计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190605
57(7): 1490-1507, 2020

一种基于智能手机传感器数据的地图轮廓生成方法

陶 涛[1] 孙玉娥[2,5] 陈冬梅[1] 杨文建[1] 黄 河[1,3] 罗永龙[4,5]
[1](苏州大学计算机科学与技术学院 江苏苏州 215006)
[2](苏州大学轨道交通学院 江苏苏州 215131)
[3](中国科学技术大学苏州研究院 江苏苏州 215123)
[4](安徽师范大学计算机与信息学院 安徽芜湖 241002)
[5](网络与信息安全安徽省重点实验室(安徽师范大学) 安徽芜湖 241002)
(ttao@stu.suda.edu.cn)

A Method of Map Outlines Generation Based on Smartphone Sensor Data

Tao Tao[1], Sun Yu'e[2,5], Chen Dongmei[1], Yang Wenjian[1], Huang He[1,3], and Luo Yonglong[4,5]
[1](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)
[2](*School of Rail Transportation, Soochow University, Suzhou, Jiangsu* 215131)
[3](*Suzhou Institute for Advanced Study, University of Science and Technology of China, Suzhou, Jiangsu* 215123)
[4](*School of Computer and Information, Anhui Normal University, Wuhu, Anhui* 241002)
[5](*Anhui Provincial Key Laboratory of Network and Information Security (Anhui Normal University), Wuhu, Anhui* 241002)

Abstract With the development of the economy, environmental maps are becoming more and more important to our daily lives. The existing mechanisms of map generation are mainly based on vehicle-driven GPS equipment for data acquisition and road network construction. However, these methods have the disadvantages of low precision and poor efficiency, and the methods cannot construct the map for some areas where the acquisition equipment is difficult to reach or the GPS signal is weak. In order to solve the problems mentioned above, this paper proposes an idea of constructing a map through mining the sensor data generated by the widely used smartphones. Based on this idea, a data fusion algorithm is proposed. Firstly, the machine learning classification algorithm and signal processing technology are used to identify the traveling state. And then, the segmentation mechanism is combined with the dynamic time warping algorithm to process the steering segment. Finally, the local map outline is generated by the fusion of the distance data and direction data of the effective segment. The experimental results based on the data collected from the real road network prove the effectiveness of the proposed method in the construction of local map outlines and the feasibility of deep mining sensor data.

Key words machine learning; state recognition; map generation; data mining; smartphone

收稿日期:2019-08-28;**修回日期**:2019-12-11

基金项目:国家自然科学基金面上项目(61672369,61873177,61572342);网络与信息安全安徽省重点实验室开放课题(AHNIS2019003);江苏高校优势学科建设工程资助项目
This work was supported by the General Program of the National Natural Science Foundation of China (61672369, 61873177, 61572342), the Open Project of Anhui Provincial Key Laboratory of Network and Information Security (AHNIS2019003), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

通信作者:孙玉娥(sunye12@suda.edu.cn)

摘　要　近年来,随着社会经济的不断发展,许多商业服务以及旅游出行活动对环境地图的依赖越来越大.传统的地图生成方法主要基于车辆驱动型的GPS设备进行数据的采集和路网的构建.然而该类方法存在精度低、时效性差等缺点,并且该类方法对于一些采集设备难以到达或者GPS信号弱的地带无法进行地图的构建.为了解决上述问题,提出了通过挖掘广泛普及的智能手机内部传感器数据进行地图构建的思想,并基于该思想提出了一种数据融合算法.该算法基于智能手机采集的行人步行数据,利用机器学习分类算法与信号处理技术进行行进状态的识别,采用分段机制结合动态时间规整算法进行转向情况的处理,通过融合有效状态下行进的距离数据和方向数据,最终生成局部地图轮廓.将所提算法在真实路网采集的数据上进行实验,实验结果证明了所提方法对局部地图轮廓构建的有效性以及深入挖掘传感器数据的可行性.

关键词　机器学习;状态识别;地图生成;数据挖掘;智能手机

中图法分类号　TP391

在智慧生活的背景下,智慧城市、智能交通受到了政府工作者和各领域科研学者的广泛关注.其中,地图信息发挥着重要的作用[1-2].例如餐饮实体店位置信息获取的准确性决定了顾客的数量;外卖、快递等服务完成的时效性取决于服务过程的路线规划(依赖起讫点之间的地图数据)情况;城乡布局的调整(例如大型商业中心或者停车场的建立、道路的扩建、乡村的改建等)以及旅游景区的规划对目标周边的地形地图等信息也具有很强的依赖性.目前,国内外比较领先的地图平台主要有高德地图、百度地图以及Google地图等,这些平台虽然拥有海量的地理信息,但它们无法实时地监控局部区域的整改扩建情况,导致其数据库中存在大量不完整的路网信息.而且这些情况往往会造成累积趋势,需要较长的时间才会得到更新,因此容易引发地图数据更新慢与智慧生活要求快2方面的矛盾,所以一种能够实时生成局部地图的方法极为重要[3-4].通过对以往经典地图生成工作的调研发现,目前常见的地图生成方法可分为基于车载GPS经纬度数据的构建方法和基于视频图像数据的构建方法.而考虑地图的时效性要求,本文提出了一种基于便携智能手机传感器数据的地图生成方法.通过对经典传感器[5-7]工作的研究,本文分析了基于传感器数据工作的主要模式,并利用大量智能手机传感器采集的数据,进行了地图轮廓的构建.近年来,基于智能手机传感器的工作越来越得到研究学者的重视,其中也存在一些基于传感器构建环境地图的工作,例如文献[5]采用分布式结构,提出利用多机器人进行地图的构建,建立了不精确传感器的混合模型,但该工作对设备要求较高,从而无法广泛应用.本文工作致力于对经典的地图生成方法无法监测到的区域进行地图轮廓数据的生成.相对于已有的地图生成方法,本文利用智能手机的便携性,通过对行人步行产生的持续传感器数据进行挖掘与重组,从而弥补地图上缺失路段的轮廓数据.

首先,本文通过对传感器数据的预处理,提取常见步行的转向数据段,并通过对常见转向情况的分析,对提取的数据段进行筛选,将其处理成方向转变的参考模板;通过真实环境下采集得到的数据集,利用信号处理方法进行特征的计算,然后利用机器学习的方法进行特征的筛选、训练数据的生成以及分类模型的构建.接着,结合动态时间归整(dynamic time warping, DTW)算法及随机森林(random forest algorithm)分类算法从传感器数据中识别出表征有效地图数据的分段,最后利用步数检测方法以及时域积分方法——辛普森法则(Simpson's rule)进行数值计算,并以二元组序列形式将生成的地图轮廓数据进行记录.本文的工作在证明所提方法有效性的同时,也证明了基于智能手机传感器数据的深入挖掘工作的可行性,为后续的基于智能手机传感器的工作提供了系统的分析步骤和处理方法.

1　相关工作

相关的工作主要可以分为3类:经典室外地图生成工作、室内地图生成工作以及基于智能手机传感器的工作.

1.1　经典地图生成工作

早期的室外地图生成工作主要以测绘为主,通过实地的测量进行数据的记录与地图的绘制.但这种方法由于周期长、速度慢而逐渐被淘汰,后续的地图生成工作主要包括对图像数据进行挖掘提取路网

信息以及利用 GPS, WiFi 或全球移动通信系统(global system for mobile communications, GSM)等设备进行数据的采集和地图的绘制.

在基于图像数据的地图生成工作中,摄像设备普遍存在成本高、延时长等缺点,而且设备的部署以及数据的存储均需消耗大量资源,例如文献[8]利用移动车辆结合图像信息进行拓扑图构建,其中数据的采集极易受到设备的影响,具有较大的局限性.除此之外,基于数字图像数据的地图生成工作还存在3个问题:1)数据涵盖的实际路网信息占比较少,利用率较低;2)数据涵盖的额外信息量大,可能具有隐私泄露的风险;3)数据的采集容易受到外部环境的影响,例如在光线昏暗的时空下进行数据采集容易导致错误成像,无法保证生成地图的精度.

基于 GPS, WiFi/GSM 数据的地图生成工作主要利用流动的汽车携带 GPS 或者 WiFi 通信设备[9],结合"众包"技术,对某城区路段进行广泛覆盖的位置数据采集,然后利用采集所得经纬度数据进行地图路线的生成.该工作主要有3个问题:1)该类设备采样率较低,从而容易在数据采集中导致路段数据的缺失;2)该类数据偏差较大,例如错误的GPS数据容易造成待测路段的严重扭曲从而导致最终路段数据无法准确刻画,产生不可估量的误差;3)在信号较弱的偏远地带,该类通信设备的采集工作较为困难,此外,基于车辆的 GPS 设备采集工作在广阔的路段具有较好的效果,但在乡村小路、景区路段中,由于其对普通车辆的限制,数据采集工作甚至无法进行.

1.2 室内地图生成工作

室内地图生成工作一般用于室内定位,而室内定位的工作可以大致分为2种.

1) 通过 WiFi/GSM 接入点或者雷达系统[10]进行定位.这种方法可以利用随身携带的设备与发射点的信号交互判断目标距离发射点的距离,从而利用多发射点进行综合判断,锁定目标的位置[11].这类工作虽然不需要进行室内地图的生成,但是往往需要通过部署一定数量的高成本设备,因此该类方法显然无法用于室外地图的数据生成工作.目前新近的工作中也存在一些需要生成室内地图的工作,但是该类工作也主要依赖于室内环境中广泛存在的无线电等信号,甚至有些工作需要结合室内平面图[12].

2) 室内定位方法,通过利用智能手机传感器设备进行室内轮廓数据的生成从而对目标进行定位.然而,由于室内建筑具有占地面积小、整改频率低等特点,所以室内轮廓数据的处理工作中往往采用较为复杂耗时的算法以提高室内轮廓的生成精度[13-14],然而尽管如此,很多工作在以步长计算距离的过程中累积误差对精度的影响依然较大.而另一些工作,例如文献[15]利用 iBeacon 技术结合传感器数据进行室内轮廓的生成工作,虽然精度比仅仅使用传感器数据有所提高,但是这类工作在室外地图的构建过程中依然面临设备部署时效性差等限制.

综上所述,室内定位处理过程中采用的技术在室外地图生成工作方面的实用性较差,其处理方法无法在实时性、高效性、经济性以及部署的简便性方面同时满足室外地图的轮廓生成要求,导致这类工作无法在大范围的室外范围进行推广普及.

1.3 智能手机传感器相关工作

在智慧生活的背景下,基于智能手机的内置传感器的工作越来越多,其原因主要可归结于2方面:1)智能手机用户群体大,数据来源广;2)智能手机具有较高的计算能力和存储能力,可以进行实时的数据采集和处理.目前广泛利用智能手机传感器的工作主要有动作识别、环境感知、情绪检测、医疗保健等方面.文献[16]通过将智能手机结合"众包"技术,利用汽车进行了城市道路坑洞检测方面的研究;文献[17]利用智能手机传感器中的温度、湿度等传感器进行环境检测以监控某区域的温度状况防止中暑现象的发生,该方法可以构建部分区域的环境温度图,但该图无法代表该区域的路网图;文献[18]利用智能手机传感器数据开发了一款检测跑者呼吸节奏的应用程序用来协调跑者的呼吸速率和步伐大小;文献[19]利用智能手机程序结合传感器数据帮助病人维持健康;文献[20]利用智能手机传感器通过检测用户睡觉过程产生的声音和动作来测评记录用户的睡眠质量.

