



用于 WLAN 室内定位的 PCA 聚类算法

杨明极, 刘恺悻, 邵丹

(哈尔滨理工大学测控技术与通信工程学院, 黑龙江 哈尔滨 150080)

摘要:在 WLAN 室内定位系统中, 针对接收信号强度(RSS)的时变特征降低室内定位精度的问题, 提出一种基于主成分分析(PCA)白化 RSS 的聚类算法, 该算法首先对信号强度进行 PCA 白化处理, 去除 RSS 信息的相关性, 提高聚类中心的可靠性和合理性, 然后通过 *K*-means 聚类方式对 RSS 信息进行聚类, 能够有效地提高聚类精度, 以此来提高定位精度。实验结果表明, 该算法相比于没有经过 PCA 的传统聚类算法, 能够使定位误差在 2 m 内的概率提高 44.8%, 性能更优良。

关键词:WLAN; 室内定位; 去相关性; 主成分分析; 聚类算法

中图分类号: TN929.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016186

PCA clustering algorithm for indoor positioning in WLAN

YANG Mingji, LIU Kaiyi, SHAO Dan

School of Measure-Control Technology and Communication Engineering, Harbin

University of Science and Technology, Harbin 150080, China

Abstract: In WLAN indoor location system, aiming at the problem of time-varying characteristic of received signal strength (RSS) which reduces indoor positioning accuracy, a clustering algorithm based on principal component analysis (PCA) albino RSS was put forward. The algorithm firstly treated the RSS with PCA whitening treatment to remove the correlation and improve reliability and rationality of the cluster centers. Then, *K*-means clustering method was used to cluster the RSS and the clustering accuracy was improved effectively, so as to improve positioning accuracy. Experimental results show that compared with the traditional clustering algorithm without PCA, probability of positioning error within 2 meters has improved 44.8% in positioning accuracy, and the performance of positioning system has been more excellent.

Key words: WLAN, indoor positioning, remove the correlation, PCA, clustering algorithm

1 引言

在通信技术迅速发展的今天, 人们对位置的需求越来越大, 无线定位服务在很多方面都有广泛的应用前景^[1]。目前应用最为广泛的定位方式是 GPS 定位和蜂窝网定位, 通常能够对室外环境有效地定位, 但是对于建筑严重遮蔽的

室内环境, 这两种定位方式都达不到理想的效果, 因此对室内定位的研究成为热门方向^[2]。

当前主要研究的室内定位技术有射频识别技术、超宽带技术、ZigBee 技术、红外线技术、超声波技术和 WLAN 技术等。这些技术中, 射频识别技术对信噪比要求高, 受限较严重; 超宽带技术虽然定位精度高, 但伴随的成本也较高,



不宜普及; ZigBee 技术成本较低,但是范围小,难以推广;红外线技术定位精度高,但传播距离短,易受荧光和房间灯干扰;超声波技术需要专门硬件,成本高且受非视距和多径传播影响大^[3]。与这些技术相比,基于 WLAN 室内定位有覆盖面积大、无需额外定位设备、部署方便等优点,所以近些年对 WLAN 定位技术的研究越来越多^[4]。

基于 WLAN 位置指纹的定位方法的核心是信号强度^[5]。一般来说,系统的定位精度是随着可利用的 AP 数量的增加而提高的,但并不是随 AP 的增加而无限增加^[6]。过多的 AP 数目,一方面导致 RSS 数据维数增加,影响解算速度,另一方面导致临近 AP 的 RSS 相关性较大^[7]。从位置指纹定位技术精度和效率的影响因素入手,提出一种将指纹数据库中 RSS 先进行 PCA 去相关处理,然后采用 K-means 聚类算法将指纹数据库分类的技术,从而使得聚类精确度更准确,较大地提高了定位效率和精度^[8]。

2 PCA 聚类算法

基于 WLAN 的室内位置指纹定位是利用定位空间中多个 AP 的 RSS 数据来描述位置信息^[9],图 1 为基于 PCA 聚类算法的 WLAN 室内定位系统架构。

该方法包括两个阶段:离线阶段和在线阶段。在离线过程中,第一步确定各个指纹的物理位置,然后在每个指纹点的位置上采集 RSS 数据,将位置信息和 RSS 信号值结合起来,形成对应的关系,该关系就是位置指纹,把位置指纹保存在数据库中,建立位置指纹数据库 (radio map, RM)^[10,11]。将位置指纹数据库中的 RSS 数据进行 PCA 去相关处理,然后进行 K-means 聚类。在线阶段,对定位终端接收到的 RSS 信息进行 PCA 去相关处理,通过最近邻 (nearest neighbor, NN) 法得到待测点所在的聚类类别,在该类别中用 KNN 算法匹配出测试点的估计位置。

