

LiveData—基于智能手机传感器的用户数据采集系统

汪仲伟 孙广中

(中国科学技术大学计算机科学与技术学院 合肥 230027)

摘 要 智能手机中各种内置传感器的出现使得用户能采集、分析和挖掘传感器数据中的有用信息。介绍了 Android 平台上用于采集传感器数据的应用 LiveData,它可用于记录用户活动时的数据。针对应用 LiveData 采集的总量为 28 万条的传感器数据记录,提取了若干特征属性来识别用户的行为,并进一步分析了不同传感器以及不同数据采集环境对实验结果的影响。

关键词 智能手机,传感器,行为识别

中图法分类号 TP39 文献标识码 A DOI 10.11896/j.issn.1002-137X.2014.10.006

LiveData—A Data Collecting System Based on Sensors in Smart Phones

WANG Zhong-wei SUN Guang-zhong

(School of Computer Science and and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract With the emergence of every build-in sensor in the smart phone, users can collect, analyze and mine more useful information. We introduced LiveData, an application based on Android platform which is used to collect sensor data. Using 280 thousand data records collected by LiveData, we distinguished the behavior of users by extracting some attributes. We also analyzed the impact of different sensors and different data collection environments on experiment results.

Keywords Smart phone, Sensor, Activity recognition

1 引言

当前主流的智能手机中包含多种传感器,如图 1 所示。对用户行为识别与理解是人机交互领域中重要的研究问题,集成在手机中的传感器对于提高用户交互体验已经发挥了重要的作用。客观地量化用户的行为对于提供个性化的服务以及理解用户有着重要意义。

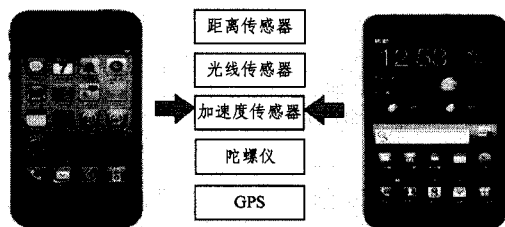


图 1 当前智能手机中包含传感器类型

本研究利用手机传感器来采集数据,并将这些数据用来识别用户的行为。我们开发了一个 Android 平台上用于采集传感器数据的应用“LiveData”。LiveData 可以运行在内置有传感器(加速度传感器、陀螺仪、GPS)的 Android 智能手机上,在不影响用户使用体验的情况下采集传感器数据,并且在手机连接 Wi-Fi 的时候提示用户上传采集的数据至服务器。

LiveData 采集以下 3 种传感器数据:

- 加速度传感器:加速度传感器感受到用户的加速度,并转换成可用的输出信号,LiveData 记录的是用户在 X、Y、Z 轴 3 个方向的加速度信息。
- 陀螺仪:陀螺仪测量偏转和倾斜时转动的角速度,它可以很好地感受到偏转的动作,从而精确分析用户的实际动作。
- GPS:GPS 是基于定位卫星实现定位导航的系统,可以测量出用户的速度等信息。

本文第 2 节介绍了相关工作;第 3 节介绍了 LiveData 系统;第 4 节介绍识别用户行为的方法;第 5 节是实验部分,比较了使用不同传感器数据识别用户行为的准确度以及采集程序的运行环境对于实验结果的影响。

2 相关工作

与本文相关的工作可以分为以下几方面:

- 数据感知

Hong 等人^[1]针对加速度传感器、麦克风以及 GPS 3 个传感器的不同内容,设计了一个传感器系统“The Jigsaw Continuous Sensing Engine”,该系统考虑了临时活动的影响、手机在人身体的不同位置,以及对加速度传感器产生的数据处理,另外系统还对麦克风和 GPS 的采样频率和准确性进行了

到稿日期:2013-06-28 返修日期:2013-08-10 本文受安徽省自然科学基金项目(1208085QF106),国家自然科学基金项目(61202064, 61303047)资助。

汪仲伟(1991—),男,硕士,主要研究方向为数据挖掘,E-mail:wzw095@mail.ustc.edu.cn;孙广中(1978—),男,博士,副教授,主要研究方向为信息检索、高性能计算及应用、组合优化算法及应用。

