融合深度图和三维模型的人体运动捕获*

肖秦琨 谢艳梅
(西安工业大学电子信息工程学院 西安 710032)

摘 要：对于当前热点的运动捕获方法存在的一些缺点，提出了一种融合深度图和三维模型的人体运动捕获方法。利用Kinect采集深度图像，经过对深度图去除背景，提取轮廓信息，建立轮廓数据库。从深度图中提取三维人体骨架，建立骨架三维模型数据库。输入1组深度图动作序列，经过去背景、提取轮廓特征后与轮廓数据库中的轮廓进行匹配，计算出最小距离所在的匹配序列，输出相应的骨架作为动作捕获的结果。实验证明了这种方法的有效性和可行性，该方法能较精确的得到运动捕获数据。

关键词：运动捕获；深度图；轮廓信息；三维骨架；DTW
中图分类号：TN **文献标识码**：A **国家标准学科分类代码**：510.99

Human motion capture based on depth and 3D model

Xiao Qinkun Xie Yanmei
(College of Electronic Information and Engineering, Xi'an Technological University, Xi'an 710032, China)

Abstract: The method based on depth and 3D model for motion capture is proposed in this paper for some disadvantages in the current method of motion capture. It collects the depth images by Kinect, then, it removes the background of depth image and extracts the contour information. After, it creates a database of contour. Then it extracts skeleton of human body through depth image so as to create 3D model database of skeleton. Finally, for a set of sequence of depth image with motion as inputting, it will match with the database of contour after the process of removing the background, extracting the contour information. Then it calculates the minimum distance for matching sequence, and outputs the corresponding skeleton as a result of the motion capture. The result of experiment shows the availability and feasibility of this approach, it also can get motion capture data more accurate.

Keywords: motion capture; depth image; contour information; 3D skeleton; DTW

1 引 言

人体运动捕捉(Motion capture)技术是指获取人体的运动信息并采集记录的技术，现在人体运动捕获技术已经被应用于实际工程应用中[1]，如影视动画、游戏、新一代人机交互、动作识别等。到目前为止常用的运动捕捉技术可分为机械式、声学式、电磁式和光学式。机械式运动捕捉依靠机械装置跟踪和测量运动轨迹。Shiratori 等使用任天堂的游戏控制器进行了动画角色的行为控制[2]。机械式运动捕捉系统对象的动作限制较大，使用不便。声学式动作捕捉利用超声波的穿透性解决了人体的遮挡问题，但捕捉有较大的延迟，误差较大。电磁式动作捕捉系统主要由电磁发射源、接收传感器和数据处理单元组成，它对高速运动捕捉效果失真度较高。光学式动作捕捉为目前应用较为广泛的方案[3]，其中基于视频的动作捕捉也是常用的运动分析方式[4]。基于深度图像的运动捕捉方式近年来得到较大的发展[5]。Ishigaki 等使用光学动作捕捉设备获取用户的动作[6]。光学式动作捕捉的使表演者活动的动作幅度大，无机械装置对动作的束缚，实时性表现也很好，但是在使用时还是需要在表演者身上粘贴标记点，或穿上特制的表演用服装，对表演场地也有一定要求，实现难度较大，算法复杂。

提出的融合深度图和三维模型的人体运动捕捉技术，使用深度摄像机采集深度图像，由于深度图是三维的，解决了身体的自遮挡导致信息缺失的问题，保证了信息获取的完整性和可靠性，提高了动作识别的准确性。在特征匹

收稿日期：2014-11
* 基金项目：国家自然科学基金(61271362)项目

配上，使用轮廓信息，它能较好的反应目标形状，数据量小，减小了计算量。同时使用的 DTW 方法用于序列匹配，具有算法鲁棒的优点。

2 算 法

2.1 深度图的采集

本文使用 Kinect 采集深度图像。使用 Kinect 可以方便的获取物体的 RGB 信息和深度信息[5-7]。Kinect 有 3 个摄像头：中间的 RGB 彩色摄像头、两边的红外发射器和 CMOS 摄像机。其中 RGB 镜头可以采集到彩色图像，分辨率为 640×480；红外发射器和红外 CMOS 摄像头可以采集到深度图像，分辨率为 320×240。图 1(a)为采集到的深度。

2.2 轮廓数据库的建立

对于物体形状的描述有很多种方法，而轮廓信息能比较好的描述目标的形状，反应目标的重要特征，所以本文使用轮廓信息来描述人体动作。

2.2.1 去除背景

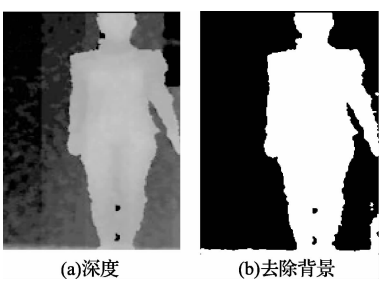
深度图像中包含目标区域和背景区域，为了得到人体动作的轮廓信息，必须先去除背景[8-9]。这里将采用阈值法对背景进行去除。