2.1 PCA 变换

由于当前大部分室内 AP 部署密集,很有可能导致相同指纹位置邻近 AP 的 RSS 信息相关性较大,导致聚类过程中出现很多相关性高的聚类中心。因此,在聚类之前,对 RSS 做去相关处理、提高聚类中心的可靠性是很有必要的。

PCA 算法的基本原理是:根据霍特林变换在另一空间中确定一组正交向量,该正交向量可以最大限度地表现原始数据,从原来的 p 维空间将原始信息映射到该组正交向量形成的 $n(n < p)$ 维空间上,从而完成对数据的去相关和维数的压缩。

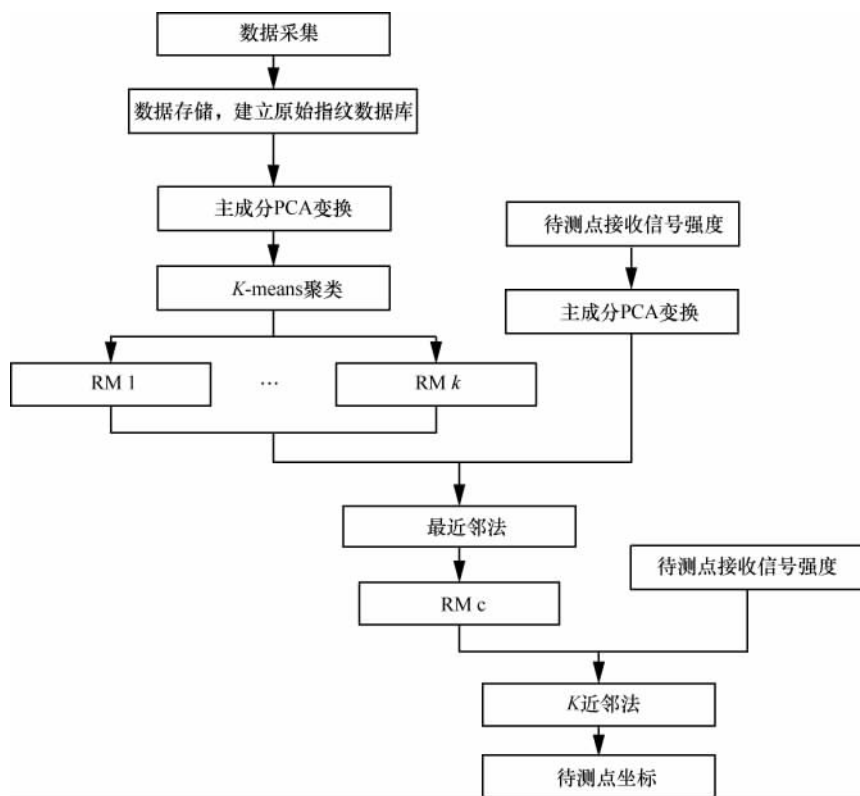


图 1 基于 PCA 聚类算法的 WLAN 室内定位系统架构

设样本集为 $R=\{(S_1, P_1), (S_2, P_2), \dots, (S_p, P_p)\}$, $S_t=(\bar{S}_{t1}, \bar{S}_{t2}, \dots, \bar{S}_{tp})$, $\bar{S}_{tp}$ 表示在指纹点 t 的位置接收到来自第 p 个 AP 的 M 次采集的 RSS 均值, p 为能够接收到的有效 AP 的个数, $P_t=(x_t, y_t)$ 是指纹点 t 的物理位置坐标。令 $X=[S_1, S_2, \dots, S_p]^T$, 则 PCA 变换为:

$$Z=A^T(X-\bar{X}) \quad (1)$$

其中, $\bar{X}=E[X]$, $A=[u_1, u_2, \dots, u_k]$, 是由 X 的协方差矩阵 $C_x=E\{(X-\bar{X})(X-\bar{X})^T\}$ 的特征值 γ_k 按由大到小排列后对应的特征向量 u_k 组成的。

