DOI: 10.13718/j.cnki.xdzk.2015.08.022

基于 Kinect 传感器的动态手势实时识别①

刘　瑶,　余　旭,　黄智兴

西南大学 计算机与信息科学学院,重庆 400715

摘要: 本文提出了一种基于 Kinect 传感器的动态手势实时识别方法,在预处理阶段通过 OpenNI/NITE 快速获取人体骨架,并从中得到关节点数据用于建立动态手势的运动轨迹特征描述子.提出了一种在全局约束条件下的权重化多维数据的动态时间扭曲算法(WM-DTW)来对手势轨迹序列进行训练和识别.实验结果表明,本文的识别方法比 LDA 算法和动态时间扭曲(DTW)算法有更高的识别率.

关　键　词: Kinect 传感器;深度数据;手势识别;动态时间扭曲

中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673-9868(2015)08-0132-06

手势是人机交互中最为自然且直观的控制方式,如今基于视觉的手势识别已经成为了实现新一代人机交互的关键性技术.基于视觉的识别,是指计算机对图像颜色、形状、深度等特征进行处理,从而达到识别、索引、模拟、分割等目的[1-4].然而,由于手势本身具有多样性、多义性以及时间和空间上的相异性等特点,基于视觉信息的手势识别变得愈加困难[5-6].

2010 年微软公司体感摄像头 Kinect 面世,在后续不到 1 年的时间里,基于该设备的系统被研发出来.J. L. Raheja 的文章[7]中讲述了一种利用 Kinect 跟踪手指尖和手掌中心的方法,该方法基于深度数据[8]的阈值找出手部区域,在该区域内找到深度值最大的点即为手掌的中心点.利用预先设定的阈值与深度数据进行比较,去掉手掌,保留手指,从而找到指尖.G. F. He 等人[9]在基于 Kinect 深度数据分割出手部区域后,利用 Graham Scan 算法[10]找出手部区域的凸包集合,从而找出凸点,通过计算候选点间的夹角度数来辨别各个手指.Y. Li[11]在其基础上提出了一种 3 点对齐的方法进一步找到更精确的指尖,并构造出一种三层分类器来识别出一些常见的手语字母.Z. Ren 等人[12-13]提出了一种改进的地球移动距离(Earth Mover's Distance, EMD)来匹配手势,并将其应用到手势的加减乘除运算和石头剪刀布的游戏中去.C. Yang等人[14]提出了基于 Kinect 获取深度信息来识别手势的系统,并将其应用在了多媒体播放器的控制上.2012 年香港中文大学设计了一种基于 SVM 方法的手语翻译和识别系统,利用 Kinect 捕获的人体关节点作为训练器输入,并利用 SVM 对提取的特征向量进行分类.2013 年微软教育峰会上 X. Chai 等人提出了一种基于轨迹匹配方法识别手语的翻译系统,利用 Kinect 捕获手语的 3D 轨迹特征,线性重采样方法将数据正规化并作为轨迹匹配的输入,但缺点是该系统需要进行大量的专家输入和训练.

本文工作主要有以下 2 部分组成:第一,结合运动关键骨骼点轨迹匹配对动态手势进行识别;第二,提出了一种在全局约束条件下的多权重多维数据时间扭曲算法,给出了全新的代价函数定义方式.实验结果表明,本文所提出的方法的识别准确率达到了 89.64%,且实时性较高.

---

① 收稿日期: 2014-04-16

基金项目: 国家自然科学基金(61372138);中央高校基本科研业务基金(No. SWU1309265, No. XDJK2014B012);重庆市自然科学基金(CSTC2012JJB40012).

作者简介: 刘　瑶(1991-),女,辽宁辽中人,硕士研究生,主要从事模式识别的研究.

通信作者: 黄智兴,副教授.

1 实验预处理

1.1 骨骼追踪与手势关键关节点

Kinect获取的深度数据以像素为记录单位，默认设置为640×480像素，转化为点云即有307 200个点，这样配置下摄像头每一秒中大约就会产生17.6M大小的数据，在这些原始数据中包含很多无用的信息(比如背景信息等)，要处理或者大量存储这样的数据显然不合实际，因此首先应从原始数据中找出有用的手部信息. 本文利用OpenNI/NITE来跟踪人体的关节点，从而定位用户手部的位置. 经过分析，只有6个点对于本文中运动手势的描述子建立是至关重要的，分别是：左手(LH)、右手(RH)、左肘(LE)、右肘(RE)、左肩(LS)、右肩(RS). 本文读取用户骨骼的流程图如图1所示.