通过对以上相关工作的学习研究和总结,本文所提地图轮廓生成模式主要有4点改进和贡献:

1) 提出并验证了深入挖掘传感器数据的可行性.通过对实际路段上持续采集得到的传感器数据的处理和挖掘,验证了本文所提方法的有效性以及提高传感器数据利用率的可行性,为后续的工作奠定了基础.

2) 利用信号处理方法和机器学习技术,提出了一种基于状态识别的数据融合算法,该方法可以检测出表征有效地图轮廓数据的分段,并将分段的数值和方向信息进行融合,生成对应的轮廓序列.

3) 在转向情况的处理中,结合动态时间规整方法提出了转向模板的概念,使得转向情况的处理较以往工作更贴近实际情况.在数值计算过程中采用

了时域积分方法进行有效分段的数值计算,并且采用了分段计算机制,减少了经典步长的计算方法中普遍存在的累积误差现象,保证了计算的速度和精度.

4) 本文的工作是目前第一份仅利用传感器数据进行室外地图生成的工作.相比经典的地图轮廓生成方法,本文的方法具有更好的普适性和抗干扰性,在对数据的挖掘深度方面,本文的工作较以往基于传感器的工作具有更好的数据利用率.

2　数据采集及处理目标

本文使用的数据来自智能手机内置传感器生成的实时数据.不同的传感器数据具有不同的形式,图 1 展示了手持智能手机步行 60 s 产生的数据.常用传感器的数据一般以三元组的形式进行表示,例如,加速度传感器(accelerometer, acc)、陀螺仪(gyroscope, gyr)等,而线性加速度传感器(linear accelerometer, lacc)的数据相对加速度传感器而言,去除了由地球重力产生的加速度数据(gravity),可以描述为 $V_{acc}=V_{gravity}+V_{lacc}$.本文工作主要采用了加速度传感器、陀螺仪传感器、线性加速度传感器以及表示智能手机所处方向的方位传感器(orientation sensor, ori),其中,方位传感器的数据包括偏航角(yaw)、俯仰角(pitch)以及滚转角(roll).偏航角指示从某点的指北方向线起,以顺时针方向到目标方向线之间的水平夹角;而俯仰角和滚转角则分别刻画了智能手机的 x 轴与 y 轴相对于水平面的俯仰程度.

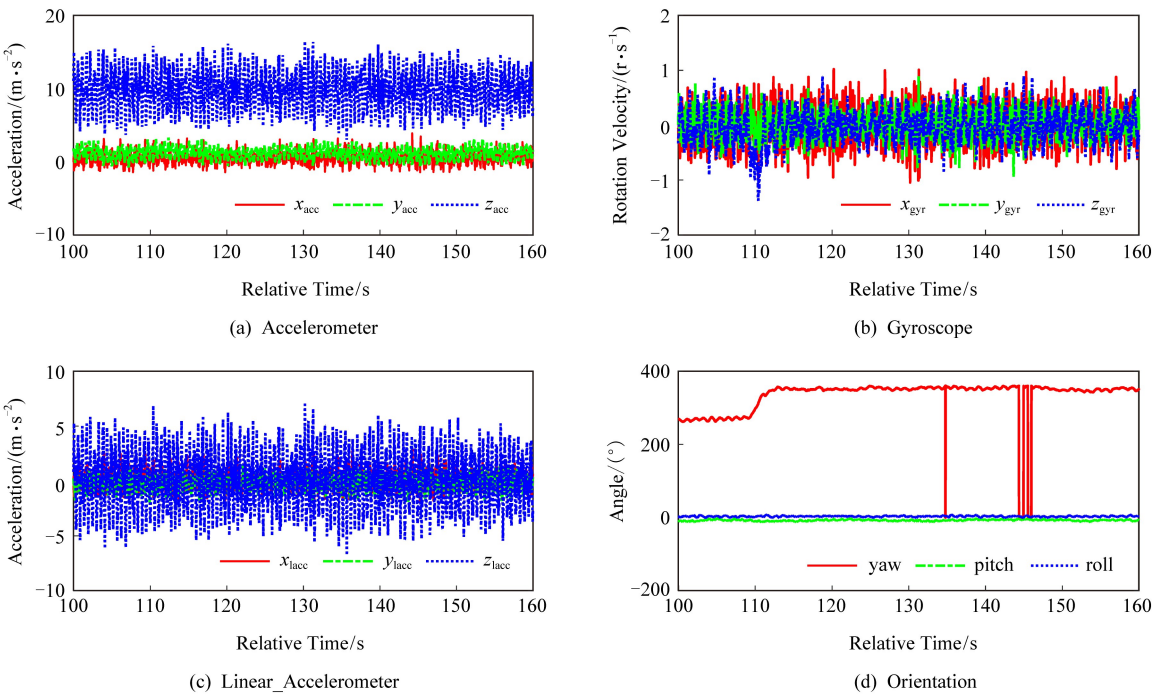


Fig. 1　The sensor data generated by walking 60 s

图 1　步行 60 s 产生的传感器数据

在所有基于传感器的工作中,数据的完整性是工作质量的重要保障之一.经典的传感器主要以定点部署为主,其生成的数据通过不同的传输渠道发送到终端进行处理.而这一过程普遍具有 2 个问题:1)设备依赖性较大,即传感器的部署、数据的发送对设备所处环境以及本身性能依赖较大,例如磁力计等传感器无法在磁场较强的环境下进行数据采集;2)实时性较差、数据易缺失,例如在数据包的发送过程中,容易受到外部的干扰而导致数据的成段缺失,这类情况虽然在日常的动作识别工作中影响较弱,但在持续利用传感器数据的工作中具有较大影响,这也是导致基于传感器数据的工作中数据利用率低的主要原因.由于传感器数据反映的动作种类较多,导致采集的数据波动也较大.所以不同的动作引起的数据无法通过插值等方法进行补充,数据在时序上的缺失会降低实验的可信度,甚至导致整个实验无法正常进行.本文利用智能手机在应用程序开发和测试方面的便携性,以安卓系统为平台开发了一款普适的传感器数据采集软件.该软件可以同时进行数据的采集、存储以及分析,在传感器种类的选择

和采样频率等参数的设定方面具有很强的自主性，保证了整个实验数据的实时性以及完整性.

本文的工作主要在于提出一种地图轮廓原型的构建方法以弥补传统的地图生成方法无法构建的路段数据，根据智能手机内置传感器的分析及本课题研究目标的综合考虑，本文首先利用方位传感器数据中的偏航角数据进行转向模板的提取，然后利用加速度传感器数据结合陀螺仪传感器数据进行有效路段的识别，最后利用线性加速度传感器数据和方位传感器数据计算表征有效地图轮廓的数据段的数值.通过在真实环境下采集的路段数据上的测试，验证本文方法的可行性和有效性，主要的处理流程如图 2 所示：

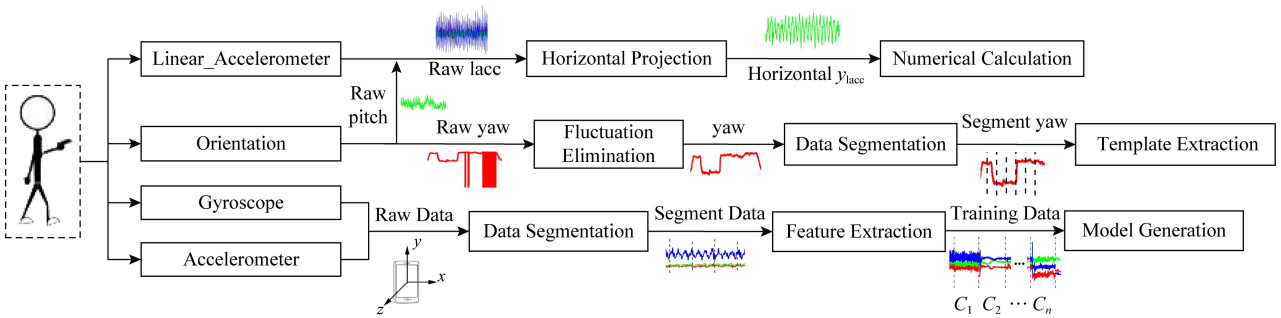


Fig. 2 Overview of processing sensor data

图 2 数据处理工作流程

基于本文的研究目标，数据采集的方式主要为行人手持智能手机行走在待构建路段上进行数据采集，并且采集过程中遵从 3 个合理的假设：

1) 行人将手机持在手中，手机朝向行进方向；

2) 行人行走步伐正常，假设不会出现左右不规则晃动、摇摆等动作；

3) 行人行进过程正常，假设行走方向不会频繁改变，不会倒着走.

参考文献[21]实验设置，本文将采集数据的频率设定为 50 Hz，并且在数据处理过程中使用 50% 的重叠率.本课题的研究目标主要为地图原型轮廓数据的生成，为方便地图数据的表述和存储，本文将地图数据以二元组序列形式进行表示，其中，单个的序列元素以二元组$\langle r,s\rangle$表示，符号 s 表示距离长度(单位为 m)；符号 r 数据则指示转向情况，根据不同情况主要用 2 种数据表示：1)当行进方向不变时，该数值为一固定角度值，指示手机方向与正北方向偏离的程度；2)当行进方向处于转变过程时，该值为序列模板的指示符号，指示当前转变过程对应的模板序号.为简化问题描述，本文将行进过程中的转向情况综合表述为表 1 所示的 f_1 与 f_2 共 2 种情况.

通过表 1 所述的转向情况的表示可以发现，通过细化偏转角度以增加转向模板的数量可以达到提高转向情况处理精度的目的.本文的工作主要致力于提供一种地图轮廓构建方法的原型及其探究过程，为验证所提方法的有效性，本文将利用方向传感器数据分析结果进行常见转向模板的提取.

Table 1 Simplified Representation of the Steering

表 1 转向情况简化表述

Steering Situation	Description
f_1	Shift σ_1 degrees to the right from current position, The change process of direction angle is $\gamma\rightarrow\gamma+\sigma_1$
f_2	Shift σ_2 degrees to the left from current position, The change process of direction angle is $\gamma\rightarrow\gamma-\sigma_2$

Note: The initial direction is set to deviate from due north by γ degrees.