通过 PCA 处理后, 基本消除了 Z 中各数据间的相关性, 保留 Z 中包含 X 的大部分信息的前 n 个主成分, $Z'=[S'_1, S'_2, \dots, S'_n]^T$, 其中, $S'_i=(s_{i1}, s_{i2}, \dots, s_{in})$, 因为 $n < p$, 因此同时完成了对数据的压缩。

2.2 聚类算法

RM 中的每个位置指纹均值通过 PCA 白化处理后, 新样本集为 $T'=\{(S'_1, L'_1), (S'_2, L'_2), \dots, (S'_n, L'_n)\}$, $Z=[S'_1, S'_2, \dots, S'_n]^T$, $S'_i=(s_{i1}, s_{i2}, \dots, s_{in})$ 。

白化的 K-means 算法的工作流程说明如下。

步骤 1 整个 RM 中选取 k 个指纹作为初始聚类中心 $(R_1, R_2, \dots, R_k)$, $R_k=[r_{k1}, r_{k2}, \dots, r_{kn}]$ 。由于初始聚类中心的随机选择可能导致聚类过于集中或过于分散, 因此在选择 k 个聚类中心时要尽量地在定位范围内均匀选择。

步骤 2 对于除 k 个聚类中心之外的其他所有经过 PCA 白化处理的 RSS 均值, 它们与每个聚类中心的欧式距离 d_i 为:

$$d_i=\sqrt{(s_{i1}-r_{i1})^2+(s_{i2}-r_{i2})^2+\dots+(s_{in}-r_{in})^2} \quad (2)$$

将它们分别分配给与其欧式距离最近的聚类。

步骤 3 将所有的指纹都分配完成后, 获得新的聚类, 因为 K-means 聚类算法可能会出现空的聚类, 为避免这种情况, 在每次迭代过程中, 将每次聚类中的全部 RSS 取平均值, 当作新的聚类中心, 该方法被称为聚类中心的抑制更新方法。

步骤 4 不断重复步骤 2 和步骤 3, 直到 k 个聚类中心固定不变, 终止迭代。

聚类过程结束, 每个位置指纹都被划分至与之最近的聚类中心, 每个聚类都被当做一个子区域。

2.3 匹配算法

离线定位阶段, 每个聚类中心和相对应的位置指纹数据形成一个独立的子位置指纹数据库。在线定位阶段,

将测试点处获得的 RSS $\{x_{i1}, x_{i2}, \dots, x_{in}\}$ 做 PCA 去相关处理, 然后与各个聚类中心进行比较分析, 利用最近邻法获取待测点所在的聚类类别。

$$d_k=\sqrt{(x_{k1}-r_{i1})^2+(x_{k2}-r_{i2})^2+\dots+(x_{kn}-r_{in})^2} \quad (3)$$

将 $d_1, d_2, \dots, d_k$ 从小到大进行排序, 选出最近的聚类中心所在的类族作为参考点的类族。选取的类族记为 $\{C_1, C_2, \dots, C_m\}$, m 表示该类中带权值的位置指纹参考点的个数。

在所确定的类别中, 各参考点与测试点进行算法匹配, 进而利用 KNN 算法实现对待测点的位置估计。

步骤 1 应用欧几里德距离计算测试点与聚类指纹数据集中的每个参考点 $\{C_1, C_2, \dots, C_m\}$ 之间的距离。其中, $C_m=[c_{m1}, c_{m2}, \dots, c_{mm}]$ 。

$$d_m=\sqrt{(x_{i1}-c_{m1})^2+(x_{i2}-c_{m2})^2+\dots+(x_{im}-c_{mm})^2} \quad (4)$$

步骤 2 对步骤 1 得到的 $\{d_1, d_2, \dots, d_m\}$ 由小到大进行排序, 选取前 k 个参考点。

步骤 3 对步骤 2 得到的 k 个指纹参考点的位置坐标求平均值, 把得到的值作为待测点的估算位置坐标。

改进后的算法在定位阶段只需跟位置指纹库中的某个子集进行计算比较, 使得 KNN 算法的匹配效率得到极大的提高, 更能满足客户实时性的要求。

3 实验测试

3.1 实验环境建立

为了验证 PCA 聚类算法的有效性, 在哈尔滨理工大学 8 公寓 3 楼进行实验, 3 楼局部平面示意图 2 所示。实验采用 13 个 CMCC 信号发射器作为 AP, 使用联想 Z460 的笔记本电脑采集各 AP 的 RSS 信号值, 信号采集软件为 inSSIDer。实验选取了 253 个指纹点和 50 个测试点, 相邻指纹点间距为 1 m, 50 个测试点的位置是随机选择的, 每个位置能接收到来自 5~7 个 AP 的 RSS, 采集 30 次为一组, 采集 2 组, 采集的时间间隔为 2 s。定义定位误差为:

$$err_i=\sqrt{(x_i-x'_i)^2+(y_i-y'_i)^2} \quad (5)$$

其中, (x_i, y_i) 表示第 i 个测试点的实际物理坐标, (x'_i, y'_i) 表示第 i 个测试点的估算物理坐标。

3.2 结果分析

在室内传播过程中 RSS 不可避免地受遮挡、多径效应等影响, 在同一位置上接收到 AP 的 RSS 会随时间变化存在程度不一的波动, 导致距离较近的指纹点之间的 RSS 信

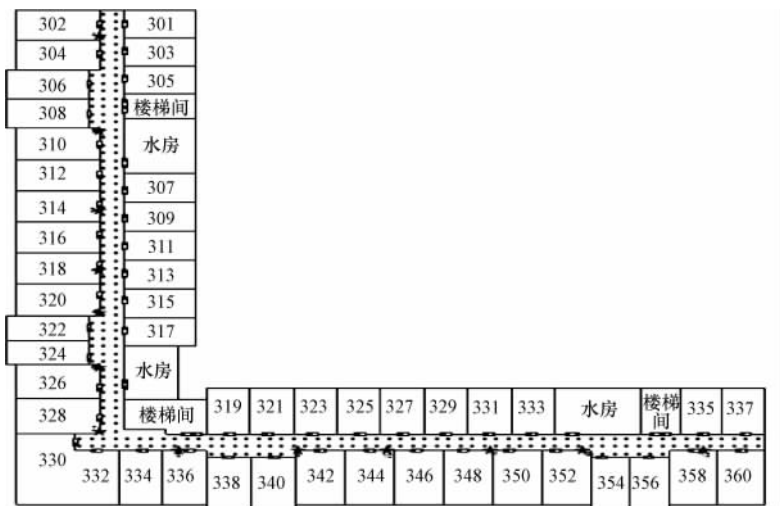


图2 WLAN室内定位环境

号均值相近^[15]。此时通过PCA白化处理数据,不仅可以解决以上问题,还将数据维度进行了压缩,增加K-means聚类算法的可靠性,同时为后面的聚类 and 匹配算法降低了复杂程度。在实验过程中,原始RSS数据的维数为6,经过PCA白化处理后,RSS值保留90%的主要信息,维数降为4($n=4$)。

首先,为了证明对RSS进行PCA白化处理有利于聚类的准确性,比较了有无经过PCA处理的K-means聚类的分块精度,如图3所示。

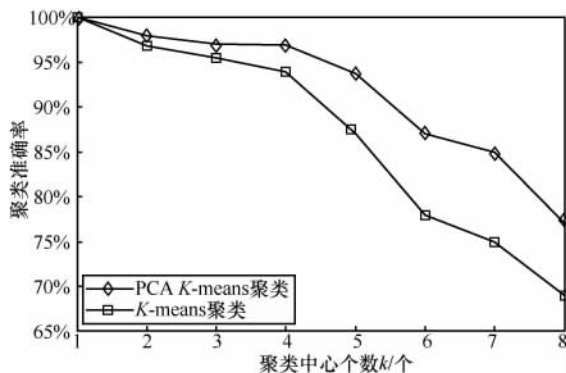


图3 K-means聚类算法经PCA前后聚类精度比较

图3为聚类中心由1到8时,经过PCA处理和没有PCA处理的聚类分块精度对比。首先,信号经过PCA处理后的K-means算法在聚类精度上要高于没有经过PCA处理的信号的聚类精度。特别是当k值慢慢增大,提高得就更加明显,因为在定位范围一定时,聚类中心k值越大,则定位子区域越多,因此相邻的子区域之间RSS强度相关性越大,通过PCA变换,消除信号之间的相关性,可以较大

地提升聚类精度。其次,从图3可以看出,k的取值越小,聚类的准确率越高,但是k取值过小,类内的信息相似程度低,不利于提高室内定位精度和降低计算复杂度。当 $k=4$ 时,聚类的准确率达到96.9%。随着k值增大,所有PCA前后的K-means聚类算法的聚类精度都在快速下降,因为k的取值在大于5以后,k值越大,分块越多,每个类的定位子区域越小,各个类之间的相似度越大,不但会降低聚类准确率,还会降低室内定位精度。