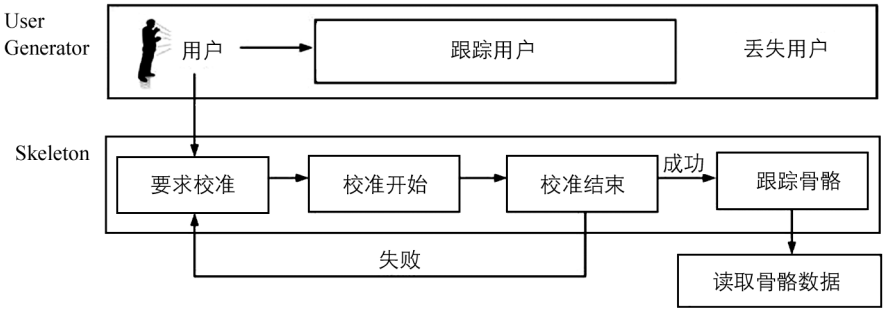


图1 骨骼追踪流程图

1.2 手部数据的标准化

分析用户运动的整个过程可以发现，在用户位置变化的过程中虽然用户的骨架大小和兴趣点在笛卡尔坐标系上的值在发生变化，但是对于同一用户而言，关节兴趣点与用户的躯干关节点的相对位置基本不变. 因此利用球面坐标系代替笛卡尔坐标系来保证用户位置在变化过程中的相对性. 但基于多空间分割的球面坐标系运动特征描述子计算过于繁琐[15]，因此，本文仅考虑通过球面坐标系对用户手部数据进行标准化.

球面坐标系是三维坐标系的一种，以坐标原点作为参考点(本文中坐标原点定义为躯干节点)，由距离、方位角和仰角构成. 假设$p(x, y, z)$为空间中一点，点p可以用3个有序数(r, θ, φ)确定. 其中r为球面点到原心O的距离($r\in[0, +\infty)$)，θ为仰角，是有向线段OP与z轴正方向间的夹角($\theta\in[0, \pi]$，本文中$\theta\in[0, \pi/2]$)，φ称为方位角，是x正半轴逆时针转到OM的夹角($\varphi\in[0, 2\pi]$，M是p在XY平面内的投影).

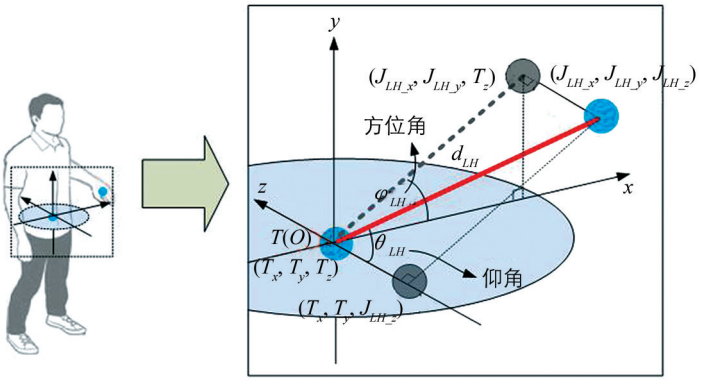


图2 用户左手关节点球面坐标示例

根据上述定义，将文中人体关节点转到球面坐标系，由于躯干关节点T在整个运动过程中对于人体其他关节的相对位置保持不变，所以将T视为球面坐标系中的原点O. 由此计算出其余各个关节点的球面坐标位置. 以用户左手关节点LH为例(图2)，躯干关节$T(T_x, T_y, T_z)$为球形坐标系中的原点O，假设左手关节点的坐标为$LH(J_{LHx'}, J_{LHy'}, T_z)$，则该点在$XY$平面中的投影$M$为$(J_{LH_x'}, J_{LH_y'}, T_z)$，由此可以计算出点$LH$的球面坐标$S_{LH}(\gamma, \theta, \varphi)$，即

$$\begin{cases} d_{LH} = \sqrt{(J_{LH\_x} - T_x)^2 + (J_{LH\_y} - T_y)^2 + (J_{LH\_z} - T_z)^2} = \gamma \\ \theta = \arctan\left(\sqrt{(J_{LHx} - T_x)^2 + (J_{LHy} - T_y)^2} / (T_z - J_{LH\_z})\right) \\ \varphi = \arctan\left((J_{LHy} - T_y)/(J_{LH\_x} - T_x)\right) \end{cases} \tag{1}$$

