第25卷 第1期 计算机工程与设计 2004年1月
Vol. 25 No. 1 COMPUTER ENGINEERING AND DESIGN Jan. 2004

文章编号：1000-7024(2004)01-0122-04 中图分类号：TP391 文献标识码：A

基于计算机视觉的人运动捕获综述

于 广

(华南理工大学，广东 广州 510640)

摘 要：人体运动的捕获是计算机视觉领域的热点之一，它是指从大范围上从图像序列中提取并描述人体轮廓的运动，然后对其进行跟踪识别。介绍了运动捕获的潜在应用，从初始化、检测、跟踪、姿态评估、识别5个方面分析了有关运动捕获的分类及研究现状，最后简要探讨了该领域面临的难点问题及发展趋势。

关键词：计算机视觉；运动捕获；图像

Survey of computer vision-based human motion capture

YU Guang

(South China University of Technology, Guangzhou 510640, China)

Abstract: Human motion capture is currently one of the most active research topics in the domain of computer vision. It attempts to extract and describe human motion, then track and recognize, from image sequences involving humans. The potential application areas are introduced and a comprehensive survey of the state-of-the-art is gaven, from five major issues: initialization, detection, tracking, pose evaluation and recognition. Finally the research challenges and future directions is briefly discussed.

Key words: computer vision; motion capture; image

1 引 言

对人体的运动分析是目前在计算机视觉领域最热门的应用之一，其中一个重要的部分就是人运动的捕获。这一术语是指从大范围上描述人体轮廓的运动，包括头、四肢、躯干，基本不涉及人脸识别及手语等。其基本任务是从摄像机摄取的视频图像序列中，跟踪一些关键点或部分(关节)，将其转换成有用的数学术语，然后合并恢复人体的结构参数，并对人的行为进行识别、判断、跟踪与理解，进而实现计算机的智能监控、身份识别、虚拟现实应用等。其研究内容涉及计算机视觉、人工智能、图形学、模式识别等多学科的交叉应用。事实上，捕获的可以是运动的任何实体，包括车辆、人体、机器人等。而对人体的捕获是最富挑战性的，比照对其它物体的跟踪：人的捕获必须是高精度、实时性、鲁棒性的；必须纪录全自由度的人体各部分的运动；对在不同角度的摄像机的定标及所得的数据要进行融合和遮挡的处理；必须分析处理人运动(包括单人及多人)的不确定性和复杂性等。因此对人运动的捕获，必须采用专门的技术与设备，大体可分为传感技术与分析技术。用于人体捕获的传感器的类型大体可分为两类，即被动式(采用摄像机)和主动式，包括电磁式(利用电磁场)、机电式(利用人体上的电位仪、加速计、声学设备)等。前者分析技术更复杂，而后者传感器技术更复杂，它们各有其优缺点及适用范围。本文只涉及基于计算机视觉(被动式)的人运动捕获研究。

2 应用领域

国外对有关人运动捕获的技术十分重视。从美国国防部的VSAM(Visual Surveillance and Monitoring)项目到美英的各大学甚至Microsoft这样一些大公司都在开展相关研究，该领域也成为一些国际权威期刊和学术会议的重要议题之一。这显然是由于其具有广阔的应用前景，学者们大都认为其应用将体现在以下几个方面。

2.1 智能监控

在对安全要求敏感的军事区域、国界、机场、停车场、银行等地方都需要智能监控设施。而基于计算机视觉的人运动捕获技术则可以实现24小时实时监控，并且能自动对摄像机捕捉的数据进行分析，若发现可疑人可以及时发出警报，同时也减少了雇用大批监控人员的费用；而用于交通管理、公共场所、超市等则可以自动监控以实现流量控制等。例如Haritaoglu的论文中提到的W4实时视觉监控系统就可实现户外条件下人的定位与识别，并能

收稿日期：2002-11-06

作者简介：于广 (1970-)，男，山东人，硕士生，研究方向为多媒体应用、嵌入式系统等。

实现多人跟踪。

2.2 运动分析

在运动捕获中所做的步态分析、关节运动模型研究可应用于个性化的体育运动和舞蹈训练，也可用于基于内容的快速搜索；而在医学领域，则可以分析病人的行为和步态，判断其疾病或受伤情况，以便做出有效的治疗；另外人运动分析与生物特征识别相结合的视觉监控，特别是非接触式远距离的身份识别可以实现远距离情况下人的检测、分类和识别，从而增强国防、民用等场合的保护能力。

2.3 高级人机接口

在此领域中，对人运动捕获的研究可以帮助我们实现智能的人机交互。例如在高噪音场合(例如机场、车间等)可以进行比语音方便精确的交流或信息输入。也可用于使机器人绕过障碍物等。