根据数据分析的过程描述以及定义的地图数据的表述形式，本文将地图轮廓生成问题分解为转向情况的处理、数据分段的识别以及有效段的数值计算 3 个子问题，然后提出了一种数据融合算法将上述子问题的处理结果进行结合，以达到生成最终地图轮廓数据的目的.其中，转向情况的处理主要包括对方向数据中偏航角的跳变现象的处理以及转向模板段的提取；有效段的识别主要用机器学习的方法对加速度传感器数据和陀螺仪数据进行分析，从而达到分类识别的目的，主要过程包括特征的计算、筛选、训练集的生成以及分类模型的建立，分段数值的计算主要指结合方向传感器中的角度信息以及线性加速度传感器数据进行相关的数值计算.

3 数据的处理方法

本节主要描述传感器数据具体的处理过程及其对应的方法.主要包括对方位传感器数据进行处理，

提取转向模板;对加速度传感器数据和陀螺仪数据进行处理,训练用于识别有效段的分类模型;利用方位传感器数据和线性加速度传感器数据进行有效数值的计算.

3.1　转向模板的提取

通过第 2 节的课题目标的分析及问题的分解,本节主要致力于地图轮廓构建过程中方向数据的处理过程.

在方向传感器所测得的 3 个数据分量中,偏航角数值的取值范围为[0°,360°],俯仰角数值的取值范围为[−180°,180°],滚转角数值的波动范围为[−90°,90°].图 3 所示为一段正常行走过程采集得到的偏航角数据,为方便观察,所测数据分别以离散点型和连续线型进行绘制和展示,通过绘制的 2 幅图可以清晰地看出偏航角在整个过程中会出现跳变现象.

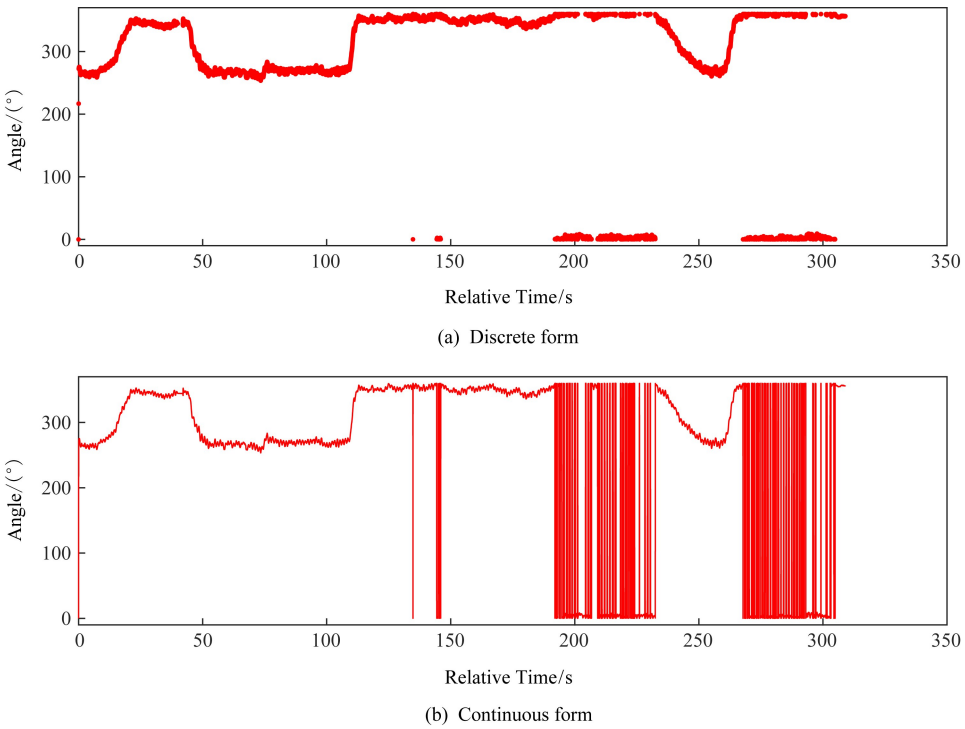


Fig. 3　The fluctuations in yaw data

图 3　yaw 的连续跳变现象

由实际采集情况和观测分析可得,偏航角的连续跳变现象产生的原因主要在于偏航角的数据具有固定的范围.即由于偏航角表征手机正向偏离正北方向的水平夹角,当手持手机以任意方向开始持续旋转一周以上便会产生一次跳变的数据.若恰好手机持向沿正北方向行进,则由于步行运动的不规则性以及身体的抖动会导致采样数据频繁介于 0°和 360°这 2 个数据点附近跳跃,最终导致如图 3 所示的跳变现象的产生.在传感器数据的处理工作中,通常需要将离散数据在时域上进行连续表示以简化问题的处理,而上述的跳变情况会增加问题分析的难度.据此,本文提出了一种针对此现象的跳变消除方法,并基于该算法进行了转向模板的提取操作,具体步骤为:

1) 从采集得到的方向传感器中获取偏航角数据,并且删去由操作采集软件生成的无效数据;

2) 由原始范围内的偏航角数据分别向上、向下构造 2 个具有相同走势的数据段,用越界数据补充跳变数据;

3) 选取原始数据中的第 1 个数据点作为新序列的初始数据点,然后以数据跳变阈值为界限,利用数据点密度间距进行传递,顺序链接当前数据的下一个数据点,直至获取与原始数据长度相等的新数据,即为跳变消除后的数据;

4) 利用平滑样条插值方法(smoothing spline)对消除跳变后的数据进行平滑操作,并对平滑后的数据进行分段,分段过程保证段与段之间具有一定的重叠率,以保证分段数据能够完整覆盖转向过程;

5) 从所有分段数据中提取间距最接近常见转向数据跨度的段作为转向模板段,保存模板段的数据并加以编号.

上述方法主要包括跳变数据的消除和转向模板的提取2个阶段,在数据平滑之前的操作不仅消除了跳变现象,同时也维持了原始数据的波动走势.在转向模板的处理中,为简化问题描述,本文主要考虑常见的转向角度跨度值作为转向模板的阈值,例如常见的转向有直角转弯,对应的跨度值为90°.上述过程对应的伪代码可描述如下:

算法1. 结合跳变消除方法的转向模板提取算法.

输入:原始偏航角数据 $\boldsymbol{E}^{s}_{yaw}$、相邻数据点之间的密度阈值 φ、无效数据段长 D_{seg}、采样频率 f、分段段长 $winLen$、重叠率 o、转向模板跨度阈值 ω;

输出:转向数据模板段 T.

阶段Ⅰ. 消除原始偏航角数据中的跳变现象:

① $\boldsymbol{\theta}^{s}_{yaw}=\boldsymbol{\theta}^{s}_{yaw}(D_{seg}\times 1000f:\text{end})$;

② 将 $\boldsymbol{\theta}^{s}_{yaw}$ 中的每个数值都加上360,得到对应原始数据相同走势的上界数据 $\boldsymbol{\theta}^{up}_{yaw}$,将 $\boldsymbol{\theta}^{s}_{yaw}$ 中的每个数值减去360,得到对应原始数据相同走势的下界数据 $\boldsymbol{\theta}^{lo}_{yaw}$;

③ $\theta^{res}_{yaw}[0]=\theta^{s}_{yaw}[0]$;

④ for $i=1$ to $length(\boldsymbol{E}^{s}_{yaw})$ do

⑤ 分别计算 $\theta^{res}_{yaw}[i-1]$ 与 $\theta^{s}_{yaw}[i]$, $\theta^{up}_{yaw}[i]$, $\theta^{lo}_{yaw}[i]$ 三者之间的差值;

⑥ 使用差值小于阈值 φ 的 $\theta_{yaw}[i]$ 值更新 $\theta^{res}_{yaw}[i]$ 的值;

⑦ end for

阶段Ⅱ.提取转向模板:

⑧ $SmoothingSpline(\boldsymbol{E}^{res}_{yaw})$, $T=\{\text{null}\}$;

⑨ $\boldsymbol{S}=DataSegment(\boldsymbol{E}^{res}_{yaw},winLen,o)$;

⑩ for $i=1$ to $length(\boldsymbol{S})$

⑪ if $\max(\boldsymbol{S}_i)-\min(\boldsymbol{S}_i)\approx\omega$/* 此处约等符号表示差值介于 ω 数值附近的较小范围之间,根据实际的情况可以调整该波动的范围 */

⑫ 将满足条件的段 $\boldsymbol{S}_i$ 加入模板段 T 中;

⑬ end if

⑭ end for

⑮ return T.

算法1主要包括对方向数据消除跳变以及对处理后的数据进行分段处理2部分.通过对智能手机方向传感器数据数值进行上下界的计算,本文所提跳变消除算法通过一次遍历即可识别信号异常跳变部分,即算法1可以保证在时间复杂度为 $O(N)$ 的情况下记录对应方向传感器的无跳变数据.空间消耗只需额外记录对应于方向传感数据同等长度的2段数据,分别保存 $\boldsymbol{E}^{up}_{yaw}$ 以及 $\boldsymbol{E}^{lo}_{yaw}$.

算法1首先将偏航角范围内发生跳变的数据段用范围外的数据进行替换,保护了数据的整体走势及其完整性,然后将处理后的偏航角数据进行分段,并提取出段内间距接近既定阈值的数据段作为转向模板.由处理过程可以发现,本文提出的跳变消除算法在原始时间序列不变的情况下保存了数据的基本走势,在后续的工作中,本文将利用该数据走势结合DTW方法将待计算的转向数据段与已提取的模板段进行比对,从而判别当前状态下的转向情况.

此外,虽然第2节所提假设可以排除行进过程手机异常偏离水平面的情况,但由于步行过程产生的震荡依然会对每一步的行进数据造成一定程度的影响.针对该问题,本文主要通过分析方向传感器数据中表征智能手机偏离水平面程度的俯仰角分量进行行进方向上的速度合成,该部分的处理将在数值计算部分进行叙述.

3.2 识别模型的构建

本节主要分析加速度传感器数据以及陀螺仪数据的处理过程.主要内容包括训练数据的生成以及识别模型的构建2部分.由于文献[21]中的工作已被证实在状态识别方面具有一定的效用,所以在对本节数据的处理过程中,本文参考了文献[21]工作的数据处理方法和参数的设定.

3.2.1 相关定义

在手持智能手机行进过程中,行人会做出不同的动作,在传感器上则体现为不同的状态.通过分析数据采集过程中的实际运动情况,本文将主要出现的状态分为3类:行进状态、静止状态以及上下楼梯状态.其中,对最终地图轮廓数据具有构建作用的主要状态包括行进状态和上下楼梯状态,然而,在室外局部地图轮廓的构建过程中,上下楼梯的情况相对少很多,基于以上分析,本文进行了相关概念的定义:

定义1. 有效状态.本文所述有效状态的处理主要为行进状态的识别与提取,关于上下楼梯状态仅以识别为主,在最终生成的轮廓数据中进行简单标识.

定义2. 有效段.基于有效状态的定义,本文将使用机器学习的方法对已分段的数据进行状态识别,满足有效状态的数据段视为有效段.

