|  |  |
| --- | --- |
| IEEE SENSORS JOURNAL, VOL. 16, NO. 3, FEBRUARY 1, 2016 16，no. 3，FEBRUARY 1,2016 | 773 |

A Real-Time Human Action Recognition System Using Depth and Inertial Sensor Fusion

基于深度和惯性传感器融合的实时人体动作识别系统

Chen Chen, *Student Member, IEEE*, Roozbeh Jafari, *Senior Member, IEEE*,

陈，学生会员，IEEE，Roozbeh Jafari，IEEE 高级会员,

and Nasser Kehtarnavaz, *Fellow, IEEE*

和 Nasser Kehtarnavaz，研究员，IEEE

***Abstract*— This paper presents a human action recognitionsystem that runs in real time and simultaneously uses a depth camera and an inertial sensor based on a previously devel-oped sensor fusion method. Computationally efficient depth image features and inertial signals features are fed into two computationally efficient collaborative representative classifiers.**

***本文提出了一种基于先进的传感器融合方法，同时使用深度摄像机和惯性传感器的实时运行的人体动作识别系统。将计算效率高的深度图像特征和惯性信号特征输入到两个计算效率高的协同代表性分类器中。***

1. **decision-level fusion is then performed. The developed real-time system is evaluated using a publicly available multimodal human action recognition data set by considering a comprehensive set of human actions. The overall classification rate of the developed real-time system is shown to be** *>***97%, which is at least 9% higher than when each sensing modality is used individually. The results from both offline and real-time experimentations demonstrate the effectiveness of the system and its real-time throughputs.**

**然后进行决策级融合。开发的实时系统是评估使用一个公开可用的多模式人类行动识别数据集，考虑到一个全面的人类行动集。实验结果表明，所开发的实时系统的总体分类率大于97% ，比单独使用每种传感方式至少高出9% 。离线和实时测试的结果表明了该系统的有效性和实时吞吐量。**

***Index Terms*— Human action recognition, real-time humanaction recognition system, depth camera sensor, wearable inertial sensor, sensor fusion.**

***索引术语ー人体动作识别、实时人体动作识别系统、深度相机传感器、可穿戴惯性传感器、传感器融合。***

1. INTRODUCTION

引言

**H**UMAN action recognition is finding its way into commercial products and is of benefit to many human-computer interface applications. Example applications include hand gesture interaction, smart assistive living, and gaming. Different sensors have been used to perform human action recognition. These sensors include conventional RGB cameras, e.g. [1]–[3], depth cameras, in particular Kinect, e.g. [4]–[7],

人体动作识别技术正在商业化生产中得到广泛应用，对于许多人机交互应用具有重要意义。例如应用程序包括手势互动，智能辅助生活和游戏。不同的传感器已被用于执行人的行为识别。这些传感器包括传统的 RGB 相机，例如[1]-[3] ，深度相机，特别是 Kinect，例如[4]-[7] ,

and inertial sensors, e.g. [8]–[10].

和惯性传感器，例如[8]-[10]。

In our previous works [11]–[13], it was shown that improvements in recognition rates can be achieved by com-bining or fusing the information from a depth camera and an inertial sensor over the situations when each of these sensors is used individually due to the complementary aspect of the information provided by these two differing modality sensors. In [13], we reported a human action recognition method which involved the development of depth motion map features and the utilization of a collaborative representation classifier. How-ever, the experimental analysis reported in [13] was conducted

在我们以前的工作[11]-[13]中，由于这两个不同的模态传感器所提供的信息互补，将来自深度相机和惯性传感器的信息相结合或融合，可以提高识别率。在[13]中，我们报道了一种人类行为识别方法，该方法涉及到深度运动图特征的开发和协同表示分类器的使用。[13]中报道的实验分析究竟是如何进行的

Manuscript received June 30, 2015; revised September 9, 2015; accepted October 2, 2015. Date of publication October 5, 2015; date of current version January 12, 2016. This work was supported by the National Sci-ence Foundation under Grant CNS-1150079. The associate editor coor-dinating the review of this paper and approving it for publication was Prof. Danilo Demarchi.