对于一个深度图 $F(x,y,d)$(其中 x，y 分别为像素坐标系下的横坐标和纵坐标，d 为深度信息)，假设基于深度信息分割背景区域和目标区域的阈值：

$$T=\frac{\max DepthVulue+\min DepthVulue}{2} \tag{1}$$

式中：$\max DepthVulue$ 和 $\min DepthVulue$ 分别为图像深度值的最大和最小值，将 T 记录在 T_0 中，根据阈值 T，将 $F(x,y,d)$分为前后两部分，分别求出两部分深度值的平均值 u_1 和 u_2。

重新计算阈值 $T=\frac{u_1+u_2}{2}$，判断 T 与 T_0 的差值是否在可接受的范围内，如果差值过大，那么将 T 记录在 T_0 里，对 T_0 进行更新。重复上面步骤，直到 T 与 T_0 的差值在可以接受的范围内，终止算法。把最后得到的 T 作为最佳阈值对 $F(x,y,d)$ 进行分割，去除背景，得到完整的人体动作深度信息 d_0。图 1(b) 为对深度图去除背景后的效果。



(a)深度 (b)去除背景

图 1 采集到的深度和去除背景的结

2.2.2 提取轮廓

描述轮廓的描述子基本上有统计不变矩、傅里叶描述子和基于曲率的描述子 3 大类。傅里叶描述算法建立在傅里叶分析的理论之上，它考虑了边界的全局和局部信息，对噪声不敏感，而且便于实现，是描述轮廓的比较好的算法。

假设图像的轮廓表示为 $\{x(m),y(m);m=0,1,\cdots,M-1\}$，其复数形式：

$$z(m)=x(m)+jy(m)m=0,1,\cdots,M-1 \tag{2}$$

对于复数序列 $z(m)$ 的傅里叶变换：

$$Z(k)=\frac{1}{M}\sum_{0}^{M-1}zm\exp\left(\frac{-j2\pi mk}{M}\right)$$

$$k=0,1,\cdots,M-1 \tag{3}$$

式中：k 表示数字化频率，$Z(k)$ 表示复数序列 $z(m)$ 的频谱系数[10]。$z(m)$ 的低频系数对应轮廓的全局形状，高频系数对应轮廓的细节。但是直接得到的频谱系数 $Z(k)$ 作为描述子不具有旋转、平移、尺度的不变性，需要对傅里叶描述子进行归一化处理。归一化傅里叶描述子 FD 定义：

$$FDk=\begin{cases}1,k=0\\ \dfrac{|Zk|}{|Z1|},k=1,2,\cdots,M-1\end{cases} \tag{4}$$

如图 2 所示为傅里叶描述子提取的人体轮廓。对采集到的深度图，重复以上过程，建立轮廓数据库。

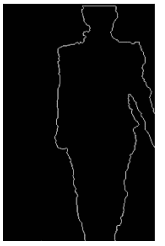


图 2 轮廓

2.3 三维人体骨架数据库的建立

现有的运动捕获的结果大多用骨架来表示。对于提取骨架的方法，近年来有很多专家学者进行过研究，最典型的是基于“非脊点下降”的方法和“冲刷模拟”的方法[11]。对于骨架的提取，理想的效果是能够确定骨架的关节点，然后按照人体本身的顺序将其连线。本文算法就是基于此种思路进行的具体方法。

1)确定身体的范围。人体的动作主要是靠四肢来体现，身体区域基本上不变，所以将身体上身区域近似视为四边形，肩关节 a1 和 a2 是四边形上方的 2 个顶点，髋关节 d1 和 d2 是下方的两个顶点。人身体区域的显著的特征是面积大，所以可以采用腐蚀膨胀的方法，将身体部分提取出来。对整个人体区域经过一定次数的腐蚀，面积相对较小的腿部、头部以及没有被遮挡的手部都会被腐蚀掉，而身体区域中心部分会保留下来，再使用相等次数的膨胀操作，就可以恢复原来的身体部分，身体部位区域之外的部分则被完全消除。图 3(a)～(c)是腐蚀膨胀的过程。

对图 3(c)中身体部分，利用近似四边形特性，统计身

体区域宽度的平均值，得到该区域的平均宽度。根据这个宽度 W，就可以确定两肩之间的近似宽度 W_s。然后对身体区域从上至下搜索扫描，当区域某行宽度大于并接近于 W_s 时，可以认为检测到了两肩的位置。同理，从下往上搜索扫描，当身体区域宽度小于并接近于 W_s 时，确定出两髋关节的坐标。确定出的 4 个坐标点如图 3(d)所示。