3.2.2　训练数据的生成及识别模型的构建

本文首先利用5种不同型号的安卓手机进行了大量的数据采集,然后提取其中的加速度传感器数据和陀螺仪数据进行分析.处理步骤可简化为图4的4步.其中分类模型的建立部分,本文选取了Random Forest算法进行识别模型的构建.在文献[21]的工作中,Random Forest算法的性能已经与其他常用的7种分类算法:Naive Bayes,Logistic,Simple Logistic,SMO,AdaBoostM1,J48,RandomTree进行了对比.8种常见分类算法的性能在多种不同分类指标下均进行了详细地分析,实验结果表明在综合考虑加速度传感器数据和陀螺仪数据进行动作识别工作中,Random Forest算法较其他7种算法性能更加突出.

1) 数据的分段及特征的计算

从图4所述模型的构建过程可以发现,进行模型构建之前首先需要进行分段段长的确定.关于数据分段长度的选取在基于传感器的工作中已经被许多学者进行了深入的研究[22],结果表明段长的选取主要根据待监测的目标动作的特性而定.本文在转向数据段的提取、识别模型的构建中均需要进行数据的分段操作.对于本文研究目标而言,数据段长如果太短,容易导致数据的过度截取,会严重影响数据处理的速度;数据段长如果过长则容易导致转向过程的错误识别,从而影响构建的地图性能.通过对前人工作基础以及实际采集数据过程的分析,在50 Hz的数据采集频率下,本文将分段段长设定为12 s.关于分段方法,本文选用End-Point Detection方法[23]

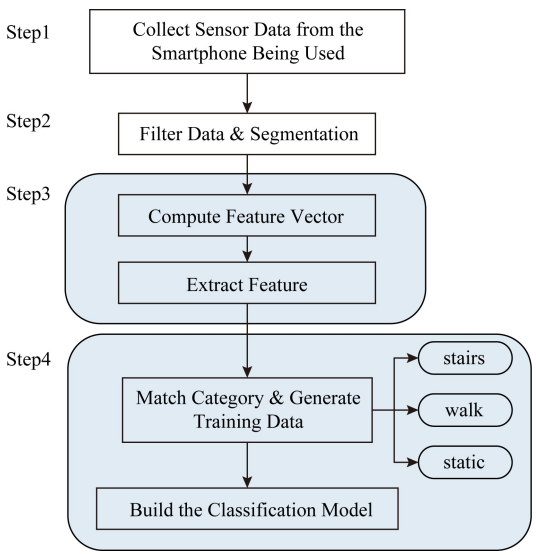


Fig. 4　The workflow to build classification model

图4　模型的构建流程

进行分段.对于数据分段特征的计算,本文分别从时域、频域以及统计域3个方面进行了综合刻画.

① 时域信号特征

时域信号特征的特点主要是能够直观、清晰地刻画采样数据近似的形状,而且对信号数据进行时域方面的特征计算时通常具有较低的时空复杂度.本文则主要选取能够直观度量各个数据分量及数据合成加速度的最大值(max)、最小值(min)以及平均值(average)这3种常见特征来表述不同状态下的加速度传感器数据和陀螺仪数据在时域方面的特点.其中,时刻t对应的合成加速度以及分量加速度分别以符号$a(t)$,$a_x(t)$,$a_y(t)$,$a_z(t)$进行简单表示,$a(t)$的计算为

$$a(t)=\sqrt{a_x^2(t)+a_y^2(t)+a_z^2(t)}. \tag{1}$$

② 频域信号特征

频域信号特征的优势主要在于其能够刻画待处理数据中无法通过直接观测得到的一些周期性特点.在频域特征的计算中,考虑信号的自然特性,本文将对时间序列上采集所得的数据进行快速傅里叶变换(fast Fourier transform, FFT),以方便进一步分析待处理数据的频域特点.本文选取的频域特征主要采用了傅里叶变换后的直流分量(DC component)、峰值幅度(peakAmp)以及峰值频率(peakFreq)这3个特征.DC component主要用来反映当前数据段的平均能量信息,peakAmp特征主要用来反映数据段中主要重复分量的能量,peakFreq特征主要用来反映数据段中最大谱幅出现的频率,计算过程为:

首先对待处理的数据进行分段,每段共有n条记录,然后进行快速傅里叶变换,并用$F_j(a_i)$表示第i个分段的第j个频率分量的幅度;用$f_{ij'}$表示第i个分段的第j'个频率分量,即第i个分段的最大谱幅,其计算为

$$\begin{cases} peakfreq_i=f_{ij'}, \\ j'=\arg\max\limits_{j\in(1,n]}|F_j(a_i)|. \end{cases} \tag{2}$$

③ 统计域信号特征

由于人类活动较为复杂,所以当人类的行为作用在传感器上时,经常容易产生常规的时域特征和频域特征无法刻画的特殊效应,所以本文对待处理数据段在统计域方面的特征也进行了计算,主要选取的特征包括跨度范围(range)和峰度(kurtosis)这2个特征.range主要用以表征数据段中最大幅值和最小幅值之间的跨度范围,kurtosis表示的是数据段中均值附近峰值高低分布的平面度,其计算过程为:

将待处理数据分为 n 段后,用 a_{ij} 表示第 i 个分段的第 j 个采样点,用 $\bar{a}_i$ 表示第 i 个分段的所有采样点的平均值,则每个分段的峰度 kurtosis 计算为

$$Kurtosis_i=\frac{n\sum_{j=1}^{n}(a_{ij}-\bar{a}_i)^4}{\left(\sum_{j=1}^{n}(a_{ij}-\bar{a}_i)^2\right)^2}. \tag{3}$$

2) 训练集的生成及模型的构建

通过匹配已分割的数据段特征及其对应的状态,本文进行了训练数据集的生成工作.并基于该训练集以及 10 倍交叉验证方法,本文进行了识别模型的训练工作,其主要步骤可描述为:

① 针对既定的有效状态进行数据采集,并标记每份数据所属状态.步行标记为 walk,上下楼梯标记为 stairs,静止标记为 static;

② 删减掉由于操作采集软件而产生的无效数据,然后根据既定的分段长度进行数据分段.与方向传感器数据保持一致,本节处理的传感器数据分段的重叠率也设定为 50%;

③ 对分段后的数据进行特征提取,主要包括时域方面的特征:max, min, average;频域方面的特征:DC component, peakAmp, peakFreq 以及统计域方面的特征:range, kurtosis;

④ 根据计算得到的特征向量匹配对应的状态,生成训练数据集,然后基于 10 倍交叉验证方法利用 Random Forest 算法建立分类模型.

根据上述步骤进行操作即可得到全特征集下的分类模型,然而结合实际情况考虑发现,在训练集的生成过程中,往往因较多冗余特征的存在而导致训练模型的构建速度以及分类模型的识别精度大幅度受损.为了避免这类情况的产生,本文同时进行了冗余特征的筛选工作.目前对于冗余特征的筛选方法大致分为 2 种:1)通过对初始特征集的所有子集进行评估,得到一个对分类精度贡献最大的特征子集作为最终筛选后的特征集;2)通过逐个评估初始提取的特征集合中的单个特征对最终正确分类率所作的贡献,将所有特征进行排名,然后再根据问题需求选取 Top-k 个特征组合为特征集作为最终筛选后的结果.由集合论以及幂集概念可知,第 1 种方法在求解子集方面具有不可忽视的时空复杂度,尤其在待处理特征较多的情况下,子集的计算时间开销不容忽视,甚至可能造成整个模型构建速度的严重下降.综合考虑本课题的研究目标以及待识别状态的复杂性,本文选用第 2 种特征筛选方法,并利用信息增益 $InfoGain$(information gain)的概念进行特征集的评估,其评估的原理可描述为

$$InfoGain(Class,Attribute)=H(Class)-H(Class|Attribute). \tag{4}$$

至此,基于加速度传感器数据和陀螺仪数据的模型构建过程描述完毕,对应的伪代码如算法 2 所述,其中数据的分段操作与前述的转向模板的提取类似.

算法 2. 训练集的生成及分类模型的构建算法.

输入:已完成分段操作的具有标签的原始数据 $R_{\text{data}}=\{\boldsymbol{S}_i\}, i\in[1,n]$;

输出:分类模型 $Model$.

① 定义矩阵 $\boldsymbol{Q}_{\text{data}}$,用于存放提取的加速度传感器和陀螺仪各维度数据,以及 2 个传感器基于各自的 3 维数据计算得到的合成加速度数据,初始时 $\boldsymbol{Q}_{\text{data}}$ 为 null;

② for each $\boldsymbol{S}_i\in R_{\text{data}}$

③ 提取 $\boldsymbol{S}_i$ 中的加速度传感器和陀螺仪对应的每一维数据分别存入 $\boldsymbol{Q}_{\text{data}}$ 中对应位置;

④ for $j=1$ to $length(\boldsymbol{S}_i)$ do

⑤ 计算加速度传感器的合成数值 $a(S_{ij}(1), S_{ij}(2), S_{ij}(3))$以及陀螺仪的合成加速度数值 $a(S_{ij}(4), S_{ij}(5), S_{ij}(6))$并将结果追加到 $\boldsymbol{Q}_{\text{data}}$ 对应的列中;

⑥ end for

⑦ 利用得到的 $\boldsymbol{Q}_{\text{data}}$ 数据计算时域特征和统计域特征;

⑧ 对 $\boldsymbol{Q}_{\text{data}}$ 执行快速傅里叶变换后,进行频域特征的计算;

⑨ end for

⑩ $C\{\boldsymbol{S}_i\}=\{\boldsymbol{Q}_{\text{data}}$ 中所有不同维度上的特征值$\}$;

⑪ if $\boldsymbol{S}_i=$stairs

⑫ $C\{\boldsymbol{S}_i\}=\{C\{\boldsymbol{S}_i\}$, stairs$\}$;

⑬ else if $\boldsymbol{S}_i=$walk

⑭ $C\{\boldsymbol{S}_i\}=\{C\{\boldsymbol{S}_i\}$, walk$\}$;

⑮ else $C\{\boldsymbol{S}_i\}=\{C\{\boldsymbol{S}_i\}$, static$\}$;

⑯ end if

⑰ end if

⑱ 使用 10 倍交叉验证法处理训练数据;

⑲ $Model=RandomForest(C\{R_{\text{data}}\})$;

⑳ return $Model$.

算法 2 主要对已分段数据进行特征的计算操作.段内数据主要包括 4 种不同传感器在 3 个维度方面的数据,在计算特征时,通过向量操作,每个维度的特征均可在 $O(N)$时间内完成对应的特征计算.所以,特征计算的时间复杂度为 $O(s\times N)$,其中 s 为总的分段数目,计算过程不需要额外消耗存储空间.

3.3　有效段的数值计算

根据 3.2.1 节对有效段概念的定义,本文将有效段的数值计算主要分为统计步数以及计算步长 2 部分.