手稿收到2015年6月30日; 修订2015年9月9日; 接受2015年10月2日。出版日期: 2015年10月5日; 当前版本: 2016年1月12日。这项研究得到了 CNS-1150079国家科学基金会的支持。负责协调本文审查并批准发表的副主编是 Danilo Demarchi 教授。

C. Chen and N. Kehtarnavaz are with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: chenchen870713@gmail.com; kehtar@utdallas.edu).

C. Chen 和 n. Kehtarnavaz 供职于德克萨斯大学电气工程系，位于 TX 的达拉斯。

R. Jafari is with the Biomedical Engineering, Computer Science and Engineering, and Electrical and Computer Engineering Departments, Texas A&M University, College Station, TX 77843 USA (e-mail: rjafari@tamu.edu).

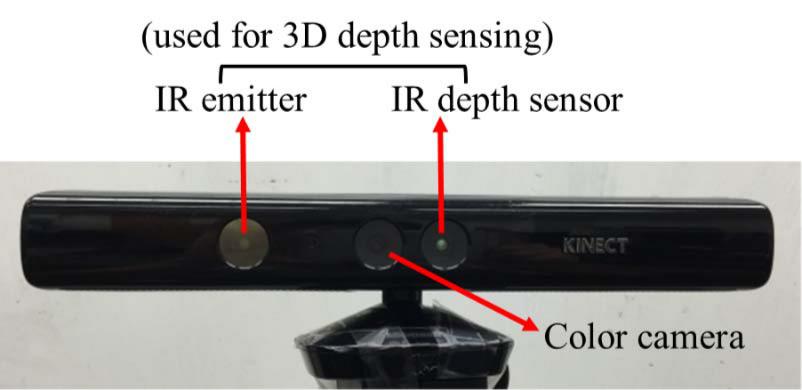
本文作者 r. Jafari 供职于美国生物医学工程、计算机科学与工程、电气与计算机工程系，德州农工大学，TX 77843。

Digital Object Identifier 10.1109/JSEN.2015.2487358

DOI 10.1109/JSEN. 2015.2487358

Fig. 1. Microsoft Kinect depth sensor.

图1. 微软 Kinect 深度传感器。



based on the data that were collected simultaneously from the sensors. In this paper, we have made modifications to the method reported in [13] in order to produce a human action recognition system which runs in real-time. These modifica-tions include (i) adding a module to automatically detect the start and end of an action in real-time, (ii) modifying the fusion approach to reduce the computational complexity for real-time operation, (iii) carrying out extensive experimentations in offline and real-time manner for both subject-generic and subject-specific scenarios.

基于同时从传感器收集到的数据。本文对文献[13]中的方法进行了改进，构建了一个实时运行的人体动作识别系统。这些改进包括(i)增加一个模块来实时自动检测一个动作的开始和结束，(ii)改进融合方法以降低实时操作的计算复杂性，(iii)对主题通用和主题特定的情景进行大量的离线和实时实验。

The rest of the paper is organized as follows. In section II, an overview of the sensors and techniques used in our fusion method is provided. In section III, the modifications made in order to produce a real-time human action recognition system are presented. The experimental results for both offline and real-time recognition are included in section V. Finally, the conclusion appears in section VI.

文章的其余部分组织如下。在第二部分，综述了我们的融合方法中使用的传感器和技术。第三部分介绍了为实现实时人体动作识别系统所做的改进。第五节介绍了离线识别和实时识别的实验结果。最后，结论出现在第六节。

1. OVERVIEW OF SENSOR FUSION METHOD

传感器融合方法综述

In this section, an overview of the sensors and techniques used in our fusion method in [13] is stated so that the stage is set for the modifications made in the next section towards enabling real-time operation.