将用户位置标准化后，每个关节点用它相对于T点的距离和与这个距离相关的2个角度θ，φ表示. 给定用户躯干点投影坐标$T(T_x, T_y, T_z)$，手势关键关节点集合$JS=\{J_{LH}, J_{RH}, J_{LE}, J_{RE}, J_{LS}, J_{RS}\}$，则关节点球面坐标$S_{LH}(\gamma, \theta, \varphi)$为

$$\begin{cases}\gamma = d(j)_{j\in JS} = \sqrt{(j_x - T_x)^2 + (j_y - T_y)^2 + (j_z - T_z)^2} \\ \theta = \theta(j)_{j\in JS} = \arctan(\sqrt{(j_x - T_x)^2 + (j_y - T_y)^2}/(T_z - j_z)) \\ \varphi = \varphi(j)_{j\in JS} = \arctan((j_y - T_y)/(j_x - T_x))\end{cases} \tag{2}$$

获取到关节点的坐标信息后，利用该信息来建立手势在运动过程中的特征描述子.

2　动态手势的特征建立和识别

2.1　运动轨迹描述子的建立

图 3 给出了本轨迹描述子建立的示意过程. 每个运动手势由一个个手势帧组成，即帧序列. 文中对每一帧构造一特定的轨迹描述子来代表该帧的手势特征，那么运动手势则可以通过帧轨迹描述子的序列表示. 其中，每帧轨迹描述子用 3 个层次构造起来分别对应 r 层、θ 层和 φ 层，每层中的数据即为该帧中关键关节点对应的球坐标值. 每帧捕获的关节点有 6 个(手，肘，肩)，每个节点对应 r，θ，φ 这 3 个值，因此每帧轨迹描述子为一个三维数据，且每维中含有 6 个坐标值.

2.2　动态手势的识别

动态手势识别常用的方法有隐马尔可夫(Hidden Markov Model，HMM)[16]序列匹配方法、支持向量机(Support Vector Machine，SVM)的训练分类方法、基于图理论匹配[17]和动态时间扭曲(Dynamic Time Wrapping，DTW)[18]等方法.

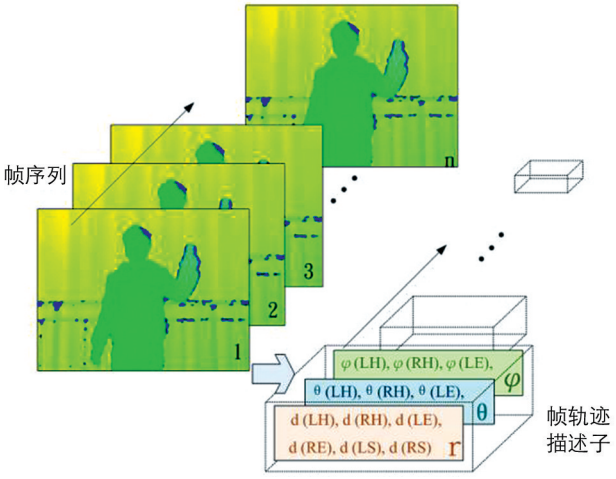


图 3　运动轨迹描述子

传统的动态时间扭曲算法能很好地解决 2 个不同长度序列的匹配问题，而且思路清晰，训练时不需要额外的数据，但存在以下问题：问题 1，DTW 的思想是基于动态规划的(Dynamic Programming，DP)，在算法执行的过程中需要不断的计算迭代，运算量会很大，如果待比较的 2 个序列很长，那么计算迭代量会变大；问题 2，对于运动手势匹配常利用建立公共模板库的方法，通过匹配模板库将动作转化为索引序列，并将此序列作为 DTW 的输入. 但在新添加手势时公共模板库会迅速增加，使动作转为索引序列的计算量变大.