2.4 虚拟现实技术

对人运动捕获的研究所获得三维物理模型及其综合与再现技术可以用来产生逼真的人体形象，配合面部表情、头手动作等可以用于视频游戏、动画制作、远程控制、虚拟聊天室、模拟训练、远程会议系统等。

3 分类及研究现状

近年有关基于计算机视觉的人运动捕获的论文显著增长。适当地分类有助于我们了解目前该领域的发展状况。基于不同的标准，可以给出不同的分类：按照如何对目标进行检测与识别有基于模式的方法和无模式的方法以及二维方法和三维方法、分布式和集中式等。也有人以基于形态模式或是基于运动特性加以分类，对于如何进行特征提取则有人以光流(Optical Flow)和运动一致性(Motion Correspondence)来区分。Moeslund 和 Granum 则从系统中各部分功能的角度，认为任一个人运动分析系统都可分为以下4个部分，即初始化、跟踪、姿态评估、识别。当然也有的系统是上述4个部分的简化版或合并版。但该方法显得过于笼统，例如跟踪还可进一步区分为检测、分类、跟踪3部分。事实上大多数系统都基本遵循初始化、检测、跟踪、姿态评估、识别与理解几个步骤，因此不失一般性，我们按此分类对基于计算机视觉的人运动捕获的研究现状加以描述。

图 1 人运动分析系统的基本组成

3.1 初始化

所谓初始化是指系统运行之前的配置工作。需要指出的是，到目前已知的研究系统都是建立在某些约束或者假设的基础上。基本有两个方面：①运动方面，即要求试验对象(通常是人)的运动是均匀的、平面的、并在一定范围内，基本没有遮挡；②外观上的，指光照均匀，背景固定一致，摄像机的参数和人的姿态已知，穿紧身衣服或身上有标志物。这些约束的目的是为减少干扰，简化任务。当然，随着技术发展及系统复杂性的提高，约束条件也将减少。

初始化包括3方面的内容：①对摄像机参数的调校，包括离线与在线的校准；②场景的设置，这些设置是基于在外观上的假设，它有利于运动分割和背景消除；③模型的选择，包括人的初始姿态和姿态描述，通常用于基于模型匹配的系统，当然基于不同的方法和研究目的，有的系统并不进行模型初始化。

3.2 检 测

检测的目的是从图像序列中将变化区域从背景中提取出来。检测基本属于底层处理，主要工作是图形数据分割(Figure-Ground Segmentation)和特征抽取(Feature Extraction)。Moeslund 在其论文中将有关图像信息分为运动数据(Motion Data)、空间数据(Range Data)和外观数据(Appearance Data)。运动数据是指在静态场景中运动变化的部分，而这通常就是人的运动，其处理一般有两种方法：光流与变化检测；空间数据是指观察场景中的三维信息，这有助于从二维到三维的模型重建；外观数据则是指人身上的衣物或标志相对于场景的变化，处理方法则有阈值法与统计法。事实上从上述的划分可看出用于运动检测的常用方法基本有背景减除(Background Subtraction)、光流(Optical Flow)、统计法(Statistics)、时间差分(Temporal Difference)等。

背景减除是最常用的算法，特点是简单，但对动态场景的变化(如光照)等敏感。通常用于简单的均匀或静止背景中，利用当前图像与背景图像的差分来检测运动物体。常用的方法有阈值法、帧差分(Frame Differencing)、纹理斑点分割(Texture Blob Segmentation)、红外线背景照明(Infrared Back-lighting)、立体深度减除(Stereo Depth Subtraction)等。例如 Nakazawa 等通过先记录静止场景的图像，然后作为参考以实现背景减除。麻省理工学院媒体实验室(MIT Media Lab)开发的人体运动跟踪系统，实现了静态复杂背景下的实时用户运动跟踪，并且能够克服短暂的光照变化和遮挡等因素的扰动。

光流利用运动目标在图像序列间也即随着时间而变化的特性，通过计算帧间像素的位移来提取人的运动。即使在摄像机运动时也可检测出运动目标，但计算量大、反应慢、抗噪性差。例如 Yamamoto 等通过光流的计算来得到人体的运动参数，并且将之与其运动模型相比较。Bregler 将每个像素用其光流来描述，而光流是由运动相关和混合多元高斯模型来描述的。

统计法则是基于像素的统计特性而从背景中提取运动信息，它首先计算背景像素的统计信息(颜色、灰度、边界等)，再将现有像素与之比较，然后归类。该方法比背

景减除更强壮，但涉及大量计算和变换。而另一种较新的采用斑点(Blob)的方法则将相同颜色的像素归类为一个斑点，人则由一组斑点组成，它通过比较图像序列中斑点的位移得到人的运动。