较好的折中。Hattori^[2]设计了一个名为“ALKAN”的系统来采集用户的加速度传感器数据,并上传至云端分析。系统的特色在于使用一系列激励措施来促使用户进行数据采集。KK Rachuri^[3]提出了一个传感器数据采样方案,其根据历史的上下文事件来动态选择采样函数,同时对传感器采集数据的频率采用动态适应机制,与以前的方法相比有了更好的权衡。

• 行为识别:

微软亚洲研究院^[4,5]从用户的 GPS 轨迹数据出发,首先分析人在不同交通工具下的轨迹特征,并从中推断其选择的交通方式(步行、公交、自行车),进一步根据多个用户的轨迹数据分析人与人之间的关联性等。E. Miluzzo 等人^[6]设计了个人感知应用“CenceMe”,该系统通过 GPS、麦克风等传感器获取数据,采用分类技术识别用户的基本活动信息,基于规则学习算法提取用户的移动轨迹,在此基础上提取更好层次的语义。Kazushige Ouchi^[7]通过手机上的加速度计和麦克风来识别人们的家庭生活的行为(比如刷牙、剃须、吹头发、洗碗等),识别率达到 75.8%,如果排除其他数据干扰,精度能达到 85.9%。Nicholas D. Lane 等人^[9]提出了 CSN(Community Similarity Networks),通过 CSN 对用户行为模式进行分类,做了相似性研究。L. Liao 等人^[10,11]对用户在固定地点以及在路程中的行为加以理解,例如分析用户所采用的交通方式以及预测用户可能的目的地。其主要的方法是从多个用户的数据中提取特征,然后训练出一个分类模型,使用该模型来识别其他人的行为。

3 LiveData 系统

LiveData 有以下的特点:

- 精确的数据采集:应用可以在前台或者后台工作,并且以一定的频率同时记录多个传感器的数据,保证采集到的数据的完整性。LiveData 使用 Android 的 Service 服务,当程序被切换至后台时,Service 就会被启动,使采集程序仍然能继续工作。

- 友好的用户界面:用户可以为即将发生的行为定义一个标签,如步行或者乘车等。同时可以标注程序运行状态,如前台还是后台。另外,应用在后台工作时,不会影响用户对手机使用的体验。

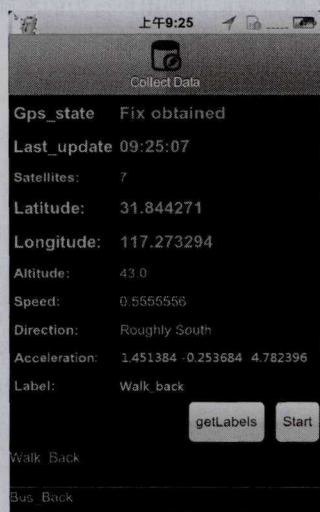


图 2 LiveData 运行界面图

LiveData 的运行界面如图 2 所示。实际上,从手机传感器获得数据到最终数据出现在服务器端,经历了两个过程:首先是将传感器数据存储在本地,在此过程中,为了降低系统 IO 对资源的使用,LiveData 设置了一个数据缓冲区来缓存数据,只有在缓冲区满时才将缓冲区的数据写入到本地文件,避免了频繁写文件,降低了对系统资源的占用;其次是将本地的传感器数据上传至服务器,由于此过程可能会消耗用户的手机流量和电量,并且容易出现错误,因此只在晚上当用户连接至 Wi-Fi 并且接通充电器时,LiveData 才会提示用户是否上传数据,这样既保证了数据正确传输,也不会过多消耗手机的流量和电量。

4 用户行为识别模型

4.1 问题定义

用户建立识别模型的用户行为数据可以用集合 $S = \{s_1, s_2, \dots, s_j\}$ 来表示,其中 $s_i = \{mode, acc_x_i, acc_y_i, acc_z_i, gyro_x_i, gyro_y_i, gyro_z_i, lat_i, lon_i, speed_i\}$ 代表一组传感器数据, s_i 中 $mode$ 表示该组传感器数据对应的行为方式,如步行、骑车等, $acc_x_i, acc_y_i, acc_z_i$ 表示 3 轴加速度数据, $gyro_x_i, gyro_y_i, gyro_z_i$ 表示 3 轴陀螺仪数据, $lat_i, lon_i, speed_i$ 代表 GPS 经度、纬度和速度数据。需要利用集合 S 中的数据训练出用户行为识别模型,用以识别用户行为。