在这一节中，概述了在我们的融合方法中使用的传感器和技术[13] ，以便为下一节中为实现实时操作所做的修改设置阶段。

*A. Sensors*

*传感器*

Kinect is a low-cost RGB-Depth camera sensor introduced by Microsoft for human-computer interface applications. It comprises a color camera, an infrared (IR) emitter, an IR depth sensor, a tilt motor, a microphone array, and an LED light. A picture of the Kinect sensor or depth camera is shown in Fig. 1. This sensor can capture 16-bit depth images with a resolution of 320×240 pixels. Two example depth images are depicted in Fig. 2. The frame rate is approximately 30 frames

Kinect 是微软为人机界面应用而推出的一种低成本 rgb 深度摄像头传感器。它包括一个彩色照相机，一个红外线发射器，一个红外线深度传感器，一个倾斜马达，一个麦克风阵列，和一个 LED 灯。图1显示了 Kinect 传感器或者深度相机的图片。该传感器可以捕捉16位的深度图像，分辨率为320240像素。图2中描绘了两幅深度图像的例子。帧速率大约是30帧

1558-1748 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

1558-17482015 IEEE. 允许个人使用，但重新发布/再分发需要 IEEE 许可。

See http://www.ieee.org/publications\_standards/publications/rights/index.html for more information.

更多信息请参见 http://www.ieee.org/publications\_standards/publications/rights/index.html。

774 IEEE SENSORS JOURNAL, VOL. 16, NO. 3, FEBRUARY 1, 2016

774 IEEE SENSORS JOURNAL，VOL. 16，no. 3，FEBRUARY 1,2016

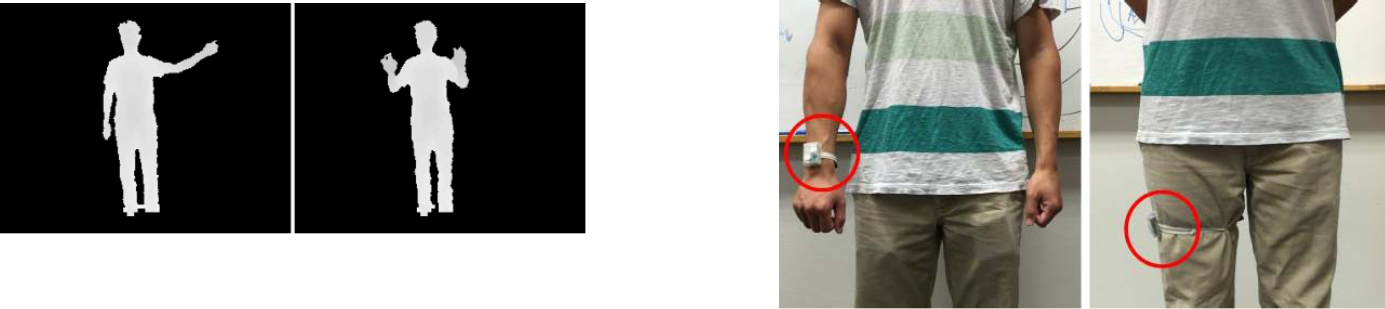


Fig. 2. Example depth images from Kinect depth sensor.

