度摄像头与 RGB 摄像头存在视差，因此需要将深度图像下的坐标值转换为 RGB 图像下的坐标值。

通过记录指尖点位置和数量等信息，我们可以相应地识别出固定的 6 种手势。通过跟踪手指的轨迹状态并对轨迹进行分析，可以识别如单击、移动、多点滑动等其他手势。

4 实验

本文实验所使用的计算机为普通桌面 PC 机，采用 Intel Pentium 双核 2.7GHz CPU，2G RAM，微软 Windows7 32 位操作系统。RGB-D 摄像头采用的是微软 Kinect 传感器，采集频率为 30Hz，分辨率为 640×480 像素。为验证方法对手指跟踪与手势识别的精确性与鲁棒性，我们设计了 6 种手势，分别代表了 6 种不同的手指状态，如图 3 所示。

我们的实验分为静态测试和动态测试。在静态测试中，随机选择了 5 名测试对象，每人分别用左右两只手各采集 6 段长度为 400 帧的图像，每段图像中分别做出一种手势，手在交互空间内做任意的位置移动，总计采集 24000 帧图像，然后将采集的图像逐帧输入，每一帧进行单独的处理和识别，以测试我们方法的准确率和单帧处理时间。在动态测试中，我们让系统在普通的办公室环境下实时运行，测试人员在交互区内按上述手势操作，以测试系统的实时性和鲁棒性。跟踪的指尖点标记如图 5 所示，图中第一行为 RGB 图像中的 6 种手势，第二行为深度图像中的 6 种手势，二者都对手指位置进行了标记。

图 5 指尖标记图

通过实验可以看出，在准确性方面，本文方法与 Zhou[12] 等人的工作相比识别精度基本相当，均达到了较高的准确率。但在实时性方面，我们的方法处理速度分别是 Zhou[12] 两种方法的 6 倍和 48 倍，在上述系统环境下实时运行，测试对象按上述手势进行实时测试，单帧处理时间平均为 83.73ms，处理速度可达 10～15 帧/s，体现了较好的实时性。精确度与运行时间对比结果如图 6 所示。在便捷性与鲁棒性方面，与 Kenji[7] 的方法相比，他们使用了昂贵复杂的红外热像仪，对温度的敏感使其使用环境也受到限制；而我们所需的设备更易获得，对环境的要求更低，场景布置更加简单、便捷，更适合普通家庭、办公室等场景的交互使用。

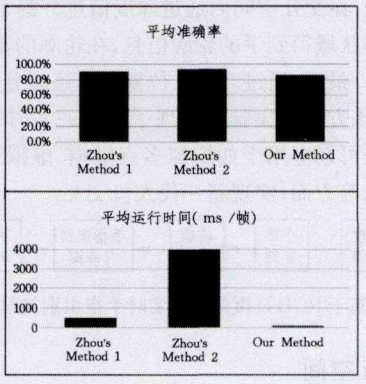

图 6 准确率与运行时间对比

结束语 本文介绍了一种基于 RGB-D 摄像头的实时手指跟踪与手势识别方法，通过 RGB-D 摄像头得到场景的深度信息，在我们设计的 3D 交互空间中对人的手部进行分割与跟踪，然后采用提出的方法对手部进行处理，对手指进行识别与跟踪。该方法在精确性、实时性、鲁棒性方面均有较好的表现，有效地解决了传统方法中复杂背景、环境光照变化等问题。对手指进行精确跟踪后，将来可以对更多人类自然手势进行识别和理解，为人机交互、多媒体技术、家庭娱乐等领域的发展提供了新的研究思路。

目前，该方法仅能对简单的手指手势进行识别与跟踪，因为现有 RGB-D 摄像头的精度仍然较低，识别精度也不能令人完全满意。在将来的工作中，如果 RGB-D 摄像头的精度能够更高，识别精度也能进一步提高。在跟踪方面，可以引入遮挡处理机制实现更多的双手交互；在手势识别方面，可以引入 HMM、有限状态机等方法对更复杂的手指动作进行识别和理解，以实现更自然、更丰富的人机交互方式。

参 考 文 献

[1] Vladimir I P, Rajeev S, Thomas S H. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7)

(下转第 61 页)

握手这一动作。若判断用户双方不全为握手手势，则不做后续处理；若双方用户均为握手手势，则应用接着会请求 GPS 信息和时间信息，并把信息发送到服务器。服务器会根据这些信息判断正在握手的两个用户，并把各用户的电子名片分别发送到对方的智能手机，并最终在智能手表上展示。

结束语 本文针对可穿戴设备协同使用的局限性提出并实现了一个支持可穿戴设备间数据及服务协同的移动中间件——Scudware Mobile。Scudware Mobile 以智能手机为中心，把各个可穿戴设备的数据和服务拆分并聚合，以统一的接口提供给上层应用，并加入了同步机制，以此实现数据和服务的协同。在 Scudware Mobile 的基础上，我们实现了两个应用：个人数据门户和 shaking e-card。

但是目前 Scudware Mobile 集中于个人的可穿戴设备，所有的设备协同也以个人服务为目标。未来我们希望搭建一个云平台来接入 Scudware Mobile，该平台具有两个功能：收集和存储个人数据；安全地把个人可穿戴设备的数据和服务共享出去，允许平台上的其他设备调用这些数据和服务，并通过把 Scudware Mobile 接入我们已有的 PSB[14] 总线，实现更大范围内的数据和服务分享。