4.2 特征提取

从 4.1 节中了解到,初始获得的传感器数据(原始数据)条目复杂、维度多,并不能直接进行分类选择,因此首先需要对原始数据进行特征提取,剔除不相关和冗余的属性以加快后续的分类速度。本文使用一个长度为 25 的窗口(即窗口内有 25 组数据)去截取行为数据,进行特征提取,同时为了避免数据丢失造成误差,相邻窗口之间有 50% 的重叠。本文的研究主要用到了加速度传感器和 GPS 数据。用集合 $W = \{s_1, s_2, \dots, s_N\}$ 来表示长度为 N 的窗口, s_i 表示每次采样数据,一些研究者^[6,8]指出,手机摆放的位置和方向都会对实验结果产生影响,因此本文不把加速度传感器的 3 个轴(X、Y、Z)分开考虑,而是综合考虑其 3 个轴上的分量,这样加速度 $accValue$ 可以定义为:

$$accelValue_i = \sqrt{acc_x_i^2 + acc_y_i^2 + acc_z_i^2} \quad (1)$$

对于 GPS,本文只使用其速度数据 $speed_i$ 。从窗口提取以下特征:

- 速度平均值和加速度平均值:速度平均值和加速度能直接反映用户的某种行为,如步行的平均速度和加速度肯定比骑车或者乘车的慢,但是存在着一些特殊情况,如在交通拥堵时乘车的速度会比骑车甚至是步行还慢,因此平均值只能作为一个参考属性。

- 速度方差、平均能量和加速度方差、平均能量:用户在活动时不可能保证传感器是固定不动的,传感器会随着用户的动作而发生摆动,因此产生的数据存在误差。方差和平均能量对数据的离散情况有着很好的体现。

另外由于用户的行为一般都呈现出周期性,因此对窗口内的数据进行快速傅立叶变换,得出其在频域上的数据以反映出用户行为的规律分布^[12]。

$$accelFFT_k = \sum_{n=0}^{\frac{N}{2}-1} accel1_n W_N^{kn} + \sum_{n=0}^{\frac{N}{2}} accel2_n W_N^{(2n+1)k},$$

$$k=0,1,\dots,N-1 \quad (2)$$

式中, $accel1$ 和 $accel2$ 分别表示 $accelValue$ 奇数和偶数序列, $W_N^{2kn} = e^{-j\frac{2\pi}{N}2kn}$ 。对于离散后的数据,提取以下特征:

- 速度幅值平均值和加速度幅值平均值。
- 速度幅值加权平均值和加速度幅值加权平均值。

4.3 数据分类

在对原始数据提取特征后,就可以将特征数据送入分类器,本研究使用 weka (<http://www.cs.waikato.ac.nz/ml/weka/>) 对特征数据进行分类。weka 是一个基于 Java、用于数据挖掘和知识发现的开源项目,使用 weka 进行数据分类需要进行以下两个步骤:

- 数据预处理:将需要分类的特征数据预处理成指定的文件格式。
- 数据挖掘:weka 工具已经实现了一些常用的数据挖掘算法,如 c4.5, KNN 等,它可以直接选择这些算法进行分类,另外 weka 还支持使用用户自己实现的算法来进行分类。

5 实验分析

5.1 实验数据集

本实验征集了 5 名志愿者,采用半人工标注的方式采集实验数据,在每次采集数据之前设置好标注,并且将不同的行为分开执行以确保每次只采集一种行为的数据,每次采集行为持续时间在 10 分钟左右,采集频率为 5 次/秒。每次采集加速度传感器和 GPS 数据,总共考虑了 3 种行为状态:步行、骑自行车和乘公交车,另外考虑到采集程序运行在前台或者后台会影响数据采集精度,对于每种行为,前台和后台采集分开进行。实验数据集的情况如表 1 所列。