图2. 来自 Kinect 深度传感器的深度图示。

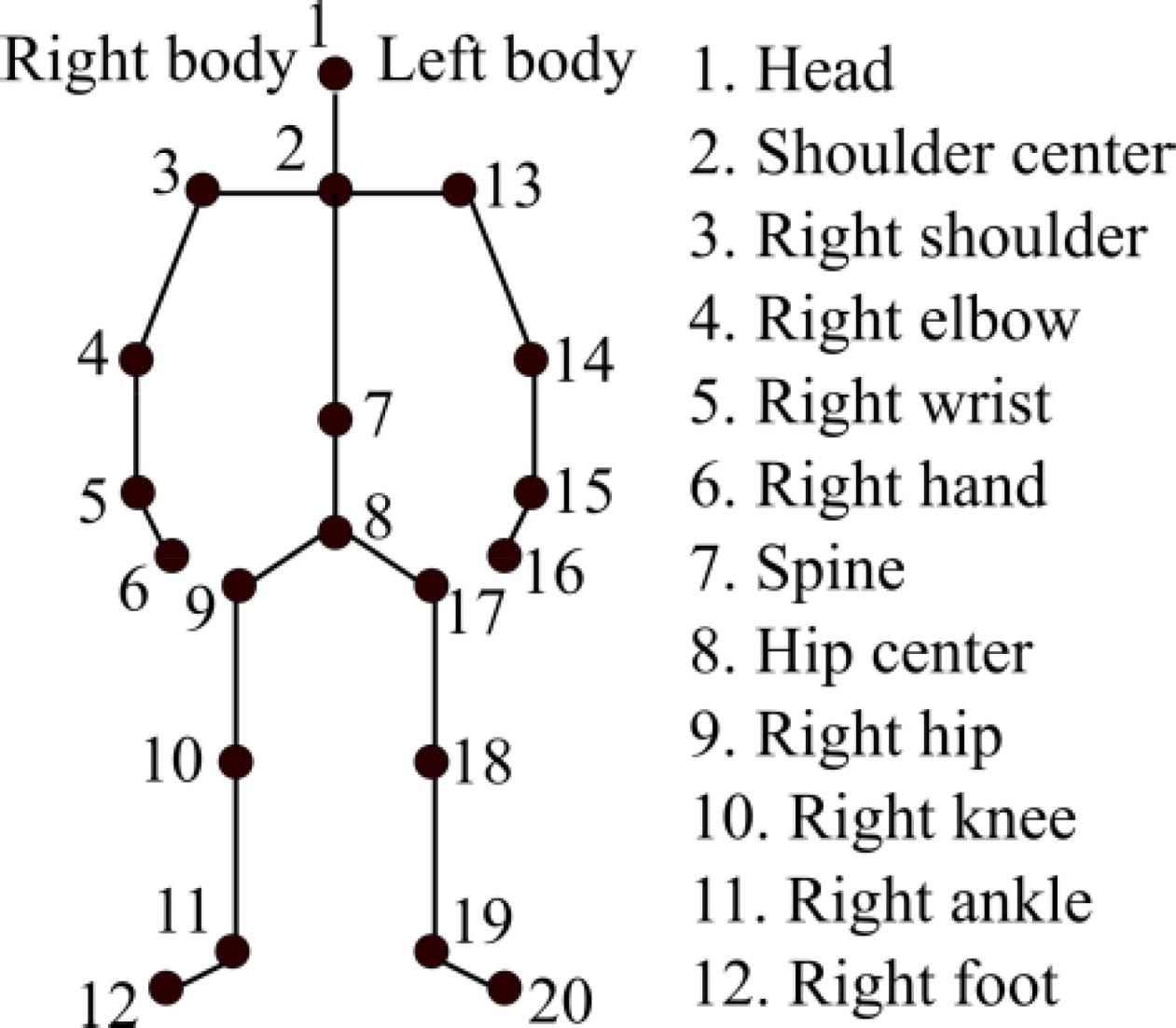


Fig. 3. Skeleton joints provided by Kinect depth sensor.

图3. Kinect 深度传感器提供的骨架连接。

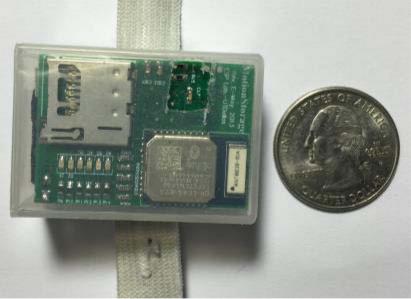


Fig. 5. Inertial sensor placements: right wrist or right thigh.