参 考 文 献

[1] Xively[OL]. https://xively.com/

[2] Funf[OL]. http://funf.org/

[3] Wu Zhao-hui, Wu Qing, Cheng Hong, et al. ScudWare: A semantic and adaptive middleware platform for smart vehicle space [J]. IEEE Transactions on Intelligent Transportation Systems, 2007, 8(1): 121-132

[4] Wu Zhao-hui, Pan Gang. Smartshadow: Models and Methods for Pervasive Computing[M]. Springer, 2013

[5] Kukkonen J, Lagerspetz E, Nurmi P, et al. Betelgeuse: A platform for gathering and processing situational data[J]. IEEE Pervasive Computing, 2009, 8(2): 49-56

[6] Beach A, Gartrell M, Xing X, et al. Fusing mobile, sensor, and social data to fully enable context-aware computing[C]// Proceedings of the Eleventh Workshop on mobile Computing Systems & Applications. ACM, 2010: 60-65

[7] Brunette W, Sodt R, Chaudhri R, et al. Open data kit sensors: a sensor integration framework for android at the application-level [C]// Proceedings of the 10th international conference on mobile systems, applications, and services. ACM, 2012: 351-364

[8] Atzmueller M, Hilgenberg K. Towards capturing social interactions with SDCF: an extensible framework for mobile sensing and ubiquitous data collection[C]// Proceedings of the 4th International Workshop on Modeling Social Media. ACM, 2013: 6

[9] Open. Sen. Se[OL]. http://open.sen.se/

[10] Wang Yi, Lin Jia-liu, Annavaram M, et al. A framework of energy efficient mobile sensing for automatic user state recognition [C]// Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services. ACM, 2009: 179-192

[11] Sae-Tang A, Catasta M, McDowell L K, et al. Semantic place prediction using mobile data[C]// Mobile Data Challenge Workshop. June 2012: 18-19

[12] Chon Y, Lane N D, Li Fan, et al. Automatically characterizing places with opportunistic crowdsensing using smartphones[C]// Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012: 481-490

[13] Wu Jia-hui, Pan Gang, Zhang Da-qing, et al. Gesture recognition with a 3-d accelerometer[M]// Ubiquitous intelligence and computing. Springer Berlin Heidelberg, 2009: 25-38

[14] Pan Gang, Zhang Li, Wu Zhao-hui, et al. Pervasive Service Bus: Smart SOA Infrastructure for Ambient Intelligence[J]. IEEE Intelligent Systems, 2012, PP(99)

(上接第 52 页)

[2] Ali E, George B, Mircea N, et al. Vision-based hand pose estimation: A review[J]. Computer Vision and Image Understanding, 2007(108): 52-73

[3] Sushmita M, Tinku A. Gesture Recognition: A Survey[J]. IEEE Transactions on Systems, MAN, and Cybernetics—Part C: Applications and Reviews, 2007, 37(3)

[4] Juan P W, Mathias K, Helman S, et al. Vision-Based Hand-Gesture Applications[J]. Communications of the ACM, 2011, 54(2)

[5] Lars B, Ivan L, Tony L. Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering, Automatic Face and Gesture Recognition [C] // Proceedings. Fifth IEEE International Conference. May 2002: 423-428

[6] Yikai F, Kongqiao W, Jian C, et al. A Real-time Handgesture Recognition Method[C]// 2007 IEEE International Conference on Multimedia and Expo. 2007: 995-998

[7] Kenji O, Yoichi S, Hideki K. Real-Time Fingertip Tracking and Gesture Recognition[J]. IEEE Computer Graphics and Applications, 2002, 22(6): 64-71

[8] Robert Y W, Jovan P. Real-Time Hand-Tracking with a Color Glove[J]. ACM Transactions on Graphics (TOG), 2009, 28(3)

[9] Jesus S, Robin R M. Hand Gesture Recognition with Depth Images: A Review[C]// RO-MAN, 2012 IEEE. 2012: 411-417

[10] Iason O, Nikolaos K, Antonis A A. Efficient Model-based 3D Tracking of Hand Articulations using Kinect[C]// Proceeding of BMVC. 2011

[11] Cem K, Furkan K, Yunus E K, et al. Real Time Hand Pose Estimation using Depth Sensors[C]// 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). 2011: 1228-1234

[12] Zhou R, Junsong Y, Zhengyou Z. Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera[C]// MM '11 Proceedings of the 19th ACM International Conference on Multimedia. 2011: 1093-1096

[13] 吴莉婷，张宇，杨一平，等. 深度图像中基于轮廓曲线和局部区域特征的 3 维物体识别[J]. 中国图象图形学报，2012，17(2)：269-278

[14] 梁秀波，张顺，李启雷，等. 运动传感驱动的 3D 直观手势交互[J]. 计算机辅助设计与图形学学报，2010，22(3)：521-526

[15] 万华根，肖海英，邹松，等. 面向新一代大众游戏的手势交互技术[J]. 计算机辅助设计与图形学学报，2011，23(7)

[16] 周瑾，潘建江，童晶，等. 使用 kinect 快速重建三维人体[J]. 计算机辅助设计与图形学学报，2013，25(6)