针对上述问题本文提出了一种权重化多维数据动态扭曲算法(Weight Multi-dimension data DTW，WM-DTW). 给定 2 个帧数据 F_1，F_2，且这 2 帧分别属于 2 个动作序列，给定序列中的 2 个点数据 P_1 和 P_2，每点包含 6 个关节点数据，假设每点有 k 维，则 2 点间的距离为

$$\text{distance}'(P_1, P_2) = \sqrt{\sum_{j}^{j\in JS}[d(\gamma) + d(\theta) + d(\varphi)]} \tag{3}$$

利用明氏距离来定义 r，θ，φ 间的差异函数，

$$d^{(p)}(x, y) = [\sum_{i=1}^{d} |x_i - y_i|^p]^{1/p} \tag{4}$$

则

$$\text{distance}'(P_1, P_2) = \sqrt{\sum_{i=1}^{3}\sum_{j}^{j\in JS}(P_{1,j,i} - P_{2,j,i})^p} \tag{5}$$

即

$$\text{distance}'(P_1, P_2) = \sqrt{\sum_{j}^{j\in JS}[(P_{1,j,r} - P_{2,j,r})^p + (P_{1,j,\theta} - P_{2,j,\theta})^p + (P_{1,j,\varphi} - P_{2,j,\varphi})^p]} \tag{6}$$

r 和 θ 可以直接使用差异函数 $d(r)$ 和 $d(\theta)$ 来计算不同元素间的差异，但是 φ 的差异函数不能简单的利用差的平方计算，在计算 $d(\varphi)$ 前需要进行判断，即

$$d'(\varphi)=\begin{cases}\sum_{j}^{j\in JS}(|P_{1,j,\varphi}-P_{2,j,\varphi}|)^{p} & \text{当 } P_{1,j,\varphi} \text{ 和 } P_{2,j,\varphi} \text{ 角度都小于 } \pi \text{ 时;}\\ \sum_{j}^{j\in JS}(\pi-|P_{1,j,\varphi}-P_{2,j,\varphi}|)^{p} & \text{其他情况的时候;}\end{cases} \tag{7}$$

为了使代价函数计算出的值更能代表实际情况，还需要考虑人体各个关节点在计算中占据的权重大小问题. 对各个关节点赋予不同的权重值，则序列中点间的代价距离函数表示为

$$\text{distance}''(P_1, P_2)=\sqrt{\sum_{j}^{j\in JS}[d(\gamma)+d(\theta)+d'(\varphi)]\times W_j} \tag{8}$$

其中 W_j 代表不同关节点的权重，且所有 W_j 的和为1，因为人体的左手和右手是对称的关系，所以可以将 W_{LH}，W_{RH}，W_{LE}，W_{RE}，W_{LS}，W_{RS} 这6个权重变为3个权重值，W_H，W_E，W_S 分别对应手、肘、肩关节点权重值，且 $W_H+W_E+W_S=1$. 综上所述，在全局约束条件下权重化多维数据动态时间扭曲算法流程如下所示：

算法 1 权重化多维数据动态时间扭曲算法

输入：2个多维数据帧序列，$A[F_1\cdots F_n]$，$B[F_1\cdots F_n]$；

输出：扭曲路径距离 $\text{DTW}'[n, m]$；

定义：$\text{distance}''(A[i], B[j])$为点 i，j 间的距离公式

Begin int $\text{DTW}[0\cdots n, 0\cdots m]$；

int i，j，distance；

For $i=1$ to m do

　　$\text{DTW}[i, 0]\leftarrow\infty$；

End For

For $j=1$ to n do

　　$\text{DTW}[0, j]\leftarrow\infty$；

End For

　　$\text{DTW}'[0, 0]\leftarrow\text{distance}''(A[0], B[0])$

For $i=1$ to n do

For $j=1$ to m do

While $(2i-j>=3)$ and $(2j-i>=2)$ do

　　$\text{distance}''(A[i], B[j])$；

$$\text{DTW}'[i, j]\leftarrow\min\begin{cases}\text{DTW}'[i, j-1]+\text{distance}''(A[i], B[j])\\ \text{DTW}'[i-1, j]+\text{distance}''(A[i], B[j])\\ \text{DTW}'[i-1, j-1]+2\text{distance}''(A[i], B[j])\end{cases}$$