如图4所示是当聚类中心k的取值由1取到8时对应的定位精度概率累积分布。

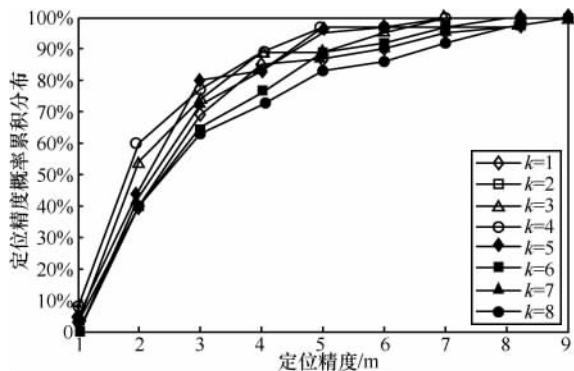


图4 k取不同值时系统定位精度概率累积分布

由图4可以发现,聚类算法可以提高定位精度,当聚类中心k的取值为2、3、4、5时都高于 $k=1$ (无聚类)时的定位精度。但是聚类中心取值大于6时,低于无聚类时的定位精度概率,聚类算法的优势渐渐消失。其中,聚类中心 $k=4$ 时,整体的定位精度高于k取其他值,虽然 $k=4$ 时,定位精度小于3m概率不如 $k=5$ 时高,但是并

不影响 $k=4$ 时系统定位的整体优势。实际上, $k=5$ 时累计概率最高为 80.1%, $k=4$ 时累计概率为 77.2%, 虽然此处 $k=5$ 时的概率略高于 $k=4$ 时, 但是由图 3 可知, $k=5$ 时的聚类精度概率为 93.8%, 明显比 $k=4$ 时的聚类精度概率 97.1% 要低, 并且 $k=5$ 时在定位精度小于 2 m 和小于 4 m 处的累积概率并不高, 整体考虑定位精度和聚类精度两个因素, 当聚类中心值 $k=4$ 时定位系统的可靠性和有效性最高。

以上分析说明, 聚类中心 k 的取值严重影响系统的定位性能, 因此为了既能提高定位精度, 缩小搜索的定位范围, 又能保证聚类精度和聚类有效性, 将聚类中心取为 4。当 $k=4$ 时, 比较了 K -means 聚类算法在 RSS 进行 PCA 处理前后的定位精度, 如图 5 所示。

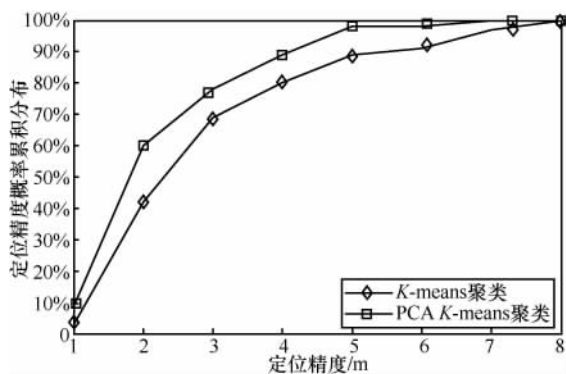


图5 最优聚类下信号 PCA 处理对定位精度的影响

从图 5 可以看出, 经过 PCA 处理后的系统在 2 m 内的定位精度概率达到 60.1%, 比没有经过 PCA 的系统定位精度概率提高了 44.8%; 经过 PCA 处理处理后的聚类定位精度在 3 m 内的概率为 77.3%, 比没有经过 PCA 处理的定位精度概率高 11.2%; 在定位精度为 1 m 时定位精度概率也有所提高。所以, 信号经过 PCA 白化处理后再进行聚类, 能够明显提高定位精度。

为了更清晰地说明 PCA 的作用, 图 6 对 PCA 处理前后系统定位误差的最大值、最小值、平均值和定位误差的标准方差做了对比。

由图 6 可发现, 最大定位误差由未经过 PCA 的 7.92 m 下降到 6.28 m, 降了 26.1%; 而信号经过 PCA 处理后系统的平均定位误差为 2.22 m, 未经过 PCA 处理的平均误差为 2.75 m, 定位误差下降了 23.9%, 此外, PCA 处理前的误差范围为 0.92~7.92 m, 经过 PCA 处理的聚类定位误差范围为 0.62~6.26 m, 由于定位误差范围越小, 则误差与均值

