# 基于智能手机传感器的用户行为识别

任谦 陈晓东 高臻 刘羽霄 李知赫 张若飞

## 1 背景

#### 1.1 项目简介

近年来，传感器数量和计算能力不断增长，不仅改变了人们的日常生活，也为许多有趣的数据挖掘应用开启了大门。从健康和健身监测，个人生物识别签名，辅助技术和老年人护理，到室内定位和导航，智慧城市以及智能交通等。这些应用背后的核心就是通过智能手机的传感器数据，了解用户的行为。

在此项目中，我们根据TalkingData公司提供的智能手机传感器数据，识别用户行为。

#### 1.2 研究背景

传感器是一种物理装置或生物器官，能够探测、感受外界的信号、物理条件（如光、热、湿度）或化学组成（如烟雾），并将探知的信息传递给其他装置或器官。以Android为例，Android平台支持三个大类的传感器。

其一是运动传感器（Motion sensors），主要指测量加速力，以及沿着三个轴的旋转力。此类别包括加速度计，重力传感器，陀螺仪和旋转矢量传感器。

重力传感器可用于手机横竖屏自动切换、拍照照片朝向选择以及各种重力感应游戏等，重力传感器一般利用压电效应实现，传感器内部一块重物和压电片整合在一起，通过正交两个方向产生的电压大小，来计算出水平方向。加速度传感器与重力传感器相同，也是压电效应，通过三个维度确定加速度方向，功耗更小，但精度低。主要用于计步、识别手机摆放位置朝向角度等。陀螺仪是另一重要的手机传感器，陀螺仪使用角动量守恒原理，一个正在高速旋转的物体（陀螺），它的旋转轴没有受到外力影响时，旋转轴的指向是不会有任何改变的。陀螺仪就是以这个原理作为依据，用它来保持一定的方向。三轴陀螺仪可以替代三个单轴陀螺仪，可同时测定6个方向的位置、移动轨迹及加速度。因此用在体感、摇一摇（晃动手机时间一些功能）、平移/转动/移动手机可在游戏中控制视角、VR虚拟现实、在GPS没有信号时（如隧道中）根据物体运动状态实现惯性导航。

其二是位置传感器，这些传感器测量设备的物理位置，主要包括GPS传感器，磁力计和距离传感器。地球特定轨道上运行着24颗GPS卫星，每一颗卫星都在时刻不停地向全世界广播自己的当前的位置坐标及时间戳信息。手机GPS模块通过天线接收到这些信息。GPS模块中的芯片根据高速运动的卫星瞬间位置作为已知的起算数据，根据卫星发射坐标的时间戳与接收时的时间差计算出卫星与手机的距离，采用空间距离后方交会的方法，确定待测点的位置坐标。常用于地图、导航、测速以及测距等。磁场传感器采用各向异性磁致电阻材料，感受到微弱的磁场变化时会导致自身电阻产生变化。可用于指南针、地图导航方向、金属探测器APP等。

其三为环境传感器（Environmental sensors），这些传感器测量各种环境参数，例如环境空气温度和压力，照明和湿度。此类别包括气压计，光度计，和温度计。

气压计分为变容式或变阻式气压传感器，将薄膜与变阻器或电容连接起来，气压变化导致电阻或电容的数值发生变化，从而获得气压数据。GPS计算海拔会有十米左右的误差，气压传感器主要用于修正海拔误差（将至1米左右），当然也能用来辅助GPS定位立交桥或楼层位置。

随着智能手机的发展普及，传感器的数量和精度逐渐提升，Google API为服务提供商提供了大量反映用户行为的真实数据。然而，由于缺乏行之有效的数据处理方法，始终无法挖掘出数据的真正价值，为用户提供更为精准的服务。幸运的是，自2010年始，大数据领域的突破为处理海量数据提供了可靠的平台和方法，使得传感器数据利用成为可能，而决策树、SVM、神经网络等机器学习算法的改进也有效解决了如何从海量数据中获得有价值信息的问题，扫除了大数据分析的最后一块障碍。