3.3.1　步数的统计

通过对经典的步数检测工作的学习研究,步数检测的方法大致可以分为 2 种:1)利用 GPS 数据计算出既定路段的长度,然后利用个人的步长数据和整个路段的时间消耗综合判断个人的步数;2)检测传感器数据在步行过程中产生的波峰、波谷等信息,利用信号处理技术去除噪声后进一步计数.第 1 种方法主要应用于一些健身运动应用程序的工作中,例如一些应用程序会根据用户沿着某路段的行进距离以及时间生成一份包含各项运动信息的报表.第 2 种经常用于基于传感器的工作中.根据本课题的研究目标,本文提出了一种交点检测法进行步数的统计.具体操作步骤可描述为:

1) 利用平滑样条插值法对删除无效段后的加速度数据进行噪声去除处理;

2) 结合线性加速度传感器的数据特性以及方向传感器的角度信息,将加速度数据进行合成,得到行进方向的加速度数据值;

3) 利用所得数据计算消除噪声后的数据平均值,将该值作为加速度数据浮动的基准数值,即加速度数据的波动现象产生在该基准数值附近;

4) 计算线性加速度数据与基线之间的交点(point of intersection, PI)个数,再由交点个数计算得到基本步数.

本文定义在行进方向上正常行走 1 步产生的加速度数据如图 5 标记点所示.经典的工作中将单步定义在相邻波峰或者相邻波谷之间.根据上述的步数统计过程,本文将相邻波峰向前推进 1/4 周期段定义为正常的单步数据,以此为单步的原因主要有 2 点:1)交点稳定,不会因为一步多峰或者一步多谷的现象影响计数;2)加速度数据走势保持先增后减趋势,有利于后续单步步长的计算.

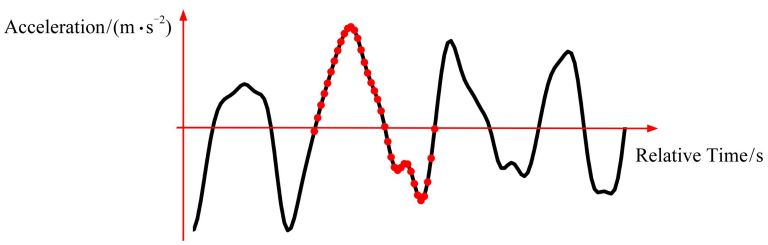
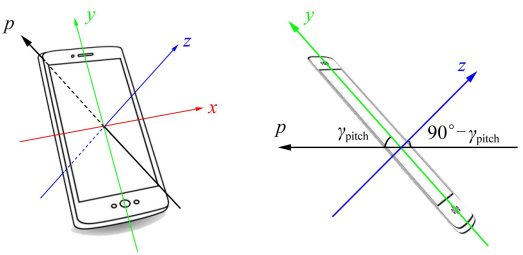


Fig. 5　The definition of single step

图 5　单步定义

关于行进方向上加速度数据的合成如图 6 所示,由于手持手机处于向前行进的过程,本文将加速度在 x 轴方向的行进距离进行忽略,所以行进方向的加速度主要分布在 y 轴、z 轴 2 个方向上,图 6 中所示 γ_{pitch} 为俯仰角 pitch 的数据读数,字母 p 指示行进方向.将最终沿行进方向的加速度分量用 p_{lacc} 表示,其数值结果计算为

$$p_{\text{lacc}}=y_{\text{lacc}}\times\cos\gamma_{\text{pitch}}-z_{\text{lacc}}\times\cos(90^\circ-\gamma_{\text{pitch}}). \quad (5)$$



(a) Coordinate representation　(b) Coordinate side view

Fig. 6　The calculation of synthetic acceleration

图 6　加速度数据合成示意图

3.3.2　步长的计算

根据对经典步长计算方法的学习和总结,以往的步长计算工作往往具有时效差、累积误差大等缺陷,其主要原因在于数据处理过程复杂、计算过程繁琐、缺乏校准工作等.例如有一些工作在计算步长时,仅仅通过用户输入个人信息进行理论的计算,而忽略了人类活动的多变性和复杂性.考虑实际情况下的行走过程,因为运动具有惯性,所以行人在固定的路段会倾向于维持一个均匀的速度前进,但每步的步长又可能因为外界环境发生改变,这些改变主要体现在每步对应的加速度数据上.根据 3.3.1 节单

步的定义以及实际情况的考虑,本文引入惯性初速度(v_0)的概念,用来表示行人普遍维持的速度,该值的确定因个人不同而有轻微区别,可以通过对个人进行平均行进速度的训练得到.基于该值的引入,本文利用时域积分方法进行行进方向上加速度数据的处理,并且在检测到转向段或状态的改变时,会用 v_0 重置下一个状态对应路段的初始速度,该过程的处理能够在一定程度上缓解以往工作中普遍出现的累积误差现象.其具体步骤可表述为

1) 基于数据预处理与识别操作(包括数据的分段、平滑、有效段的识别等工作)以及前述关于单步的定义,对已识别的待处理数据段进行步数统计;

2) 由于平滑处理可能使数据段中的数据点减少,因此本文将待处理数据段分别进行单步提取,并对提取的单步数据中的离散点进行拟合,获得原始时间序列上等量的数据点;

3) 利用辛普森法则结合惯性初速度在时域上对单步数据段中的数据点进行数值积分获取每个单步的步长;

4) 将单步计算方法扩展到每个分段数据上,继而获取待处理数据段(已被截取并识别为有效行进状态的数据分段)的总长度.

通过上述分析,本文将步数的统计过程和步长的计算过程进行了结合,其伪代码可表述如下:

算法 3. 基于辛普森法则的有效段数值计算方法.

输入:有效状态对应的数据段 $\boldsymbol{D}_v$(包括线性加速度传感器数据、方向传感器的俯仰角数值等)、训练得到的惯性初速度 v_0、采样频率 f;

输出:路段长度 $segDistance$.

① 使用式(5)计算 $\boldsymbol{D}_v$ 中每条记录对应的 p_{lacc} 值,并将得到的整列数据进行平滑处理得到新的单列数据 $\boldsymbol{D}_c$,初始化 $segDistance=0$;

② $baseValue=mean(\boldsymbol{D}_c)$; /* 计算 $\boldsymbol{D}_c$ 的平均值作为 $baseValue$ */

③ 将 $\boldsymbol{D}_c$ 中每条数据减去 $baseValue$ 得到新的差值数据列 $\boldsymbol{dif}$,定义 $\boldsymbol{Prod}=[]$;

④ for $i=2$ to $length(\boldsymbol{dif})$ do

⑤ 计算 $dif(i)\times dif(i-1)$ 并将值追加到 $\boldsymbol{Prod}$ 中;

⑥ end for

⑦ 把 $\boldsymbol{Prod}$ 中正值置为 0,负值置为 1;

⑧ $sumPI=sum(\boldsymbol{Prod})$; /* 获取交点个数 */

⑨ 找到 $\boldsymbol{Prod}$ 中值为 1 对应的下标值存入 $\boldsymbol{IntSeq}$;

⑩ if $D_c(IntSeq[1]-1)<D_c(IntSeq[1]+1)$

⑪ $startSeq=1$;

⑫ else $startSeq=2$;

⑬ end if

⑭ for $i=startSeq$ to $sumPI$ do

⑮ $\boldsymbol{Seg}_i=D_c(IntSeq[startSeq]:IntSeq[startSeq+2])$;

⑯ $\boldsymbol{vel}_i=ones(size(\boldsymbol{Seg}_i))\cdot v_0+Simpson(\boldsymbol{Seg}_i,\boldsymbol{t}_i)$;
$\boldsymbol{dis}_i=Simpson(\boldsymbol{vel}_i,\boldsymbol{t}_i)$;

⑰ $segDistance=segDistance+dis_i[length(\boldsymbol{dis}_i)]$;

⑱ $i=i+2$;

⑲ end for

⑳ return $segDistance$.

3.4 数据的融合

利用算法 3 的步骤即可进行有效段的数值计算,通过将转向情况与数值计算结果进行结合即可生成对应路段的轮廓数据.对此,本文提出一种基于状态识别的数据融合算法(state-recognition based data fusion algorithm, SrcDFA),该算法的功能包括识别数据分段的状态,以及将已识别为有效状态的数据分段的方向数据与距离数据进行融合,生成最终地图轮廓数据序列,其主要思想可描述为:

1) 通过数据的预处理操作进行数据的噪声去除处理并截断无效数据;

2) 将用于计算行进数据的线性加速度传感器数据结合方向传感器的俯仰角数据进行行进方向的数据合成;

3) 对整个数据进行分段,获取总段数;

4) 分别识别每个数据分段对应状态的有效性;

5) 利用 DTW 方法将有效数据段与已提取的转向模板库中的模板进行对比,同时利用辛普森法则计算有效段的行进距离;

6) 将有效段的转向情况和行进距离进行数据融合,并追加到地图轮廓数据序列中;

7) 最终生成地图轮廓数据序列.

上述过程的细节处理和实现已经在 3.1~3.3 节中进行了详尽的阐述和说明,下面给出上述过程的伪代码:

算法 4. 一种基于状态识别的数据融合算法(SrcDFA).

输入:原始数据 $\boldsymbol{D}_r$;

输出:地图轮廓数据序列 $Seq_{outline}$.

① 对原始数据 $\boldsymbol{D}_{\mathrm{r}}$ 进行预处理操作;

② $\boldsymbol{E}_{\mathrm{yaw}}^{\mathrm{res}}=EliminateFluctuation(\boldsymbol{E}_{\mathrm{yaw}}^{\mathrm{s}})$; /* 对原始的 yaw 数据 $\boldsymbol{E}_{\mathrm{yaw}}^{\mathrm{s}}$ 进行跳变消除操作,得到新的 yaw 数据 $\boldsymbol{E}_{\mathrm{yaw}}^{\mathrm{res}}$ */

③ $\boldsymbol{A}_{\mathrm{lacc}}^{p}=SyntheticValue(\boldsymbol{A}_{\mathrm{lacc}}^{y},\boldsymbol{A}_{\mathrm{lacc}}^{z},\boldsymbol{E}_{\mathrm{pitch}}^{\mathrm{raw}})$; /* 利用线性加速度传感器数据的 y 轴数据和 z 轴数据以及原始的 pitch 数据在行进方向 p 上合成出线性加速度数据 */

④ $\boldsymbol{Seg}_{\mathrm{total}}=DataSegment(\boldsymbol{D}_{\mathrm{r}},\boldsymbol{E}_{\mathrm{yaw}}^{\mathrm{res}},\boldsymbol{A}_{\mathrm{lacc}}^{p})$;

⑤ for each $\boldsymbol{S}_i$ in $\boldsymbol{Seg}_{\mathrm{total}}$

⑥ $State\leftarrow Model(\boldsymbol{S}_i)$;

⑦ if $isValid(State)$

⑧ $r\leftarrow DTW(\boldsymbol{S}_i^{\mathrm{yaw}},Templates)$;

⑨ $s\leftarrow Simpson(\boldsymbol{S}_i)$;

⑩ 把子序列$\langle r,s\rangle_i$追加到 Seq_{outline} 中;

⑪ end if

⑫ end for

⑬ return Seq_{outline}.