表 1 实验数据集概况

总状态数	3
数据采集频率(次/秒)	5
总数据量(条)	287326
前台采集数量(条)	167201
后台采集数量(条)	120125

5.2 实验结果与分析

本实验使用 KNN 算法进行十折交叉验证(10-fold cross-validation),即将样本集随机地分成 10 个互不相交的子集,每个子集的大小大致相等。利用 9 个训练子集,对给定的一组参数建立回归模型,将剩下的 1 个子集作为测试集,根据以上过程重复 10 次,因此每个子集都有机会进行测试。每次的实验结果都会有相应的正确率,将 10 次结果的正确率的平均值作为对算法精度的估计。

- 传感器数据种类对实验结果的影响

实验选取加速度传感器和 GPS 数据进行了 3 组实验,分别是单独采用加速度数据、单独采用 GPS 数据、同时采用两种数据,其分类效果如表 2 所列。

表 2 不同传感器数据分类效果对比

	加速度	GPS	加速度和 GPS
最高精度	0.926	0.902	0.942
最低精度	0.829	0.636	0.931
平均精度	0.917	0.825	0.939

从表 2 可以看出,综合使用两种传感器数据的分类效果最好,平均精度能达到 93.9%,而单独使用 GPS 数据时分类

效果较差。由于 Android 系统限制,GPS 更新的速率最快只能是 1 秒,而且输出数据的准确程度受天气等因素影响。通过查看原始数据,发现 GPS 速度数据平均 4 秒更新一次,最坏的情况 11 秒才更新,因此只通过 GPS 速度数据来判断用户行为,所得到的结果是不可靠的。

- 数据采集程序工作环境对实验结果的影响

在本文的第 3 节中提到,LiveData 在被置于后台时会启用 Service 服务,以保证采集数据的精确性,但是由于手机本身的系统资源是有限的,当 LiveData 工作在后台的同时,用户使用手机进行其他活动,比如接电话、听音乐等,系统为了不降低用户体验,在资源不足的时候它可能会降低传感器的输出频率,甚至会停止相应的 Service,因此后台采集到的数据会存在缺失的情况。

由于后台数据的缺失,因此在特征提取之前需要对后台数据进行修复。首先对所有的后台数据的缺失情况进行了统计,结果如表 3 所列,表中每行表示在采集频率为 5 次/秒的前提下,缺失数据条数为 i 在整体数据中所占的比例。

表 3 后台数据的缺失情况

缺失条数 i	所占比例
5	1.71%
4	2.13%
3	2.13%
2	10.87%
1	19.19%
0	63.97%

从表 3 中可以看出,数据完全缺失的情况(缺失条数为 5)所占比例相对较低,大部分的数据都只是缺失了 1—2 条,因此可以考虑使用平均值填充法,使用缺失数据所属的那一秒的所有数据的平均值来补齐缺失值。

对于修复后的后台数据进行特征提取并分类,并与前台数据相比,结果如表 4、表 5 所列。

表 4 修复后的后台数据行为分类的混淆矩阵

	步行	骑车	公交	精确度
步行	1159	82	24	0.916
骑车	115	2321	156	0.895
公交	11	281	2977	0.911

表 5 前台数据行为分类的混淆矩阵

	步行	骑车	公交	精确度
步行	1238	73	8	0.939
骑车	22	2094	133	0.931
公交	29	132	2597	0.942

表 4 和表 5 中分别是用修复后的后台数据以及前台数据进行分类的混淆矩阵,其中 $R_{ij}(i, j \geq 1)$ 表示行为 i 被分类成行为 j 的数量。对比两个表可以很直观地看出,使用算法修复后,精确度得到了一定的提高,比较接近使用前台数据进行分类的精确度。

结束语 本文介绍了一个基于传感器来采集用户数据的应用 LiveData,并且分析了采用不同传感器数据和数据采集环境对实验结果的影响。实验结果表明,采用加速度传感器和 GPS 速度数据结合的方式来识别的准确率明显高于单独使用一个传感器数据。另外,由于后台数据存在一定的缺失,如果通过一定的算法对后台数据进行修复,也可以在一定程度上提高分类的准确率。