3.3 跟踪

即建立观察对象与图像序列之间的对应匹配关系，其难度与场景和人运动的复杂度有关，特别在多人场景下。常用的数学工具有卡尔曼滤波(Kalman Filtering)Condensation算法、动态贝叶斯网络（Dynamic Bayesian Network）等。其中卡尔曼滤波是基于高斯分布的状态预测方法，不能有效地处理多峰模式的分布情况。Condensation算法以因子抽样为基础的条件密度传播算法，有较强的鲁棒性，但需要大量的抽样计算。目前就跟踪对象而言有跟踪头、手等身体部分与跟踪整个人体；就跟踪视角而言有对应于单摄像机的单视角与多摄像机的多视角之分。但就跟踪方法而言则有很多分类。常用的有：①基于模型的跟踪，该方法首先预定义一个模型然后再将实际运动与该模型匹配，例如Ju等人使用纸板人模型，它将人的肢体用一组连接的平面区块来表达，而区块的参数化运动受关节运动的约束，该模型被用来进行关节运动的图像分析；②基于运动轮廓的跟踪，利用曲线来描述目标，并在跟踪中自动更新。此外，还有基于区域、特征的跟踪等，事实上这一过程是与后续的姿态估计紧密联系的。

3.4 姿态估计

姿态估计即识别人体或四肢在场景中的运动，它既可作为跟踪的后期步骤也可作为跟踪处理的主动部分，并且根据需要有不同的精度，例如在监控应用中只需要识别出头、手甚至仅仅躯干就可，而医学运动分析则需要识别四肢的精确运动。按照系统如何使用人体模型可以将姿态估计分为3类，即无模型、直接模型与间接模型。

无模型即系统中没有一个欲设的模型，对姿态的描述采用点、简单图形、线图等方式。例如对于线图法，由于人运动实质是骨骼的运动，因此采用直线可近似表达人的各部分，并且其含有的结构信息在研究步态时十分有用。另一种完全不同的方法是直接从图形特征中映射出姿态信息。这样的系统建立在对大量可靠数据的训练基础上。例如Rosales和Sclaroff利用关节运动的三维数据训练其系统，使用扩展的EM算法和神经网络进行训练，来映射出关节的平面位置，然后使用一个圆柱体来合成出人体轮廓。

间接模型是指分析数据时以模型来估算出人体姿态，以该模型作为间接的参考或查找表。例如Leung和Yang使用U形边界来描述人体轮廓，他们使用了一个二维的带状模型，并从图像数据中查找与模型中相似部分，采用选取的“带”组成的边界区域就可表达出人体轮廓。

对于直接模型则使用一个欲设的模型来直接描述人的运动，这个模型根据观察数据不断更新，大概有40%的论文都采用此模式，其优点是较易处理遮挡，并且系统可兼容许多不同的轮廓模型。模型通常由关节和线条骨架组成，用轴来表示状态空间中的关节自由度，用状态空间来描述姿态，通常的方法是在预测、匹配、更新过程中使用分析与合成技术。其原理是先预测下一图像的姿态，再将这一预测模型分析、合成、抽象，然后与真实图像数据比较。直至找到最佳匹配的模型，然后更新系统模型。

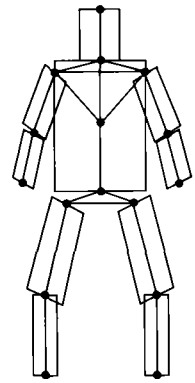


图2 人体二维模型

图3 线图法提取的人体模型

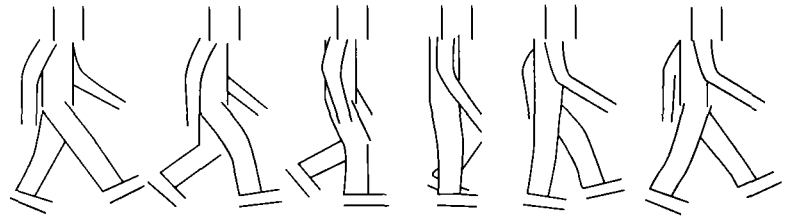
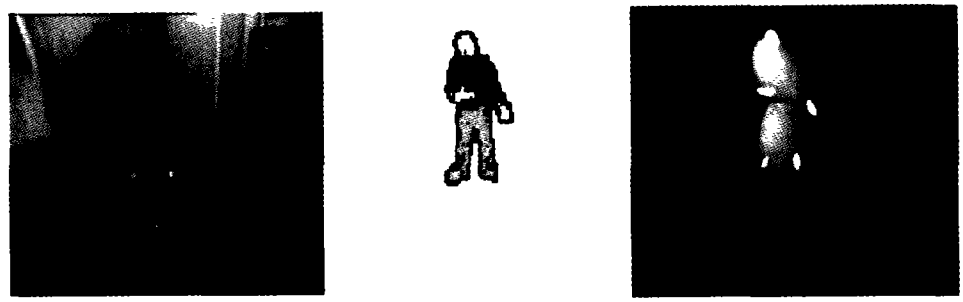