惯性传感器位置: 右手腕或右大腿。

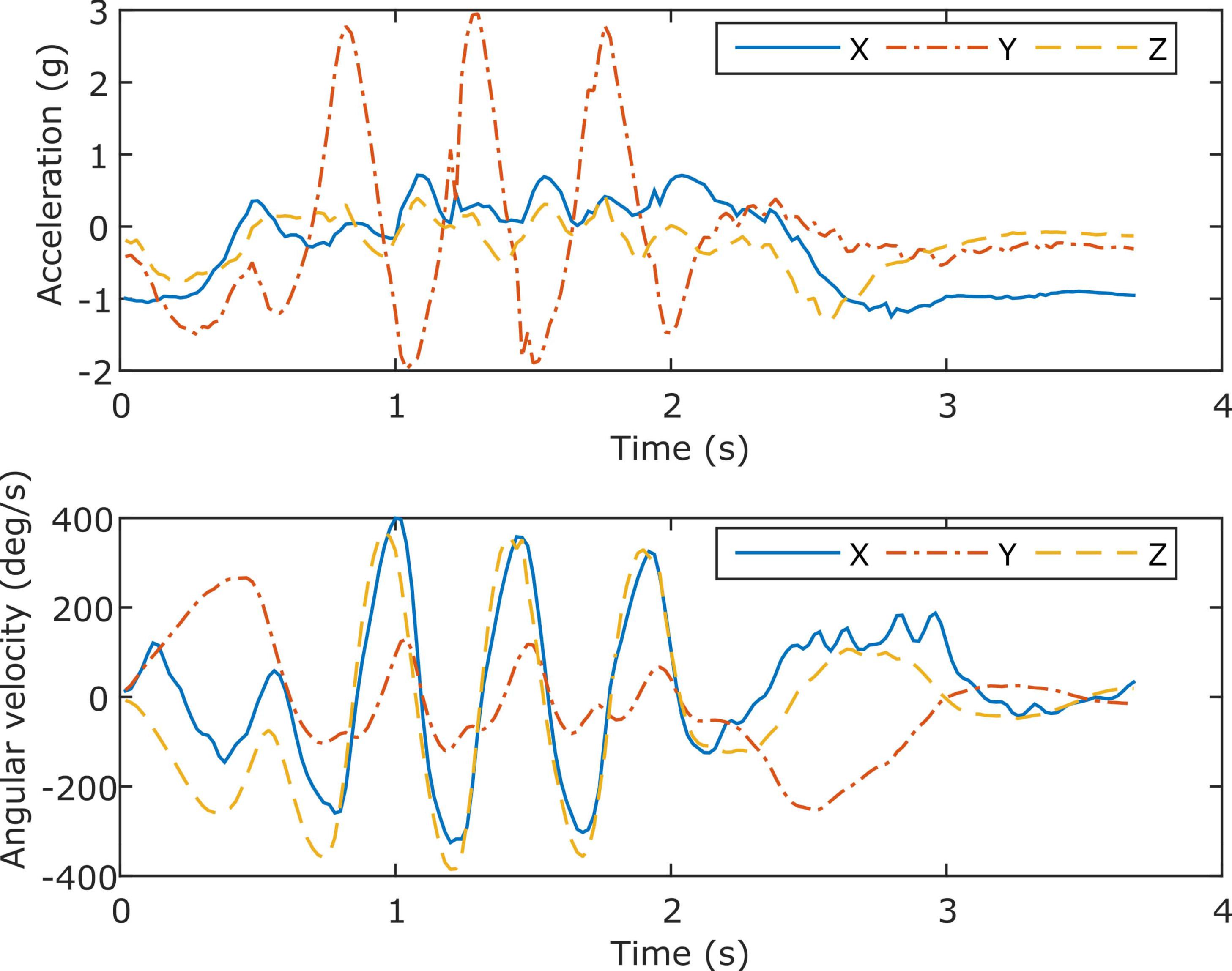


Fig. 4. Wearable inertial sensor developed in the ESP Lab.

图4. 电除尘器实验室研制的可穿戴式惯性传感器。

per second. In addition, the Kinect SDK [14] is a publicly available software package which can be used to track 20 body skeleton joints (see Fig. 3) and their 3D spatial positions.

每秒。此外，Kinect SDK [14]是一个公开可用的软件包，可用于跟踪20个人体骨骼关节(见图3)及其3D 空间位置。

The wearable inertial sensor used in this work is a small size (1”×1.5”) wireless inertial sensor built in the Embedded Signal Processing (ESP) Laboratory at Texas A&M Univer-sity [15]. This sensor captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength, which are transmitted wirelessly via a Bluetooth link to a laptop/PC. This wearable inertial sensor is shown in Fig. 4. The sampling rate of the sensor is 50 Hz and its measuring range is ±8g for acceleration and ±1000 degrees/second for rotation. It is worth mentioning that other commercially available inertial sensors can also be used in place of this inertial sensor. For practicality reasons or to avoid the intrusiveness associated with asking subjects to wear multiple inertial sensors, only one inertial sensor is considered in our work, either worn on the right wrist (similar to a watch) or the right thigh as depicted in Fig. 5 depending on the action of interest to be recognized in a particular application. More explanations about