2)经腐蚀膨胀操作后，颈部呈尖角的形状，颈关节点 b 可以取尖角的位置；沿 b 点向上，最上端就是头部节点 c 的位置；胯关节点 e 在 d1、d2 的中点位置。

3)确定手关节点和肘关节位置从 a1、a2 处开始进行搜索，若手臂伸直，端点是手关节点 g1、g2 的位置，肘关节点 f1、f2 取手和肩中点的位置；若手臂弯曲，肘部是肩和手之间的一个端点，肘部另一端的端点是手的位置。

同理，可以确定出膝关节点 h1、h2 和脚关节点 i1、i2 的位置。根据以上方法确定的关节点如图 3(e)。

4)从前面的步骤中可以得到各个关节的像素坐标。将像素坐标与深度值结合，得到各个关节点的三维坐标，再将三维坐标表示的各个关节点按顺序相连就能得到三维骨架。但是深度值表示物体到摄像机的距离，它是有量纲的，需要先将深度值归一化，归一化后的深度值：

$$zn=\frac{(dn-\min dk)}{(\max dk-\min dk)} \qquad k,n=1,2,\cdots,N \tag{5}$$

然后还要将像素坐标转换为世界坐标。坐标转换的过程中深度值是不变的，令 $z(n)=Z$，Kinect 设备内部的芯片已经对红外摄像头进行了标定，经过标定后摄像机可以看做是理想的针孔摄像头的成像模型，根据相似三角形变换原理：

$$\frac{X}{Z}=\frac{x}{f},\frac{Y}{Z}=\frac{y}{f} \tag{6}$$

从中可以计算出世界坐标系的 X、Y 值，从而得到三维坐标(X,Y,Z)。其中 f 表示摄像机的焦距，x，y 是像素坐标。

图 3 为腐蚀膨胀方法确定身体范围及关节点的效果图。图 4 为三维骨架。对于采集到的深度图像，重复以上步骤，建立三维骨架数据库。

(a)去背景 (b)腐蚀 (c)膨胀 (d)肩和髋节点 (e)关节点

图 3 确定身体范围和关节点

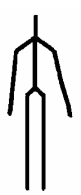


图 4 三维骨架

2.4 相似性度量

运动序列的匹配过程中序列的长度有可能不一样，欧式距离等经典的距离度量方法只能用于单个帧动作的匹配，对于序列的匹配不适用。而动态时间规整（DTW）它是一种很好的用于全局时间域对齐的方法，常用于视频序列、运动序列的匹配问题[12]。所以采用 DTW 的方法进行序列之间的相似性度量。

使用 DTW 算法的原理是：对于 2 个动作序列 X：$x_1,\cdots,x_m$ 和 Y：$y_1,\cdots,y_n$，m、n 分别表示 2 个动作序列的长度，首先要计算 X 和 Y 的各个分量之间的距离，形成一个 $m\times n$ 的矩阵，然后从(1,1)到(m,n) 进行遍历，计算得到一条具有最短距离的路径。最短距离所在的路径就是最优路径，也就是最优匹配序列。

2.4.1 两帧动作之间的相似性度量

这里单个帧动作之间的相似性度量就是指轮廓模型之间的度量。2 个轮廓之间的相似性距离可以用归一化的傅里叶描述子 $FD(k)$ 来计算。采用欧氏距离计算其相似性距离可定义为：

$$dist\,i,j=\sqrt{\sum_{k=0}^{M-1}\|FD_ik-FD_jk\|^2} \tag{7}$$

如果两轮廓 i,j 之间的相似度越高，$dist$ 的值越小；反之，$dist$ 的值越大。当 i,j 完全相同时，$dist=0$；当 i,j 完全不相同时，$dist$ 的值是非常大的。

2.4.2 序列之间的相似性度量

对于 2 个长度分别为 m、n 的动作序列 X、Y，它们通过 DTW 算法计算的距离：

$$D(i,j)=dist(i,j)+\min[D(i,j-1),Di-1,j,Di-1,j-1] \tag{8}$$

式中：$i=1,\cdots,m,j=1,\cdots,n,D(i,j)$ 的初值在 i 或 j 小于 1 时都为 0。$D(i,j-1)$，$D(i-1,j)$ 和 $D(i-1,j-1)$ 表示 3 个方向局部路径约束的值，即当从一个点$(i-1,j-1)$或$(i-1,j)$或$(i,j-1)$到下一个点(i,j)时，如果是横着或者竖着到达的话其距离为 $D(i,j)$，如果是斜着过来的则是 $2D(i,j)$。