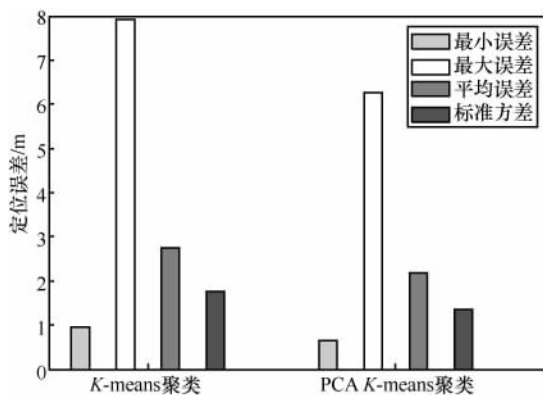


图6 最优聚类下信号是否 PCA 的定位误差比较

的偏离度越小, 定位误差的标准方差由未经过 PCA 处理的 1.86 m 降至 1.36 m, 表明 PCA 算法可以降低偏离平均定位误差的程度。

4 结束语

通过对实验结果的分析可以看出, PCA 能够有效地对接收信号强度去除相关性和进行维数压缩; 经过 PCA 的 K -means 聚类算法比没有经过 PCA 的聚类算法聚类精度更高, 并且当聚类中心 $k=4$ 时既能保证定位精度, 又能保证可靠性最高。在最优聚类情况下进行 PCA 处理能有效地提高定位精度, 给用户带来更好的定位结果。

参考文献:

- [1] FANG S H, LIN T N. A dynamic system approach for radio location fingerprinting in wireless local area networks [J]. IEEE Transactions on Communications, 2010, 58(4): 1020-1025.
- [2] BORENOVIC M, NESKOVIC A, BUDIMIR D. Space partitioning strategies for indoor WLAN positioning with cascade-connected ANN structures [J]. International Journal of Neural Systems, 2011, 21(1): 1-15.
- [3] GU Y, LO A, NIEMEGER I. A survey of indoor positioning systems for wireless personal networks[J]. IEEE Communications Surveys and Tutorials, 2009, 11(1): 13-32.
- [4] SUN Y, LA PORTA T F, KERMANI P. A flexible privacy-enhanced location-based services system framework and practice[J]. IEEE Transactions on Mobile Computing, 2009, 8(3): 304-321.
- [5] CHON M, CHA H. Life map: a smartphone-based context provider for location-based services [J]. IEEE Pervasive Computing, 2011, 10(2): 58-67.



- [6] TAGASHIRA S, KANEKIYO Y, ARAKAWA Y, et al. Collaborative filtering for position estimation error correction in WLAN positioning systems [J]. IEICE Transactions on Communications, 2011, 94(3): 649-657.
- [7] SERTTHIN C, FUJII T, OHTSUKI T, et al. Multi-band received signal strength fingerprinting based indoor location system[J]. IEICE Transactions on Communications, 2010, 93(8): 1993-2003.
- [8] KUSHKI A, PLATANIOTIS K N, VENETSANOPOULOS A N. Intelligent dynamic radio tracking in indoor wireless local area networks[J]. IEEE Transactions on Mobile Computing, 2010, 9(3): 405-419.
- [9] YOO J W, PARK K H. A cooperative clustering protocol for energy saving of mobile devices with WLAN and bluetooth interfaces[J]. IEEE Transactions on Mobile Computing, 2011, 10(4): 491-504.
- [10] CHAVALIT K, KOMWUT W, KAMOL K. Indoor localization improvement via adaptive RSS fingerprinting database[C]// International Conference on Information Networking, Jan 28-30, 2013, Bangkok, Thailand. New Jersey: IEEE Press, 2013: 412-416.
- [11] WANG T. Novel sensor location scheme using time-of-arrival estimates[J]. IET Signal Processing, 2012, 6(1): 8-13.

[作者简介]



杨明极(1971-),男,博士,哈尔滨理工大学教授,主要研究方向为计算机网络和嵌入式系统。



刘恺(1992-),女,哈尔滨理工大学硕士生,主要研究方向为无线通信技术。



邵丹(1988-),女,哈尔滨理工大学硕士生,主要研究方向为无线通信技术。