利用本节提出的 SrcDFA 算法思想,结合算法 1、算法 2 以及算法 3 的具体操作步骤即可进行最终的地图轮廓数据序列的生成工作.下面本文将给出具体的实验过程阐述本文所提模型以及采用方法的有效性.

4　实验仿真

4.1　数据采集

本文所用数据主要包括 2 部分:1)数据用来提取转向模板以及构建有效段的分类模型;2)数据主要用于验证本文所提地图轮廓生成方法的有效性,这部分数据无存储限制,即该数据可以是移动端实时采集得到的数据,也可以是已经采集并存储到移动端用于线下生成地图轮廓的数据,为了方便观察处理过程并展示处理结果,本文选择使用采集后存储的数据验证本文方法的有效性.以动作出现的大致频率为参考,采集的数据中包括 305 390 条 walk 数据、33 000 条 stairs 数据、88 975 条 static 数据.在采样频率为 50 Hz、分段段长为 12 s 的情况下,上述数据对应的分段数量分别为 897,96,269 段.

4.2　数据处理

本课题研究的目标是以提高传感器数据的利用率为目标,深入挖掘传感器数据信息进行地图轮廓的构建.因为所有的传感器数据均对应于同一时间序列,所以本文将基本的数据处理过程的参数进行了统一设定,主要的基本操作包括无效数据段的删除、转向数据以及用于构建识别模型数据的分段、噪声的去除等.相同参数的数值设定如表 2 所示:

Table 2　Settings for Basic Parameters

表 2　基本操作过程对应的参数设置

Parameters	Notation	Value	Description
Sampling Frequency	f/Hz	50	Data sampling frequency set by data collection software
Truncate Data Duration	D_{seg}/s	1	The length of the invalid data segment
Segment Unit Length	$winLen$/s	12	The length of one segment
Overlap Rate	o	0.5	Overlap rate between data segments

4.2.1　方向数据的处理及转向模板的提取

对于原始采集得到的数据,本文首先以既定的无效段长的截断阈值将原始数据前 1 s 的数据进行删除;然后将方向数据进行单独提取.通过对实际路况转向情况的分析发现,常见转向数据段的形成一般可以由如图 7 中 3 条路线对应的 6 段数据生成(每段路线均包含顺时针与逆时针 2 个方向的数据).

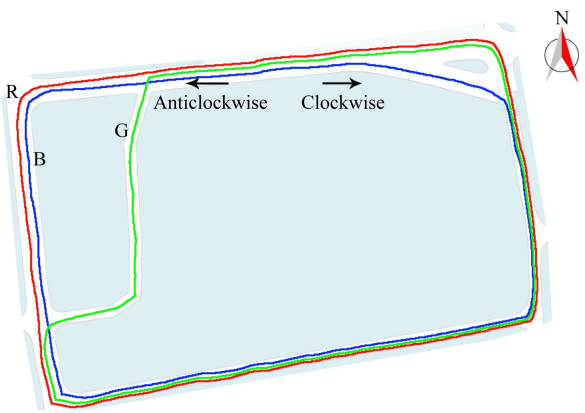
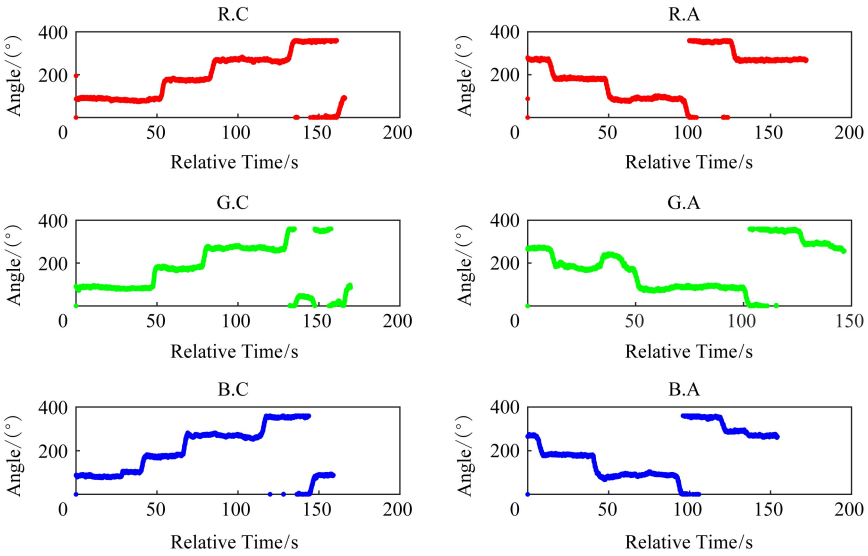


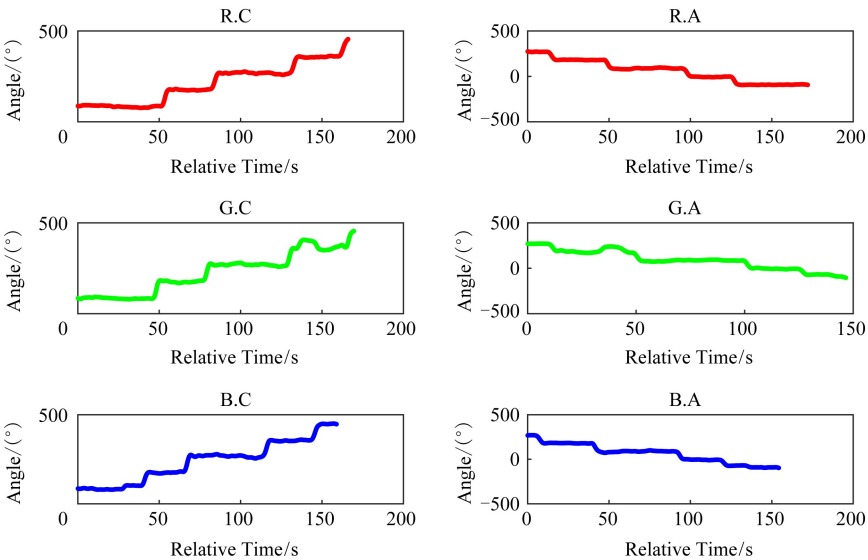
Fig. 7　Display of the actual roads

图 7　实际路况示意图

图 7 路段对应的实际偏航角序列如图 8(a)所示,其中,右向箭头指示方向为顺时针(clockwise)方向,标记为 C,左向箭头指示逆时针(anticlockwise)方向,标记为 A.则 R.C 表示顺时针沿路线 R 步行采集得到的数据,R.A 表示逆时针沿路线 R 采集得到的数据,其他 4 幅类似.由图 8(a)可以发现,当行人沿北向行进时,方向传感器数据都有大量跳变数据产生(无论顺时针、逆时针),基于合理的假设以及实际情况可知,路段的拐弯角度一般设计为直角或钝角.在本文中,密度传递的阈值可设为(90°,360°)范围内任一数值(本文选取为 300°).根据算法 1,本



(a) Corresponding raw yaw data



(b) After eliminating fluctuations

Fig. 8 The yaw corresponding to 6 routes

图 8 6 条路线对应的偏航角数据

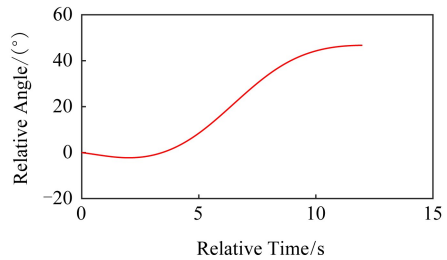
文首先消除方向传感器中偏航角的跳变现象,然后进行了数据的平滑处理,得到如图 8(b)所示结果.然后基于消除跳变后的数据进行了转向段跨度的计算,并提取了常见跨度下的顺时针、逆时针模板,本文将模板跨度阈值分布设定为±45°与±90°,其中,顺时针转向为正,逆时针转向为负.跨度计算结果如表 3 所示,其中,*Maximum* 表示每个模板序列中的最大值,*Minimum* 表示每个模板序列中的最小值,*Span* 表示该模板的数据跨度值.

根据计算所得跨度以及行进的方向,本文将转向情况进行了简化以及编号,最终形成了如图 9 所示的 4 个转向模板.图 9 中数据已进行处理,原始模板中的绝对数值被简化为相对数值.

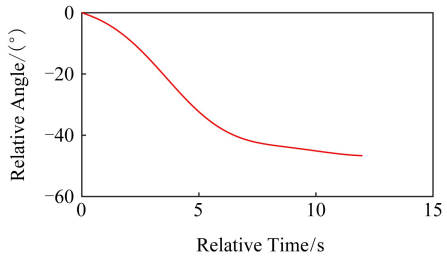
Table 3 Span Calculation Results for Common Steering Conditions

表 3 常见转向情况的跨度计算结果

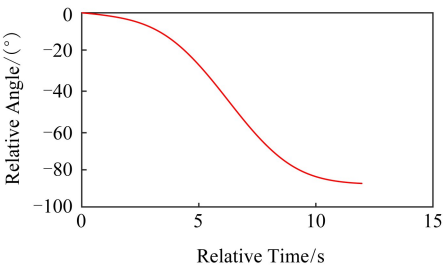
Sequence	*Maximum*	*Minimum*	*Span*
Temp1	406.553 3	357.614 3	48.939
Temp2	396.125 1	349.461 2	46.663 83
Temp3	88.278 79	0.365 605	87.913 18
Temp4	360.439 6	262.919	97.520 52



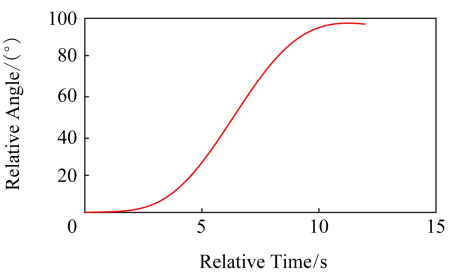
(a) Steering Template1 *Temp*1



(b) Steering Template2 *Temp*2



(c) Steering Template3 *Temp*3



(d) Steering Template4 *Temp*4

Fig. 9　The steering templates after simplifying

图 9　简化后的转向模板

4.2.2　分类模型的构建

在构建分类模型之前,本文先进行了训练集的生成工作.根据表 2 拟定的处理参数和算法 2 的构建过程,本文首先将加速度传感器数据和陀螺仪数据进行了分段操作,然后根据前文所述分类特征,本文计算了每段数据对应的特征值,所计算的特征值如表 4 所示.

通过对表 4 所述特征集的计算,本文根据算法 2 生成了全特征集下的训练数据集,继而利用 10 倍交叉验证方法,本文进行了分类模型的构建.根据前文所述,本文同时利用信息增益思想对上述特征进行了筛选并基于筛选后的特征集进行了分类模型的构建.其中,筛选后的特征集如表 5 所示.表 6 和图 10 分别给出了特征筛选前后所构建的分类模型的性能.