3 实验结果

本文使用 Kinect 采集深度图像，在 MATLAB 平台上获取轮廓信息、提取三维骨架、测量运动序列之间的相似性距离，从而实现运动捕获的结果。采集了人体常做的 50 个动作序列，经过处理得到轮廓数据库和骨架数据库，2 个数据库是一一对应的。系统输入一个动作序列的深度图，经过去除背景、提取人体动作轮廓，然后与轮廓数据库中的动作序列进行相似性匹配，计算出距离，最短距离所对应的动作序列就是最优匹配序列，系统输出骨架数据库中与之对应的骨架作为运动捕获的结果。为了测试运动捕获结果的准确性，本文进行了一些实验，图 5 为 2 组实验结果。

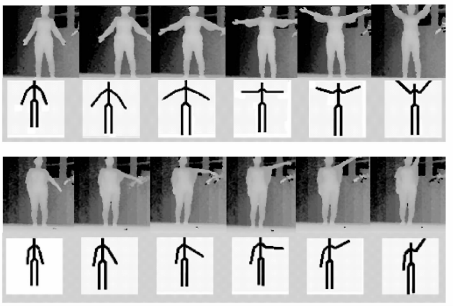


图5 2组捕获结果

使用准确程度来度量捕获的效率。准确程度描述的是最终捕获输出的骨架序列与输入的深度图的动作序列的相似程度。准确程度的度量标准是:记录运动序列的总帧数n,统计出捕获结果中动作误差明显大的帧的个数m,然后计算$p=(n-m)/n$,p就是对动作捕获的准确程度,p值越大,说明捕获的准确度越高。规定3组运动,由2个不同体型的人来做,每组运动有6~10帧动作不等,然后对每个人的每组运动捕获的准确度进行计算,实验结果如表1所示,从表中可以看出本文动作捕获的准确度较高,稳定性也比较好。

表1 实验结果比较

	运动1	运动2	运动3
表演者1	$p=83.3\%$	$p=80\%$	$p=75\%$
表演者2	$p=85.7\%$	$p=75\%$	$p=78.7\%$

4 结 论

详细介绍了利用Kinect采集深度图实现人体动作捕获的关键技术,进行了融合深度图像和三维模型的动作捕获技术的研究,并详细叙述了整体的研究和实现工作。通过实验证明轮廓信息能够较好地反映目标的形状,从而提高动作识别的准确性,使运动捕获的效率提高,并且计算量较小。三维深度图像则保证了信息的完整性和可靠性,提高了系统的鲁棒性。但是DTW的序列匹配方法在时间和空间复杂度上都是二次方的,效率不高,所以在今后的研究中可以考虑提高匹配的效率。

参 考 文 献

[1] WANG X, XIAO B X, GUO X Y. Human-like character animation of maize driven by motion capture data[J]. Information & Computational Science, 2011, 8(2): 345-353.

[2] SHIRATORI T, HODGINS J K. Accelerometer-based user interfaces for the control of a physically simulated character[J]. ACM Transactions on Graphics, 2008, 27(5): 1-9.

[3] XIAO ZH D, NAIT-CHARIF H, ZHANG J J. Automatic estimation of Skeletal motion from optical motion capture data [J]. Computer Science, 2008, 5277: 144-153.

[4] BLACKBURN J, RIBEIRO E. Human motion recognition using isomap and dynamic time warping[C]. Proceedings of the 2nd Conference on Human Motion: Understanding, Modeling, Capture and Animation, 2007: 285-298.

[5] SHOTTON J, SHARP T, KIPMAN A, et al. Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1): 116-124.

[6] ISHIGAKI S, WHITE T, ZORDAN VB, et al. Performance-based control interface for character animation[J]. ACM Transactions on Graphics, 2009, 28(3): 1-8.

[7] GANAPATHI V, PLAGEMANN C, KOLLER D, et al. Real time motion capture using a single time-of-flight camera[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2010:755-762.

[8] 陈健,郑绍华,余轮,等.基于方向的多阈值自适应中值滤波改进算法[J].电子测量与仪器学报,2013,27(2):156-161.

[9] 刘佳,付伟平,王雯,等.基于改进SIFT算法的图像匹配[J].仪器仪表学报,2013,34(5):1107-1112.

[10] 李红岩,毛征,袁建建,等.一种基于算法融合的运动目标跟踪算法[J].国外电子测量技术,2013,32(12):36-40.

[11] PLAGEMANN C, GANAPATHI V, KOLLER D, et al. Real-time identification and localization of body parts from depth images[C]. In Proceedings of ICRA, 2010. 1, 2, 7.

[12] LINDASALWA M, MUMTAJ B, ELAMVAZUTHI I. Voice recognition algoithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping(DTW) techniques[J]. Journal of Computing, 2010,2(3):138-143.

作 者 简 介

肖秦琨,1974年出生,博士,教授,主要研究方向为图像检索、三维模型描述及学习,目标检测,先进控制理论及应用。

谢艳梅(通讯作者),1989年出生,硕士,研究生,主要研究方向为通信与信息系统、信息检索、图像处理。
E-mail:1023417886@qq.com