分类号 _____ 密级 _____

UDC _____ 编号 _____

中国科学院研究生院
工学硕士学位论文

基于迁移学习的无线位置感知技术研究

孙卓

指导教师 陈益强 副研究员

中国科学院计算技术研究所

申请学位级别 工学硕士 学科专业名称 计算机应用技术

论文提交日期 2009年4月 论文答辩日期 2009年6月

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院研究生院

答辩委员会主席 陈熙霖 研究员

摘　　要

无线位置感知技术研究利用无线信号确定和跟踪移动设备的位置，是普适计算中的一项重要技术。随着Wi-Fi接入点的广泛覆盖，基于Wi-Fi的室内外定位系统已经成为热门的研究领域。本文从机器学习的方法入手对无线定位技术进行了深入研究，主要采用贝叶斯统计、流形学习等方法对信号模式进行描述与挖掘。例如：通过将传统的统计定位方法与判别规则相结合，进一步提高了单一依靠统计方法的定位精度；利用不同时间不同设备的数据共享流形分布的特性，在不同领域间进行知识迁移，解决了Wi-Fi信号分布差异导致定位精度下降的问题。较之已有工作，本文在无线定位的精确度、稳定性以及所需的训练样本数上都有不同程度的改进，并将所提出的方法应用于实际系统中。

具体包括以下几部分工作：

基于概率统计的定位方法通过采集大量数据学习信号分布特性，适用于粗纹理的快速定位，而基于规则的定位方法在局域区域建立复杂规则，适用于细纹理的准确定位。本文提出一种统计和规则结合的无线定位方法，在传统的统计方法之上，加入规则库作为约束项，能够有效提升单一方法的精度。特别是基于规则的隐马尔科夫模型，在概率转移上加入规则约束，达到准确跟踪用户轨迹的目的。

已有定位模型面临信号分布变化时定位精度会急剧下降，如何只用较少的有标记数据更新模型而能保持较高的定位精度是一个难点问题。本文从信号传播的衰减特性和信号空间的流形假设出发，提出了一种流形对齐的降维方法，通过构造两种不同分布的数据集在低维空间的对齐流形，实现不同领域间的知识迁移。实验结果表明，基于该迁移学习方法的自适应室内定位算法能够在有限的有标记数据的帮助下，有效的提高目标领域中的定位精度，从而大大减小了人工标记数据的工作量，增强了定位系统的实用性。

在已有的定位跟踪原型系统基础上，本文还实现了一种基于位置的多媒体服务原型系统，并且在实际无线环境中进行了测试，能够根据移动用户的位置提供定制化的视频服务。

关键词： 定位计算，贝叶斯统计，迁移学习，流形降维

Research on Wireless Location Estimation Techniques Based on Transfer Learning

Sun Zhuo
Directed by Prof. Chen Yiqiang

The wireless location estimation technique, which aims to determine and track the position of the mobile device through wireless network signals, is important in pervasive computing. As Wi-Fi access points become more and more widely deployed, the indoor and outdoor localization systems based on Wi-Fi signals become one of hot research topics. In this thesis, we study machine-learning-based wireless localization techniques. In particular, we utilize the statistical methods like Bayesian inference and manifold learning to characterize and mine the wireless signal patterns. For instance, we propose to combine the traditional statistical localization methods with rule-based methods to improve the localization accuracy. We also take advantage of the characteristic that different domains share a common manifold to transfer knowledge between them, and solve the problem of accuracy reduction due to Wi-Fi signal distribution changing. Compared with the previous work, our proposed methods give improvements on the localization accuracy, stability and the amount of necessary training samples. Some methods have been implemented in real-world applications.

The following gives the details of the work of this thesis:

The statistical methods, which learn the signal distribution through a great deal of data, are good at fast coarse-grained localization, while the rule-based methods, which create sophisticated rules at local areas, are more suitable for fine-grained localization. Based on the statistical approaches, we add rule terms as constraints to increase localization accuracy. Specially, the Rule-based Hidden Markov Model, in which the transition probability is restricted by rules, can track the user's traces accurately.