Table 4　Different Features on Accelerometer & Gyroscope

表 4　基于加速度传感器数据和陀螺仪数据计算的特征集

Categories	Features
Time Domain	x,y,z:min,max,average abs(x,y,z):min,max,average a:min,max,average
Frequency Domain	x,y,z:DC component,peakAmp,peakFreq a:DC component,peakAmp,peakFreq
Statistics	x,y,z:range,kurtosis

Table 5　Selected Features After Attribute Selection

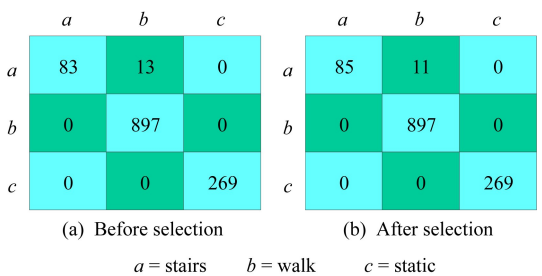
表 5　经过特征筛选后的特征集

Categories	Sensors	
	Accelerometer	Gyroscope
Time Domain	y:min z:max a:min,max	x,y,z:mean a:min,mean
Frequency Domain	y:peakAmp,peakFreq z:peakAmp a:DC component,peakAmp	z:peakAmp a:DC component
Statistics	y:range,kurtosis z:range,kurtosis a:range	y:kurtosis z:range,kurtosis

Table 6　Comparison of Basic Information About the Model Before and After Attribute Selection

表 6　特征筛选前后构建的分类模型基本信息对比

Information	Before Selection	After Selection
Total Number of Instances	1 262	1 262
Attributes	85	27
Time Taken to Build Model/s	3.73	2.39
Correctly Classified Instances	1 249	1 251
Accuracy/%	98.97	99.13



(a) Before selection　(b) After selection

Fig. 10　Comparison of confusion matrices before and after attribute selection

图 10　特征筛选前后构建的分类模型的分类矩阵对比

4.2.3　地图轮廓序列的生成

基于 4.2.1 节与 4.2.2 节所述操作过程获取的

转向模板库以及建立的识别模型,利用 SrcDFA 算法,本文进行了地图轮廓序列的生成.为了验证本文所提方法的性能,方便对比轮廓数据序列的生成效果,本文选取了已在地图上标识出的路段作为该部分的实验区域.为了降低算法 SrcDFA 在匹配路段方面的时间消耗,本文在匹配路段的过程中,引入了分类比较的方法进行代码优化,以降低算法的时间复杂度.即在分段操作时,首先计算出每个方向数据分段的跨度绝对值,然后以该值为依据,依次对比模板段数据的阈值.具体操作如下.

由于转向段的随机性使得转向模板的期望值能接近既定阈值,而无法严格生成与既定跨度值一致的模板.例如,当转向模板阈值定为 45°时,只能产生接近 45°的模板段,而无法严格生成一个跨度恰好为 45°的模板段.所以本文将模板跨度中接近 45°的转向数据段组合为一阈值$[\alpha_1,\alpha_2]$,即图 9 所示的 $Temp2$ 与 $Temp1$ 对应的跨度值;转向模板段中接近 90°的转向数据阈值为$[\alpha_3,\alpha_4]$,即图 9 所示的 $Temp3$ 与 $Temp4$ 对应的数值;则,对于跨度绝对值为 β 的数据段进行 3 种情况判定:1)当 $\beta<\alpha_1$,判定当前状态为直行,直接计算该段数值;2)当 $\alpha_1\leqslant\beta<(\alpha_2+\alpha_3)/2$,对比模板 1 与模板 2,选取 DTW 结果最相似的模板作为该段数据的转向情况,然后将其与该分段的数值组合为二元组,追加到轮廓序列中;3)当 $(\alpha_2+\alpha_3)/2\leqslant\beta$,对比模板 3 与模板 4,选取 DTW 结果最相似的模板作为该段数据的转向情况,然后将其与该分段的数值组合为二元组,追加到轮廓序列中.

通过引入分类比较的思想,转向识别过程的判定操作可由循环比较减至常数比较,时间复杂度由理论的 $O(N)$降到 $O(1)$,这在后期的转向模板增加的过程中,将能够节省大量的时间消耗.

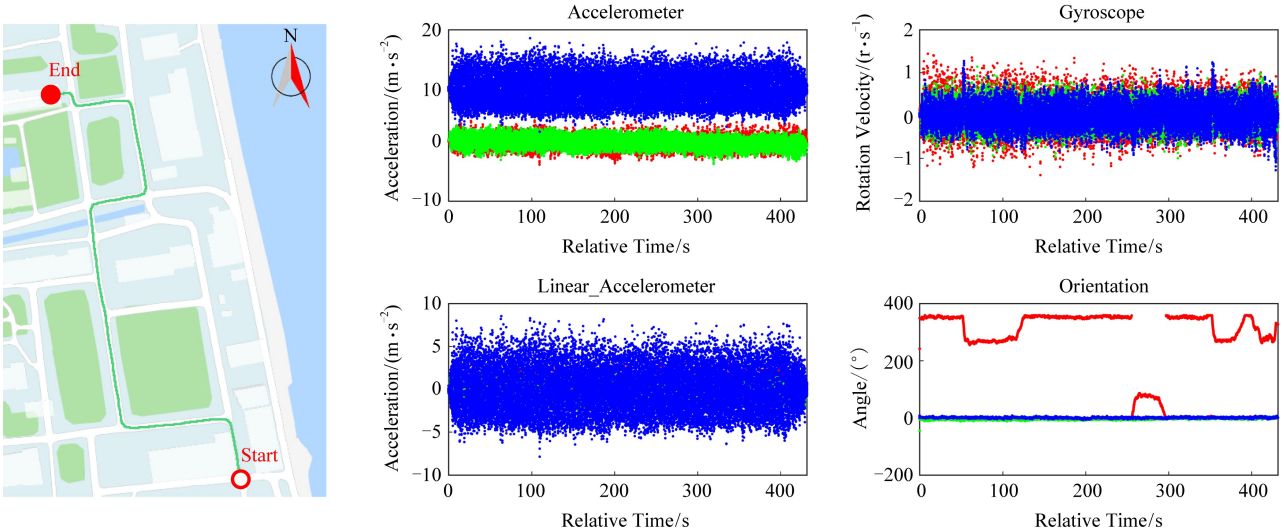
本文在该部分的实验区域主要包括 2 段(截选至百度地图,并且路段信息已经隐匿),分别表示在图 11 与图 12 中,其中,图 11(a)、图 12(a)、为路段对应的实际地图,图 11(b)、图 12(b)为路段对应的数据图,序列Ⅰ与序列Ⅱ分别为利用本文所提方法生成的地图轮廓序列.此外,本文利用百度地图自带测距工具对非转向路段的距离进行了测绘,并定义了平均准确率(mean accuracy rate, MAR)来度量本文数值计算的精度:

$$MAR=1-\frac{1}{n}\sum_{i=1}^{n}\frac{|L_t^i-L_a^i|}{L_a^i},\tag{6}$$

其中,n 表示非转向路段的总段数,L_t^i 与 L_a^i 分别表示当前第 i 分段的计算距离和实测距离.

路段 1 对应的序列Ⅰ:〈354.68, 40.00〉$_1$, 〈$Temp3$, 9.91〉$_2$, 〈267.28, 59.44〉$_3$, 〈$Temp4$, 9.63〉$_4$, 〈357.75, 119.52〉$_5$, 〈$Temp1$, 8.93〉$_6$, 〈75.64, 28.43〉$_7$, 〈$Temp3$, 9.85〉$_8$, 〈354.52, 50.00〉$_9$, 〈$Temp3$, 10.00〉$_{10}$, 〈271.29, 19.79〉$_{11}$, 〈$unknown$, 29.62〉$_{12}$, 〈276.84, 9.98〉$_{13}$.

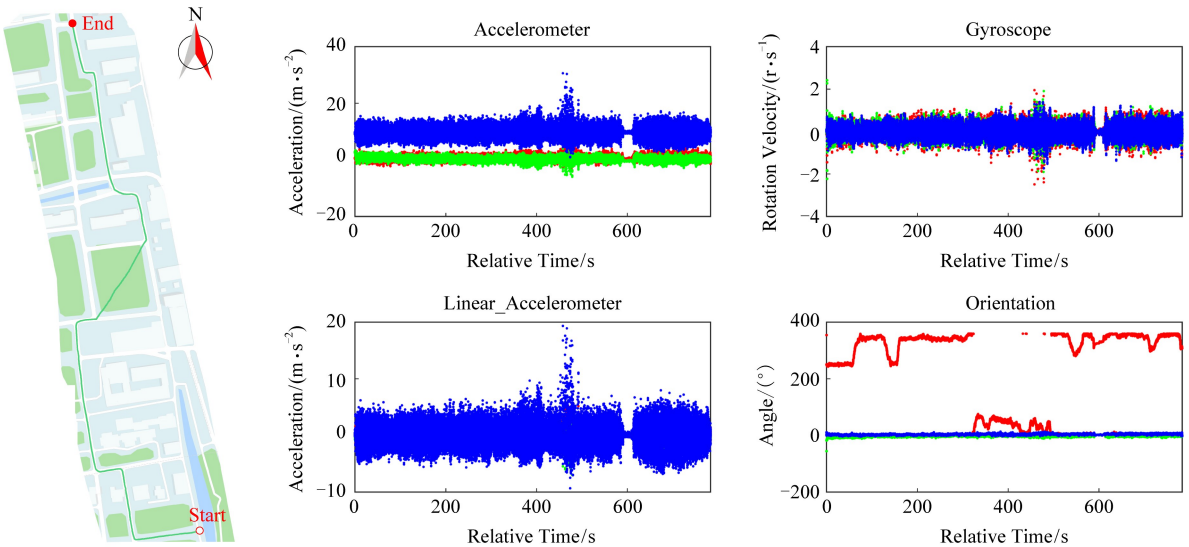
由实际观测和上述计算的序列可发现非转向路段包括段 1,3,5,7,9,11,13 共 7 段.对应的实测距离分别为 41 m,59 m,150 m,30 m,56 m,25 m,9 m.序列Ⅰ测得非转向路段的平均准确率 $MAR_{\text{Ⅰ}}$ 为 89.83%.



(a) Actual road map (b) Corresponding sensor data

Fig. 11 The information about testing road 1

图 11 实验路段 1



(a) Actual road map　　(b) Corresponding sensor data

Fig. 12　The information about testing road 2

图 12　实验路段 2

路段 2 对应的序列 Ⅱ:〈250.42, 48.75〉$_1$, 〈*Temp*4, 19.70〉$_2$, 〈344.23, 58.62〉$_3$, 〈*Temp*3, 19.25〉$_4$, 〈254.48, 10.02〉$_5$, 〈*Temp*4, 19.40〉$_6$, 〈342.12, 137.14〉$_7$, 〈*Temp*4, 19.26〉$_8$, 〈65.00, 9.79〉$_9$, 〈*Temp*2, 9.6458〉$_{10}$, 〈54.35, 80.85〉$_{11}$, 〈*Temp*2, 19.03〉$_{12}$, 〈39.06, 9.60〉$_{13}$, 〈*Temp*2, 28.29〉$_{14}$, 〈38.30, 36.91〉$_{15}$, 〈*Temp*3, 15.05〉$_{16}$, 〈*Temp*4, 19.76〉$_{17}$, 〈352.83, 126.25〉$_{18}$, 〈*Temp*1, 9.70〉$_{19}$, 〈356.84, 49.70〉$_{20}$.