如依靠加速等或陀螺仪等传感器检测到用户处于骑行或者驾驶状态，智能手机可以自行切换到静音或者飞行模式，以免发生事故。也可以通过识别用户一整天在行走和跑步等动作之间的切换，推算用户的运动量，帮助用户进行健康管理等。若能够精准的识别用户行为，则服务提供商可以通过用户的当前状态，准确的提供服务。这也是推动App更加智能化，提升生活质量的有效方法。

与此同时，利用智能手机传感器进行用户行为识别与通过智能手表，手环或者其他智能设备传感器具有极高的相似性，本项目研究所采用的有效方法将对其他种类的行为识别问题提供重要参考。

#### 1.3 研究现状

基于智能手机传感器的用户行为识别是行为识别的重要研究方向，研究人员在这方面做出了许多非常有价值的探索。一方面，研究人员采用了RNN，LSTM以及CNN等算法，不断提升了行为识别的准确率。与此同时，为研究数据来源对识别精度的影响，部分研究探索了不同的数据采集方法，分别从实验室和真实使用环境采集数据，并公布了数据集以供研究对比。下面将对该领域的研究现状做简单介绍。

###### 1.3.1 数据来源

基于智能手机传感器做用户行为识别领域的经典数据集主要包括UCI（UC Irvine）和WISDM（Wireless Sensor Data Mining ）数据集。

UCI数据集全称为UC Irvine Machine Learning Repository。是加州大学欧文分校提供的用于机器学习的公开数据集。迄今已经包括了378个细分领域的数据集。其中与智能手机传感器相关的主要包括UCI-HAR（Heterogeneity Activity Recognition Data Set）、UCI-DSA（Daily and Sports Activities Data Set）、UCI-UIFWA（User Identification From Walking Activity Data Set）以及UCI-HARUS（Human Activity Recognition Using Smartphones Data Set）五个数据集。

UCI-HAR数据集包含智能手机中常见的两个运动传感器的读数。活动包括骑自行车，坐，站立，行走，楼梯上升，楼梯下降六个类型。设备包括4颗智能手表以及8款智能手机，采集了共9位用户的数据。UCI-DSA数据集包含8名20-30岁之间不同性别的受试者19个动作的数据，每个行动总信号持续时间为5分钟。数据主要在室内采集，因为要求受试者以自己的风格进行活动，所以，一些活动的速度和幅度存在主体间的差异。

UCI-UIFWA从位于胸口口袋的Android智能手机手机数据。加速度计数据来源于22名参与者在野外通过预定义的路径手机数据。数据集同样用于活动识别研究目的，为使用运动模式的人的识别和认证提供了挑战。

UCI系列最重要的数据集为UCI-HARUS，实验已经在19-48岁的年龄段内由30名志愿者组成。每个人在腰上戴着智能手机（三星Galaxy S II），进行了六项活动（WALKING，WALKING\_UPSTAIRS，WALKING\_DOWNSTAIRS，SITTING，STANDING，LAYING）。采用嵌入式加速度计和陀螺仪，以50Hz的恒定速度捕获3轴线性加速度和3轴角速度。其中70％的志愿者被选择用于产生训练数据和30％的测试数据。 传感器信号（加速度计和陀螺仪）通过施加噪声滤波器进行预处理，然后在2.56秒和50％重叠（128个读数/窗口）的固定宽度滑动窗口中进行采样。该数据在Kaggle竞赛有相应的用户行为识别比赛。

另一经典数据集是WISDM（Wireless Sensor Data Mining），该数据集出自[Activity Recognition using Cell Phone Accelerometers, Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data]，提供了实验室Activity Prediction和真实数据集Actitracker两个数据集，分别包括1098207和2980765条加速度传感器的有标签数据。采样频率都是20Hz，具有走路、慢跑、上楼梯、下楼梯、坐、站立六种活动方式。

除此之外，还有dartmouth/cenceme数据集，该数据集关联较强，标签一致。相比之下，Columbia kinetic数据集关联一般，有加速度数据，标签重点也不在于用户行为。其他数据集还包括Jiit Accelerometer India。