在这项工作中使用的可穿戴惯性传感器是一个小尺寸(1“1.5”)的无线惯性传感器，内置于德克萨斯 a & m 大学嵌入式信号处理(ESP)实验室[15]。这种传感器可以捕捉3轴加速度、3轴角速度和3轴磁强度，通过蓝牙无线传输到笔记本电脑/个人电脑。这种可穿戴式惯性传感器如图4所示。该传感器的采样率为50赫兹，测量范围为加速度 ± 8克，旋转 ± 1000度/秒。值得一提的是，也可以使用其他商业上可用的惯性传感器来代替这种惯性传感器。出于实用性考虑，或者为了避免要求受试者佩戴多个惯性传感器带来的干扰，我们的工作中只考虑一个惯性传感器，要么佩戴在右手腕上(类似于手表) ，要么如图5所示，根据特定应用中需要识别的感兴趣的动作，佩戴在右大腿上。更多关于

Fig. 6. Inertial sensor signals (3-axis accelerations and 3-axis angular velocities) for the action *right hand wave*.

图6。惯性传感器信号(三轴加速度和三轴角速度)的作用右手波。

the placement of the sensor for different actions are stated in Section IV. Fig. 6 shows the inertial sensor signals (3-axis accelerations and 3-axis angular velocities) for the action *right* *hand wave*.

不同动作的传感器位置在第四节中说明。图6显示了惯性传感器信号(三轴加速度和三轴角速度)的作用右手波。

*B. Feature Extraction*

*B. 特征提取*

To extract features from depth images, depth motion maps (DMMs) discussed in [7] are used due to their computational efficiency. More specifically, each 3D depth image in a depth video sequence is first projected onto three orthogonal Cartesian planes to generate three 2D projected maps corresponding to front, side, and top views, denoted by *map* *f* , *maps* , and *mapt* , respectively. For a depth video sequence with *N* frames, the DMMs are obtained as follows:

为了从深度图像中提取特征，使用了文献[7]中讨论的深度运动映射(DMMs) ，这是由于它们的计算效率。更具体地说，每个深度视频序列中的3 d 深度图像首先投影到三个正交的笛卡尔平面上，生成三个2 d 投影图，分别对应于前视图、侧视图和顶视图，分别由 map f、 map 和 mapt 表示。对于具有 n 帧的深度视频序列，可以得到如下的 dmm:

1. −1

- 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| D M M D m m | { f,s,t } = { f，s，t } = | mapi+1 Mapi + 1 | mapi 麻皮 | *,* | (1) (1) |  |
|  | { f,s,t } − { f，s，t } - | { f,s,t } { f，s，t } |  |  |  |

*i*=1

*I = 1*

where *i* represents frame index. A bounding box is considered to extract the non-zero region in each DMM and the fore-ground extracted DMMs are then used to serve as features. Since foreground DMMs of different video sequences may have different sizes, bicubic interpolation is applied to resize all such DMMs to a fixed size and thus to reduce the intra-class variability. An example set of DMMs for the action *one hand wave* is shown in Fig. 7. For the system developed

其中 i 表示帧索引。提出了一种基于包围盒的非零区域提取方法，并利用前地提取的数字多目标模型作为特征。由于不同视频序列的前景数字千禧模型可能具有不同的大小，因此应用双三次插值来将所有这些数字千禧模型重新调整到一个固定的大小，从而减少类内的变化。图7显示了单手波动作用的一组 dmm 的示例。为了系统的发展

|  |  |
| --- | --- |
| CHEN et al.: REAL-TIME HUMAN ACTION RECOGNITION SYSTEM USING DEPTH AND INERTIAL SENSOR FUSION 陈等: 使用深度和惯性传感器融合的实时人类行为识别系统 | 775 |



Fig. 7. DMMs generated from a sample video of the action *one hand wave*.

图7。从动作的示例视频生成的动作单手波。

Fig. 8. Action segmentation illustration using skeleton joint positions.

图8。使用骨骼关节位置的动作分割图。

in this paper, only the DMM generated from the front view, i.e. *D M M* *f* , is processed in order to keep the computational complexity low towards achieving real-time throughputs.