图4 二维轮廓人体跟踪



输入图像　　斑点分割结果　　三维图像重建

图5 斑点分割建模过程

例如Rohr使用14个椭圆柱体来产生人体结构，坐标原点被定位在人体中心，然后利用该模型来描述人体行走的三维运动，而圆柱的描述只需一个参数，该参数确定了圆柱的相位，显然这是一种非常高效的方法。

3.5 跟踪识别

所谓识别是区分不同运动行为的能力。传统上有两类方法，一是基于重建的的识别，一是从底层数据直接识别。Johansson使用MLD(Moving Lights Displays)在200ms内就可识别出行走一类的人体运动，这是直接识别的实例。基于重建的识别与姿态评估是紧密相关的，它依靠抽象得到的高层数据进行匹配与识别。更一般的方法是将识别区分为静态识别与动态识别。每种方法都涉及对低层和高层数据的处理。

静态识别是指对空间数据的处理。例如 Oren 预先将步行分割成一列图像，并用哈尔小波变换(Haar Wavelet)产生一个通用模板，将模板与得到的图像比较匹配，就可找到步行者。对于高层数据，Campbell 和 Bobick 假设关节位置已知，来识别不同的芭蕾舞步，他们使用状态空间和二维阙值来分类和识别。

动态识别涉及对时间数据的处理。例如对于底层的没有太多处理的时空数据，Chomat 和 Crowley 使用由组成分析原理(PCA)计算得到的时空滤波器，产生运动模板，然后使用 Bayes 分类器来选择识别。

对于已经过姿态估算的高层数据，例如 Becker 和 Pentland 使用隐马尔可夫模型 HMM (Hidden Markov Models)来分类不同的手语。

4 结束语

评价一个系统的性能通常都要考虑鲁棒性、精度、速度等方面。通过分析，我们发现基于计算机视觉的人运动捕获目前正处于发展初期，这主要表现在只能应用于受限的简单场景中，只能处理单人或几个人的简单运动，对于跑动或外部干扰通常无能为力。因此距离实用还有很长一段路要走，在算法方面还有很多难点有待解决。而这些难题的解决直接影响到该领域的发展。这些问题包括多摄像机的使用、三维建模、遮挡处理、精确运动分割、动态模型匹配等。未来该领域的发展将是多学科相结合，例如计算机视觉技术与图形学的结合，运动处理将借鉴语音识别、人工智能和模式识别等方面的技术。目前一些专用的商品化软硬件产品也已开始出现，例如在计算机动画方面的应用。我国科学家也已在计算机视觉及运动捕获、视觉监控等领域开展研究，并取得了一定的成果。

参 考 文 献:

[1] Haritaoglu I, Harwood D , Davis L. W4: Real- time surveillance of people and their activities[J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2000, 22 (8): 809-830.

[2] Pentland A. Looking at people: Sensing for ubiquitous and wearable computing [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2000, 22 (1): 107-119.

[3] Begler C, Malik J. Video motion capture[C]. Computer Science Division, Univ. of California, Berkeley, Berkeley, CA, 1994.720-1776.

[4] Aggarwal J K, Cai Q. Human motion analysis[C].A Review Computer Vision and Image Understanding, 1999,73 (3): 428-440.

[5] Thomas B, Moes Lund. Computer vision-based human motion capture - A survey[R]. Technical Report LIA 99-02 University of Alborg, 1999.

(上接第 121 页)

计算对象有N个，单个对象完成其所负载的计算一次时，最大耗时 Td，在这N个对象中有依赖关系的最大层次是m，一次跨网络的通信耗时 Tn，则生成一帧图像，最坏的情况耗时大约为 (Td + 2Tn)×m + Tg。当考虑到使用流水线技术时，则生成一帧图像耗时大约为 max [Td+2Tn, Tg]。可见大大提高了计算能力。当任务负载在每个网络节点计算能力之内时，这一数值不会波动，能满足较高的多任务和多用户负载。

5 总 结

面对这种规模和要求的可视化分析系统，使用专用系统或从底层做起，都是毫无经济性和实用性可言的，只有借助最普及的、最容易的技术，灵活地运用，快速地开发和维护，才是有生命力的。经过对原型系统的测试，基本达到了设计需求和目标，能够满足大容量、多用户的计算，显示了分布式系统的优越性。

参 考 文 献:

[1] 李晓梅. 并行与分布式可视化技术及应用[M]. 北京:国防工业出版社, 2001.

[2] Wu Jie. 分布式系统设计[M]. 北京: 机械工业出版社, 1999.

[3] Box Don. Essential COM[M]. 北京:中国电力出版社, 2001.

[4] OMG. The common object request broker: Architecture and specification[M].北京:电子工业出版社, 2001.

[5] Microsoft. MSDN [DB/CD]. http://www.microsoft.com/msdn.