The localization accuracy of existing models would drop dramatically when

signal distribution changes. It is difficult to use a small amount of labeled data to update the models while maintaining a high accuracy. Considering the characteristics of signal propagation and the manifold assumption of signal space, we propose a dimension reduction method by manifold alignment. The knowledge is transferred between two domains via their aligned low-dimensional manifold. The experimental results show that the adaptive localization algorithm is able to improve the accuracy of the target domain with the help of limited labeled data. Thus, it greatly reduces the human labeling efforts and enhances the practicability of real-world systems.

Based on our localization and tracking system, a location-based multimedia service prototype system is also proposed and implemented in real wireless environment, which provides personalized video service according to the location of the mobile user.

Keywords: Location estimation, Bayesian inference, transfer learning, dimension reduction

目　　录

表　　格

插　　图

第一章　引言

位置信息的重要性和时间一样，决定着我们的日常生活。我们总是需要明确知道自己的位置信息，在什么地方，要去什么地方，应该有一条怎样的路线，这些看似简单的信息却有着重要的意义。在普适计算环境中，位置信息同样扮演了重要的角色，其中最典型的就是基于位置的服务（Location-based Service，LBS），也称作位置感知服务（Location-aware Service）或位置相关服务（Location-related Service）[23]。LBS通常定义为根据目标的位置而为其提供的增值服务，在日常生活中有着广泛的应用，包括目标跟踪[30]、人群定位[3, 50, 27]、消息递送[7]、安全加密[42]、环境监测[40]、行为识别[48, 26]等。尽管这些增值服务各有不同，它们的核心问题一样，都在于确定可靠的物理位置，即定位技术。

近年来，定位技术发展迅速，出现了各式各样的定位跟踪系统，例如全球定位系统（Global Positioning System，GPS）、无线蜂窝（Cellular-based）定位系统、红外（Infrared-based）定位系统、超声波（Ultrasound-based）定位系统等等。这些系统有着各自的优缺点：GPS采用卫星信号进行全球跟踪定位，主要应用于室外，无法用于室内定位，甚至在高密度建筑的城市中它的定位也不够精确；蜂窝定位系统利用GSM信号进行定位，精度相当有限；红外和超声波定位系统根据信号反射时间差进行精确定位，但是只能作用于很短的距离。随着Wi-Fi技术的成熟和Wi-Fi接入点的广泛覆盖，利用Wi-Fi无线信号进行室内外定位逐渐成为研究热点。一方面，Wi-Fi具有传输率高、支持多媒体传输、频段免费、成本低廉、无处不在的无线覆盖等优点，并且基于Wi-Fi定位不需要附加额外的硬件设备；另一方面，费城、新奥尔良、旧金山、多伦多、伦敦、台北等国际都市相继宣布和实施了全城Wi-Fi信号覆盖计划，一些发达地区的政府、校园和其它地方也在广泛部署Wi-Fi接入点，因而基于Wi-Fi的定位系统是未来无线定位领域的必然趋势。

基于Wi-Fi的无线定位系统通常是利用移动设备从多个无线接入点（Access Point，AP）接收到的信号强度（Received-Signal-Strength，RSS）来推断当前的可能位置。由于无线信号的传播符合这样的规律：移动设备接收到的信号越强，它与AP的距离就越近；接收到的信号越弱，与AP的距离就越远，移动终端接收