为了降低计算复杂度，实现实时吞吐量，本文只对前视图生成的 DMM (即 dmf)进行处理。

For the inertial sensor, each acceleration and gyroscope signal sequence is partitioned into *M* temporal windows as reported in [16]. Three statistical features of *mean*, *variance*, and *standard deviation* are computed for each direction per temporal window. All the features from the temporal win-dows are concatenated to form a single combined feature vector. Thus, for *M* windows, the feature vector dimensionality is 3 × *M* × 3 × 2 = 18*M*.

对于惯性传感器，每个加速度和陀螺信号序列被分割成 m 个时间窗，如文献[16]所述。每个时间窗计算每个方向的均值、方差和标准差。将时域双向图中的所有特征连接起来，形成一个单一的组合特征向量。因此，对于 m 窗口，特征向量维数为3m32 = 18M。

*C. Collaborative Representation Classifier*

*协同表示分类器*

Collaborative representation classifier (CRC) [17] is a com-putationally efficient classifier that has been used in many image classification applications. Let *C* denote the number of classes and ***X*** *j* ∈ R *D*×*n* *j* denote the training samples of

协同表示分类器(CRC)[17]是一种高效的图像分类器，在许多图像分类应用中得到了广泛的应用。设 c 表示类的个数，x j ∈ rn j 表示类的训练样本

class *j* (each column of ***X*** *j* is a D-dimensional sample). Also, let ***X*** = [***X***1*,* ***X***2*, . . . ,* ***X****C* ] ∈ R *D*×*n* denote the set of all the training samples, where *n* = *n*1 + · · · + *nC* is the total

类 j (每列 x j 是一个 d 维样本)。另外，设 x = [ X1，X2，... ，XC ]∈ rn 表示所有训练样本的集合，其中 n = n1 + + nC 是所有样本的集合

number of training samples. In this classifier, a test sample ***y*** ∈R*D*is represented as a linear combination of all thetraining samples ***X***:

训练样本数目。在这个分类器中，一个测试样本 y ∈ r d 被表示为所有训练样本 x 的线性组合:

|  |  |
| --- | --- |
| y = Xα, Y = x, | (2) (2) |

where *α* is an *n*-dimensional coefficients vector corresponding to all the training samples from *C* classes.

其中 n 维系数向量对应于 c 类的所有训练样本。

An *l*2-norm is then considered to regularize *α* based on this optimization formulation

在此基础上考虑了 l2- 范数的正则化问题

|  |  |
| --- | --- |
| αˆ = arg y − Xα 22 + λ α 22 , = arg y-x 22 + 22, | (3) (3) |

vector associated with class *j* . The classification is made by

与类 j 相关的向量。分类是由

*label (****y****)* =arg min *e j* *,* (5)

*Label (y) = arg min e j，(5)*

*j*

where *e* *j* = ***y*** − ***X*** *j* *α*ˆ *j* 2 denotes the residual error, and ***y***ˆ*j*= ***X*** *j**α*ˆ*j*indicates a class-specific representation of ***y***.

其中 ej = y-xj2表示残差，y j = xj 表示 y 的类特定表示。

* 1. MODIFICATIONS MADE FOR REAL-TIME SYSTEM

实时系统的修改

1. *Detection of Action Start and End*

*动作开始和结束的检测*

For real-time operation, it is necessary to identify the start and end of an action. Action segmentation is a challenging task. In our system, it is made a requirement that a subject performs an action naturally by completing the action without any pause in the middle of the action. Furthermore, it is required that an action begins with a static posture and ends with a static posture lasting for at least one second. For example, for the action *right hand wave*, a subject stands in front of the Kinect camera and wears the inertial sensor on his/her right wrist. The static posture is the stand still posture. These requirements allow the action segmentation to be performed in a computationally efficient manner.