###### 1.3.2 相关研究

基于智能手机传感器数据，研究人员在行为识别方面做了许多探索。Xing Su[Travel Mode Identification with Smartphone Sensors]在研究中同过传感器数据识别六种用户行为，并对基于智能手机用户识别的方法做了比较详细的阐述。

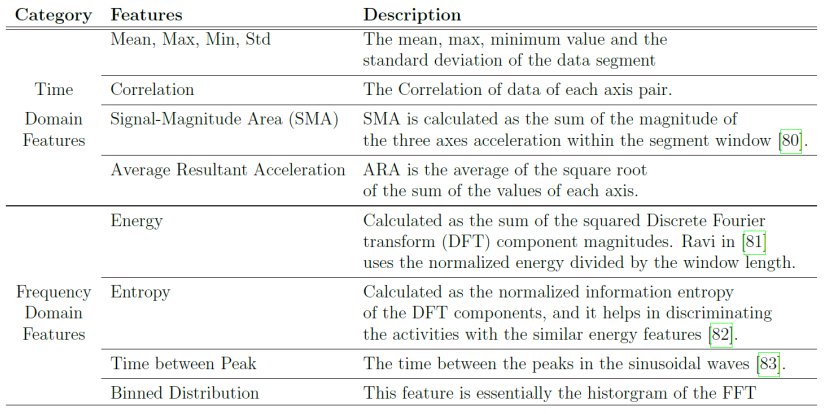


图1 常用时序数据特征

基于元数据，首先通过旋转操作将陀螺仪等传感器数据结合磁力计由机身坐标转化为地球坐标系，然后采用极值调整，高斯平滑分别进行异常值处理，低通滤波，并通过归一化解决幅度敏感的问题。

对于时序数据，首先需要根据传感器的采样频率选取合适的滑动窗口，然后进行数据特征的提取，常用的数据特征如图1所示。Xing Su在研究中采用了group-lasso的特征选择方法。

在行为识别阶段，Xing Su使用层次模型来节省能量和时间，并考虑引入个人模型来适应受测个体间的活动风格差异。第一层分为有轮（自行车、开车、公交）和无轮（走路、跑步），其精度对于有轮和无轮的分类精度可以达到近似100%，对于六种交通模式的分类精度，可以达到平均97.1%。此文利用了在线学习的方法，比离线模型更加快速，且此文模型的采样频率较低，节省电量。使用的方法仍然是SVM，只是添加了分层识别和基于SVM的online learning的因素。

在[基于智能手机的用户行为识别技术研究和应用]中，研究人员简单利用了加速度信号在XYZ轴的均值、标准差、相关系数和峰度等特征，用SVM分类器，在400组训练样本，达到平均90.4%的精度。其特征提取方式可以借鉴，SVM也是本项目重要的候选分类器之一。

在[A Survey on Human Activity Recognition using]中，研究人员将HAR（Human Activities Recognition）的挑战归纳为selection of attributes and sensors, obtrusiveness, data collection protocol, recognition performance, energy consumption, processing, 以及flexibility七个方面。

[Developing and Validating a Statistical Model for Travel Mode Identification on Smartphones]在基于严格的统计回归和逻辑推断的模型，在New Zealand公开数据集上精度为97%，在一个Australia数据集上只有79.3%，其原因是数据比较脏。该模型准确率较高，但是模型太过复杂。

上述研究对于我们的项目具有较高的参考意义。在手机平台进行用户行为识别，应充分考虑手机计算能力以及功耗的限制。行为识别应使用尽可能少种类的传感器数据，一方面方便数据的采集，另一方面能降低维度和复杂度。并且，由于使用的传感器少，传感器功耗也可以得到有效降低。在采用同样数量但不同种类的传感器能达到近似准确率时，应考虑传感器功耗差别，选择功耗较小的传感器数据。例如，若两个加速度传感器和一个加速度结合一个陀螺仪传感器都能准确的识别用户的转向动作，那么邮箱采用两个加速度传感器识别的方式，因为陀螺仪的平均能耗要高于加速度传感器。

参考上述研究，本项目在特征识别阶段可以考虑statistical 或者 structural两种滑动窗口时间序列特征提取方式，并合理选用BIC, MDL, MRMR, CFS, 遗传算法等算法。