序列 Ⅱ 的非转向路段为 1,3,5,7,9,11,13,15,18,20 共 10 段,对应的实测距离分别为:66 m,78 m,18 m,177 m,13 m,92 m,10 m,44 m,136 m,53 m.序列 Ⅱ 的平均准确率 $MAR_{Ⅱ}$ 为 81.18%.

4.3　实验结果分析

基于实际路况采集的数据和本文所提方法,如图 8 所示结果,本文提出的消除转向数据跳变的方法不仅达到了去除跳变的目的,同时也保留了原始动作的波动走势,为后续的模板对比方法提供了基础.而基于 DTW 的模板对比方法可以准确识别出转向过程所属类别,为准确刻画转向信息提供了方法;此外,利用 DTW 进行转向识别较以往的方向处理工作而言更具有实际意义,其主要原因在于,行人在路口转弯过程,往往是缓慢转变的过程,而本文将转向过程以数据段表示,利用 DTW 方法进行对比识别,更贴近于实际情况.从上述跳变消除过程到转向模板的提取以及匹配过程可以发现,转向精度的提高只需采集更多实际路况的转向数据,再利用本文方法进行模板库的扩充即可.一般地,模板库越丰富,转向精度处理的也越细致.而关于模板对比的速度,本文也给出了时间复杂度为 $O(1)$ 的操作步骤.

通过生成训练数据集以及构建分类模型的过程可以发现,使用信号处理方法计算得到的特征集能够达到有效分类的目的.本文在原始特征集上进行了特征筛选,对比筛选前后的各项信息可以发现,利用信息增益思想进行特征筛选,通过选取对有效分类贡献为 0.9 以上的单项特征进行组合,最后的特征数量较初始特征数量减少了 58 个,建模时间节省了 1.34 s,分类精度提高至 99.13%.通过该实验结果可以得出在处理传感器数据的过程中,本文所采用的分类识别处理方法具有一定的普适性和有效性.

通过观察最终生成的轮廓序列 Ⅰ 与实际路况图 11 的对比可以发现,本文所提方法可以达到准确识别路段轮廓的目的.其中序列 Ⅰ 中子序列 12 对应的路段检测结果为未知转向段,主要原因在于路段的拐弯处较集中(观察图 11 可以发现该处路况),而本文采用的分段段长为 12 s,对该连续转弯路段进行了全覆盖,所以导致了该情况的发生,但通过对子序列 12 前后序列进行数据观测可以发现,虽然该路段的转向情况未知,但依然可以通过对比前后序列的行进方向推测出该序列的整体趋向在 271°方向.

通过生成的轮廓序列Ⅱ与图 12 进行对比可以发现，序列Ⅱ中，子序列 11，12，13 中的 stairs 状态以及子序列 18 中的 static 状态均被正确识别，而且对于地图上未绘制的路段，本文的方法也能够进行初步的判别与数值的计算.

对于数值计算的精度判定方面，目前还没有公开的精准距离可供参考，但通过百度地图测距工具获得的实测距离结果可以发现，本文的数值计算在精度方面可以达到获取地图轮廓信息的目的.

5 总 结

本文通过对经典的基于传感器工作的学习研究，提出了对持续生成的传感器数据进行深入挖掘的思想.基于该思想，本文将机器学习方法以及信号处理技术应用到地图轮廓的构建工作中，通过对问题的简化以及子问题的分析，提出并实现了一种基于状态识别的数据融合算法(SrcDFA).其中，构建得到的状态识别模型可以达到 99.13%的识别精度.而基于跳变消除方法的模板提取算法以及基于数值积分计算方法的有效性在实际的路段数据处理过程中均得到了验证.在真实数据集上的实验结果表明，本文所提方法在地图轮廓构建工作上具有一定的有效性和可行性.此外，本文的工作较经典的基于智能手机传感器的工作在数据利用率上具有明显的提高，所提方法和处理步骤为后续对持续生成的传感器数据进行深入挖掘的工作奠定了基础.

在未来的研究中，我们将进一步利用传感器数据对路段的额外属性(路宽、坡度等)进行挖掘和分析，同时不断探索精确度更高的数值计算方法以及提取代表性更强的转向模板以提高地图轮廓的生成精度.

参 考 文 献

[1] Zhang Junping, Wang Feiyue, Wang Kunfeng, et al. Data-driven intelligent transportation systems: A survey [J]. IEEE Transaction on Intelligent Transportation Systems, 2011, 12(4): 1624-1639

[2] Ge Yong, Xiong Hui, Liu Chuanren, et al. A taxi driving fraud detection system [C] //Proc of the 11th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2011: 181-190

[3] Rao B, Minakakis L. Evolution of mobile location-based services [J]. Communications of the ACM, 2003, 46(12): 61-65

[4] Chen Chao, Zhang Daqing, Castro P S, et al. Real-time detection of anomalous taxi trajectories from GPS traces [C] //Proc of the Int ICST Conf on Mobile and Ubiquitous Systems. Berlin: Springer, 2012: 63-74

[5] Li Peng, Huang Xinhan, Wang Min. Multi-robot map building based on hybrid DSm model [J]. Journal of Computer Research and Development, 2009, 46(1): 70-76 (in Chinese)
(李鹏，黄心汉，王敏. 基于混合 DSm 模型的多机器人地图构建[J]. 计算机研究与发展，2009，46(1)：70-76)

[6] Lane N D, Miluzzo E, Lu Hong, et al. A survey of mobile phone sensing [J]. IEEE Communications Magazine, 2010, 48(9): 140-150

[7] Wang Feng, Wang Yasha, Wang Jiangtao, et al. Mental stress assessment approach based on smartphone sensing data [J]. Journal of Computer Research and Development, 2019, 56(3): 611-622 (in Chinese)
(王丰，王亚沙，王江涛，等. 基于智能手机感知数据的心理压力评估方法[J]. 计算机研究与发展，2019，56(3)：611-622)

[8] Lin H Y, Yao C W, Cheng K S, et al. Topological map construction and scene recognition for vehicle localization [J]. Autonomous Robots, 2018, 42(1): 65-81

[9] Cao Lili, Krumm J. From GPS traces to a routable road map [C] //Proc of the 17th ACM SIGSPATIAL Int Conf on Advances in Geographic Information Systems. New York: ACM, 2009: 3-12

[10] Bahl P, Padmanabhan V N. RADAR: An in-building RF-based user location and tracking system [C] //Proc of the IEEE INFOCOM'00. Piscataway, NJ: IEEE, 2000: 775-784

[11] Correa A, Barcelo M, Morell A, et al. A review of pedestrian indoor positioning systems for mass market applications [J]. Sensors, 2017, 17(8): 1927-1953

[12] Lee Y C, Myung H. Indoor localization method based on sequential motion tracking using topological path map [J]. IEEE Access, 2019, 7: 46187-46197

[13] Gallagher T, Wise E, Li Binghao, et al. Indoor positioning system based on sensor fusion for the blind and visually impaired [C] //Proc of the 2012 Int Conf on Indoor Positioning and Indoor Navigation (IPIN). Piscataway, NJ: IEEE, 2012: Article Number 175

[14] Kang W, Han Y. SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization [J]. IEEE Sensors Journal, 2014, 15(5): 2906-2916

[15] Chen Zhenghua, Zhu Qingchang, Soh Y C. Smartphone inertial sensor-based indoor localization and tracking with iBeacon corrections [J]. IEEE Transaction on Industrial Informatics, 2016, 12(4): 1540-1549

[16] Xue Guangtao, Zhu Hongzi, Hu Zhenxian, et al. Pothole in the dark: Perceiving pothole profiles with participatory urban vehicles [J]. IEEE Transaction on Mobile Computing, 2016, 16(5): 1408-1419

[17] Ismail M Z, Inoue M. Map generation to detect heat stroke by using participatory sensing data [C] //Proc of the 2018 Int Conf on Electronics, Information, and Communication (ICEIC). Piscataway, NJ: IEEE, 2018: 68-71

[18] Hao Tian, Xing Guoliang, Zhou Gang. RunBuddy: A smartphone system for running rhythm monitoring [C] // Proc of the 2015 ACM Int Joint Conf on Pervasive and Ubiquitous Computing. New York: ACM, 2015: 133-144

[19] Higgins J P. Smartphone applications for patients' health and fitness [J]. The American Journal of Medicine, 2016, 129 (1): 11-19

[20] Chen Zhenyu, Lin Mu, Chen Fanglin, et al. Unobtrusive sleep monitoring using smartphones [C] //Proc of the 7th Int Conf on Pervasive Computing Technologies for Healthcare. Piscataway, NJ: IEEE, 2013: 145-152

[21] Tao Tao, Sun Yu'e, Chen Dongmei, et al. PosAla: A smartphone-based posture alarm system design for smartphone users [C] //Proc of the 14th Int Conf on Mobile Ad-Hoc and Sensor Networks (MSN). Piscataway, NJ: IEEE, 2018: 115-120

[22] Kwapisz J R, Weiss G M, Moore S A. Activity recognition using cell phone accelerometers [J]. ACM SIGKDD Explorations Newsletter, 2011, 12(2): 74-82

[23] Gu Tao, Chen Shaxun, Tao Xianping, et al. An unsupervised approach to activity recognition and segmentation based on object-use fingerprints [J]. Data & Knowledge Engineering, 2010, 69(6): 533-544

Tao Tao, born in 1995. Master candidate at the School of Computer Science and Technology, Soochow University, China. Student member of CCF. His main research interests include machine learning, sensor networks.

Sun Yu'e, born in 1983. Associate professor at the School of Rail Transportation, Soochow University, China. Member of IEEE, ACM and CCF. Her main research interests include network traffic measurement, spectrum auction, privacy preserving, and wireless networks.

Chen Dongmei, born in 1994. Master candidate at the School of Computer Science and Technology, Soochow University, China. Student member of CCF. Her main research interest is machine learning. (20184227026@stu.suda.edu.cn)

Yang Wenjian, born in 1995. Master candidate at the School of Computer Science and Technology, Soochow University, China. His main research interests include transportation traffic measurement and network traffic measurement. (wj__yang@outlook.com)

Huang He, born in 1983. Professor at the School of Computer Science and Technology, Soochow University, China. Member of IEEE, ACM and CCF. His main research interests include network traffic measurement, spectrum auction, privacy preserving, crowdsourcing. (huangh@suda.edu.cn)

Luo Yonglong, born in 1972. Professor at the School of Computer and Information, Anhui Normal University, China. His main research interests include information security and spatial data processing. (ylluo@ustc.edu.cn)