End While

End For

End For

Return $\text{DTW}'[n, m]/(n+m)$；

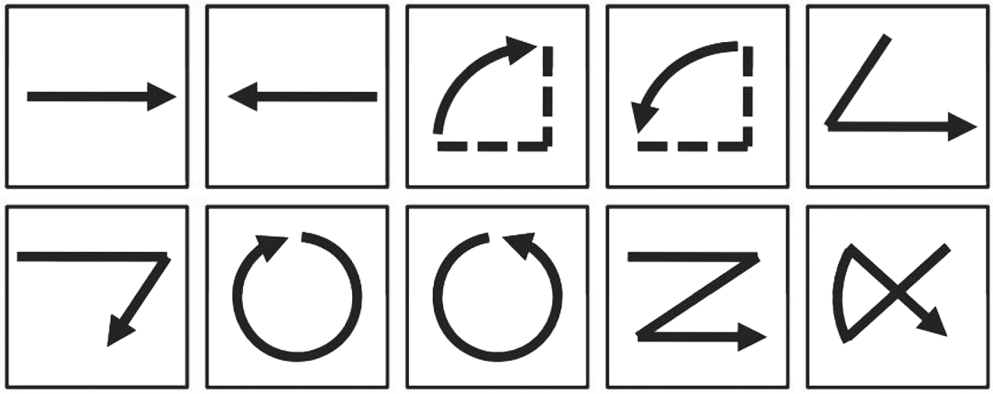
End

3 实验结果与分析

为测试动态手势的效果，使用动态轨迹手势模板库 DMT 中的10种轨迹作为本次识别的模板，定义的手势如图4所示.

本次实验采集8个人的动态手势轨迹，每人每种手势做4次输入，其中的7位用户所做手势作为测试集，测试集中测试样本总数目为280(7×10×4)个，剩余的一位同学所做手势加入训练集中，作为标准手势样本. 经过多次试验，当 $W_H=0.7$，$W_E=0.2$，$W_S=0.1$ 时识别效果达到最佳状态.

随后，对明氏距离参数 p 进行讨论. 可知，当 $p=1$ 时，计算的是输入的2个元素间的曼哈顿距离；当 $p=2$ 时，计算的是输入的2个元素间的欧式距离；当 $p=\infty$ 时，计算的是输入的2个元素间的切比雪夫距离. 随着参数 p 值的增加，其差异函数值越来越小，选 p 值为1～4，其识别结果如下表1所示. 可知，当 $p=2$ 时，手势的识别率最高.



（依次为）右移，左移，上旋转，下旋转，下右，右下，顺时针，逆时针，Z 字形，交叉.

图 4　动态轨迹手势模板库 DMT

表 1　不同 p 值的手势识别率

p 值	1	2	3	4
识别率/%	59.23	89.64	80.21	75.55

为了对比分析文中所提框架的正确性和效率，文中对比了额外的 2 种动作识别方法. 方法 1 为基于 LDA 和词袋分类的方法，实验框架借鉴文献[19]所提的方法. 并简化对比实验流程，采用 SIFT＋K-means＋LDA 的识别方法. 方法 2 选取传统的动态时间扭曲(DTW)方法. 从表 2 数据分析可知，在 DMT 测试数据集上使用本文的方法达到最大的总平均识别率 89.64%.

表 2　DMT 上用户每种手势的平均识别率

手势	方法 1：SIFT＋K-means＋LDA		方法 2：DTW		本文方法		手势	方法 1：SIFT＋K-means＋LDA		方法 2：DTW		本文方法	
	识别数	识别率/%	识别数	识别率/%	识别数	识别率/%		识别数	识别率/%	识别数	识别率/%	识别数	识别率/%
	27	96.43	28	100	28	100		25	89.29	26	92.86	27	96.43
	26	92.86	24	85.71	24	85.71		25	89.29	27	96.43	27	96.43
	18	64.29	16	57.14	20	71.43		24	85.71	26	92.86	27	96.43
	20	71.43	22	78.57	22	78.57		23	82.14	22	78.57	26	92.86
	24	85.71	22	78.57	24	85.71		22	78.57	24	85.71	26	92.86

注：总平均识别率，方法 1：83.57%；方法 2：84.64%；本文方法：89.64%.

4　总　结

目前基于三维数据信息的手势研究偏向于大量数据的计算，需要花费大量时间和对计算机硬件的要求过高，针对此问题本文提出了一种全局约束条件下的权重化多维数据动态时间扭曲方法. 利用多维数据帧描述子作为手势的基本特征，将运动手势转换为多维数据的帧轨迹特征描述子序列，再将问题转化为帧轨迹特征描述子序列的匹配. 实验结果表明，本文动态手势识别有较高的识别率. 本文进一步的工作将在多个 Kinect 上展开，以期达到更高的识别精度.