对于实时操作，必须标识操作的开始和结束。行动细分是一项具有挑战性的任务。在我们的系统中，它要求主体通过完成动作自然地执行一个动作，在动作中间没有任何停顿。此外，还要求动作以静态姿势开始，以静态姿势持续至少一秒钟结束。例如，对于动作右手波动，受试者站在 Kinect 摄像机前，将惯性传感器戴在他/她的右手腕上。静止的姿势是静止的姿势。这些要求允许动作分割以一种计算高效的方式执行。

In our real-time system, the initial skeleton frames from the Kinect and the initial samples from the accelerometer for a one second duration are used to verify a static posture. Let *Js* = *(xs , y s , zs )* denote the average 3D position of a staticposture and *As* = *a*\_*x* *s* *,* *a*\_*ys* *,* *a*\_*zs* *)* its average accelerations. Note that one can easily obtain these static posture data before the actual real-time operation. When a subject starts an action, the position of a corresponding joint will deviate from the position of the static posture. The following distance between the 3D joint position *J* = *(x* *,* *y,* *z)* in a skeleton frame and *Js* is then computed

在我们的实时系统中，来自 Kinect 的初始骨架帧和来自加速计一秒钟内的初始样本被用来验证静态姿态。设 Js = (xs，y s，zs)表示静态的平均3D 位置，As = a \_ x s，a \_ y，a \_ zs)表示静态的平均加速度。请注意，在实际实时操作之前，可以很容易地获得这些静态姿态数据。当受试者开始一个动作时，相应关节的位置将偏离静态姿势的位置。然后计算骨架框架和 Js 中3D 关节位置 j = (x，y，z)之间的下列距离

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| where λ is a regularization parameter. The l2-regularized 正则化参数在哪里? l2正则化 | | | | | |  |  |  |  |  |
| d = (x − xs )2 + (y − ys )2 + (z − zs )2. D = (x-xs)2 + (y-ys)2 + (z-zs)2. | | | (6) (6) |  |
| minimization of (3) is in the form of the Tikhonov regulariza- (3)的最小化是 Tikhonov 正则化-的形式 | | | | | |  |
| tion [18] leading to the following closed form solution: [18]导致以下封闭形式的解决方案: | | | | | | If for ms consecutive skeleton frames, all the distances are 如果对于 ms 连续的骨架帧，所有的距离都是 | | | |  |
|  |  |  |  | λI)−1XT y. I)-1XT y. |  |  |
|  | *α* | (XT X (XT x | + | (4) (4) | greater than a specified sensitivity σd , the start of an action 大于特定的敏感度 d，行动的开始 | | | |  |
|  | ˆ = ˆ = |  |  |  | is triggered. If for ms consecutive skeleton frames, all the 如果对于 ms 连续的骨架帧，所有 | | | |  |
| Let P = (XT X + λI)−1XT . Given training sample set X 设 p = (XT x + i)-1XT，给定训练样本集 x | | | | | | distances are less than or equal to the specified sensitivity σd , 距离小于或等于指定的灵敏度 d, | | | |  |
| and with λ determined via these samples, P is independent 通过这些样本测定，p 是独立的 | | | | | | the end of an action is triggered. Fig. 8 illustrates the procedure 一个动作的结束被触发。图8说明了这个过程 | | | |  |
| of a test sample y. Therefore, P can be pre-computed as a 因此，p 可以预先计算为一个 | | | | | | of using skeleton joint positions to indicate the start and end of 用骨骼关节的位置来表示 | | | |  |
| projection matrix. Once a test sample arrives, the correspond- 一旦测试样本到达，相应的- | | | | | | an action. The use of ms consecutive skeleton frames avoids 使用 ms 连续的骨架帧可以避免 | | | |  |
| ing coefficient vector αˆ can be simply found via Py, which 系数向量可以简单地通过 Py 找到 | | | | | | responding to possible signal jitters. 对可能的信号抖动做出反应。 | | |  |  |
| is computationally efficient. According to the class labels of 是有效率的。根据 | | | | | | An example of a subject performing the action right hand 一个主体右手执行动作的例子 | | | |  |
| the 的 | training samples, 训练样本, | αˆ can be partitioned into 可以被分割成 | | | C subsets C 亚群 | wave is shown in Fig. 9. Fig. 9(a) exhibits the 3D positions 波形如图9所示。图9(a)显示了三维位置 | | | |  |
| αˆ αˆ | = [ αˆ1, αˆ2, . . . , αˆC ] where = [1,2，... ，c ] | | | αˆ j represents the j 代表 | coefficient 系数 | of the right wrist over time. Fig. 9(b) exhibits the distance d 图9(b)显示右手腕的距离 d | | | |  |

= |*Macc* − *Msacc* |.

776