在下一步的工作当中,我们将寻求从系统源头提高后台采集数据的准确率的方法,真正做到在获取较高的数据准确率的同时提供较好的用户使用体验。

参 考 文 献

- [1] Lu H, Yang J, Liu Z, et al. The Jigsaw continuous sensing engine for mobilephone applications[C]//Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems. ACM New York, 2010; 71-84
- [2] Hattori, Technol Y I O, Kitakyushu, et al. A Large Scale Gathering System for Activity Data with Mobile[C]//Proceedings of the 15th Annual International Symposium on Wearable Computers, 2011; 97-100
- [3] Rachuri K, Musolesi M, Mascolo C. Energy-Accuracy Trade-offs in Querying Sensor Data for Continuous Sensing Mobile Systems[C]//Proc. of Mobile Context-Awareness Workshop. 2010
- [4] Zheng Yu, Liu Li-ke, Wang Long-hao, et al. Learning transportation mode from raw gps data for geographic applications on the Web[C]//Proceedings of the 17th international conference on World Wide Web. ACM New York, 2008; 247-256
- [5] Li Quan-nan, Yu Zheng-chen, Xie Xing. Mining user similarity based on location history[C]//Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in Geographic Information Systems. ACM New York, 2008
- [6] Miluzzo E, Lane N, Fodor K, et al. Campbell. Sensing meets mobile social networks; The design, implementation and evaluation of the CenceMe application[C]//Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems. ACM New York, 2008; 337-350
- [7] Ouchi K, Doi M. Living activity recognition using off-the-shelf sensors on mobile phones[J]. annals of telecommunications, 2012, 67(7-8); 387-395
- [8] Bao L, Intille S S. Activity recognition from user annotated acceleration data[C]//Proceedings of the 2nd International Conference on Pervasive Computing. 2004; 1-17
- [9] Lane N D, Xu Y, Lu H, et al. Enabling large-scale human activity inference on smartphones using community similarity networks (csn)[C]//Proceedings of the 13th International Conference on Ubiquitous Computing. ACM New York, 2011; 355-364
- [10] Liao L, Patterson D J, Fox D, et al. Learning and Inferring Transportation Routines[C]//Proceedings of the National Conference on Artificial Intelligence. 2007; 311-331
- [11] Patterson D J, Liao L, Fox A D. Inferring High-Level Behavior from Low-Level Sensors[C]//Proceedings of the 13th International Conference on Ubiquitous Computing. 2003; 73-89
- [12] Tukey J W, Cooley J W. An Algorithm for the Machine Calculation of Complex Fourier Series[J]. Mathematics of Computation, 1965, 19(90); 297-301
- [13] Zlot R, Bosse M. Place recognition using keypoint similarities in 2D lidar maps[C]//Experimental Robotics. Springer Berlin Heidelberg, 2009; 363-372
- [14] Guivant J, Nebot E. Optimization of the simultaneous localization and map building algorithm for real time implementation [J]. IEEE Transactions on Robotics and Automation, 2001, 17(3); 242-257
- [15] Nieto, Juan, Bailey T, et al. Scan-slam: Combining ekf-slam and scan correlation[C]//Field and service robotics. Springer Berlin Heidelberg, 2006; 167-178
- [16] Martinez-Cantin R, Castellanos J A. Unscented slam for large-scale outdoor environments[C]//IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems. 2005; 328-333
- [17] Montemerlo M, Thrun S. Simultaneous localization and mapping with unknown data association using FastSLAM[C]//the Proceedings of the 2003 IEEE International Conference on Robotics & Automation (ICRA'03). IEEE, 2003; 1985-1991
- [18] Eliazar A, Parr R. Dp-slam: Fast, robust simultaneous localization and mapping without predetermined landmarks[C]//International Joint Conference on Artificial Intelligence. Lawrence Erlbaum Associates LTD, 2003, 18; 1135-1142
- [19] Olson E B. Real-time correlative scan matching[C]//IEEE International Conference on Robotics and Automation (ICRA'09). IEEE, 2009; 4387-4393
- [20] Olson E B. Robust and efficient robotic mapping[D]. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2008
- [21] Borges G A, Aldon M J. Line extraction in 2d range images for mobile robotics[J]. Journal of Intelligent & Robotic Systems, 2004, 40(3); 267-297
- [22] Nunez P, Vazquez-Martin R, del Toro J C, et al. Feature extraction from laser scan data based on curvature estimation for mobile robotics[C]//Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA 2006). IEEE, 2006; 1167-1172
- [23] Li Y, Olson E B. Extracting general-purpose features from lidar data[C]//2010 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2010; 1388-1393

(上接第 26 页)