参考文献：

[1]　郭士会，杨　明. 基于颜色的图像检索方法的研究 [J]. 西南大学学报：自然科学版，2012，34(1)：128－133.
[2]　胡伟平，邓辉文. 个性化人脸图像模拟识别 [J]. 西南大学学报：自然科学版，2014，36(2)：178－185.
[3]　胡钦瑞，肖国强. 基于粗糙集和 MRF 的彩色图像分割方法 [J]. 西南师范大学学报：自然科学版，2014，39(4)：113－119.

[4] 施成湘. 基于二次分水岭和近邻传播聚类的彩色图像分割算法研究与实现[J]. 西南师范大学学报：自然科学版，2013，38(8)：125－129.
[5] 任海兵，祝远新，徐 光，等. 基于视觉手势识别的研究[J]. 电子学报，2000，28(2)：118－121.
[6] 胡友树. 手势识别技术综述[J]. 中国科技信息，2005(2)：41－42.
[7] RAHEJA J L，CHAUDHARY A，SINGAL K. Tracking of Fingertips and Centers of Palm Using Kinect[C]//2011 Third International Conference on Computational Intelligence Modeling Simulation. Langkawi：IEEE，2011：248－252.
[8] SHOTTON J，FITZGIBBON A W，COOK M，et al. Real-Time Human Pose Recognition in Parts from Single Depth Images[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Providence：IEEE Computer Society，2011：1297－1304.
[9] HE G F，KANG S K，SONG W C，et al. Real-Time Using Gesture Recognition Using 3D Depth Camera[C]// Software Engineering and Service Science，2011 2nd IEEE International Conference on. Beijing：IEEE，2011：187－190.
[10] GRAHAM R L. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set[J]. Information Processing Letters，1972，1(4)：132－133.
[11] LI Y. Hand Gesture Recognition Using Kinect[C]// Software Engineering and Service Science (ICSESS)，2012 IEEE 3rd International Conference on. Louisville，Kentucky：IEEE，2012：196－199.
[12] REN Z，YUAN J S，ZHANG Z Y. Robust Hand Gesture Recognition Based on Finger-Earth Mover's Distance with a Commodity Depth Camera[C]// Proceedings of the 19th International Conference on Multimedia. Scottsdale，Arizona，US：ACM，2011：1093－1096.
[13] 郑丹晨，韩 敏. 基于H-EMD的形状上下文特征形状匹配方法[J]. 控制与决策，2012，27(11)：1639－1643.
[14] YANG C，JANG Y，BEH J，et al. Gesture Recognition Using Depth-Based Hand Tracking for Contactless Controller Application[C]// Consumer Electronics (ICCE)，2012 IEEE International Conference on. Las Vegas，NV，USA：IEEE. 2012：297－298.
[15] YE M，ZHANG Q，WANG L，et al，A Survey on Human Motion Analysis from Depth Data[J]. Time-of-Flight and Depth Imaging. 2013，8200：149－187.
[16] LV F，NEVATIA R. Single View Human Action Recognition Using Key Pose Matching and Viterbi Path Searching[C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition，Oregon Convention Center in Portland. Oregon：IEEE. 2007：1－8.
[17] PANSARE J R，BANSAL M，SAXENA S，et al. Gestuelle：a System to Recognize Dynamic Hand Gestures Using Hidden Markov Model to Control Windows Applications[J]. International Journal of Computer Applications，2013，62(17)：19－24.
[18] 邹 洪. 实时动态手势识别关键技术研究[D]. 广州：华南理工大学，2011.
[19] NIEBLES J C，WANG H C，LI F F. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words[J]. International Journal of Computer Vision，2008，79(3)：299－318.

Real-Time Dynamic Hand Gesture Recognition Based on Kinect Sensor

LIU Yao， YU Xu， HUANG Zhi-xing
School of Computer and Information Sciences，Southwest University，Chongqing 400715，China

Abstract：A reliable method for dynamic gesture recognition based on Kinect sensor is proposed in this paper. Firstly，the human skeleton is extracted by using OpenNI/NITE toolbox in the pre-processing stage. Then the feature descriptor of hand area is established for dynamic gesture trajectory. Finally，an algorithm of dynamic time warping called WM-DTW is proposed to handle the weighted multi-dimensional data under the global constraint for trajectory sequence matching. The experimental results show that our WM-DTW method improves the recognition accuracy compared with Latent Dirichlet Allocation (LDA) and conventional Dynamic Time Warping (DTW) methods.

Key words：Kinect sensor；depth data；hand gesture recognition；dynamic time warping (DTW)

责任编辑 崔玉洁