到的信号强度能够反应出它的位置信息。在室外环境中，Wi-Fi信号的传播和衰减的干扰因素较少，加上定位精度要求有限，已有成型的商用Wi-Fi室外定位系统[11, 39]。而由于室内环境的复杂性，加上Wi-Fi信号在室内传播的阴影遮蔽和多径效应，基于Wi-Fi的室内定位是一个难点问题，也是一个热点问题。近年来很多研究者都对这个方向进行了深入的研究[16, 27, 50]。室内Wi-Fi定位系统通常采用机器学习（Machine Learning）的方法采集有标记数据建立定位模型，这样能够达到较好的定位精度。这一方法面临的一个实际问题是大量训练数据的采集标记工作，而一旦环境或者移动设备发生变化，已有的定位模型无法自适应的调整来避免定位精度的下降，造成已有模型的失效而导致重复采集大量新的数据。这一问题严重制约了Wi-Fi定位系统的实用性和鲁棒性。

1.1 本文的贡献

本论文在详细调研国内外研究现状的基础之上，首先提出了一种统计和规则结合的无线定位算法，能够在传统方法基础之上利用规则库约束进一步提高定位精度。其次，在分析了造成信号分布差异的主要因素后，提出了一种基于迁移学习的自适应室内定位方法，将已有定位模型的数据包含的知识迁移到新的环境下以解决有标记数据不足的问题，从而大大减小了人工标定的工作。具体来说，本文的贡献主要包括以下几点：

1. **统计和规则结合的定位方法**。本文提出的统计和规则结合的定位方法，利用规则库约束对统计算法获得的结果进一步提取以提高定位精度，特别是基于规则的隐马尔科夫模型，将规则作为观察值推测用户轨迹，达到较高的精度。

2. **基于迁移学习的自适应定位算法**。本文提出了一种基于流形对齐的迁移学习方法，通过将已有定位模型的数据包含的知识迁移到新的领域，解决了新领域中训练数据不足导致定位精度下降的问题，从而大大减小了数据的人工标记成本，实现了自适应定位。

3. **基于Wi-Fi的定位跟踪原型系统**。本文实现了一个基于Wi-Fi的定位跟踪系统，既能为用户提供定位跟踪服务，也能为其它基于位置的服务提供位置信息。基于该定位原型系统本文提出了一种基于位置的多媒体服务系统作为LBS的应用之一。

4. **Wi-Fi信号数据集**。本文中所采集的Wi-Fi数据集可以作为其它迁移学习算法的实验数据集进行实验验证。

1.2　论文的组织

本文的章节安排如下：

- 第一章是引言部分，分析了定位技术的研究意义和价值，以及室内Wi-Fi定位系统遇到的问题。
- 第二章介绍了基于Wi-Fi的定位技术的研究现状，包括已有的定位方法和定位系统，并对基于学习模型的定位方法进行了详细介绍。另外，我们还介绍了迁移学习的机器学习方法，通过将其它领域的知识迁移帮助解决目标领域的任务，可以用于解决Wi-Fi信号分布差异的问题，
- 第三章详述了定位中常见的基于概率统计的方法，考虑到它们在局部区域的定位精度有限，我们提出了一种统计和规则结合的无线定位方法，在传统的统计方法之上加入规则库作为约束，能有效提高定位精度。
- 第四章论述了我们在实际系统中遇到的信号分布差异导致定位精度降低的问题，提出了一种流形对齐的降维方法进行知识迁移，利用已有数据帮助建立新领域下的定位模型，能够有效的减少所需的有标记样本数同时保证较高的定位精度。
- 第五章描述了在我们的定位跟踪原型系统基础上的一个实际LBS应用：LMSS —— 基于位置的多媒体服务系统，根据用户的位置提供定制化的多媒体视频服务。
- 第六章对整篇论文进行了总结，并给出了将来的研究方向。

第二章　研究现状概述

Wi-Fi是由Wi-Fi联盟在1999年制定的一个品牌，其实质是IEEE 802.11无线标准（也叫做无线局域网，Wireless LAN）。Wi-Fi包括了IEEE 802.11a、802.11b、802.11g和802.11n等，但是在实际中比较常用的是802.11b/g。Wi-Fi工作在2.4 GHz的免费频段上，提供最大54Mbps的传输速率，覆盖范围从几十米到几百米。基于Wi-Fi的定位通常采用收到的信号强度RSS而不是时间量度，因为一方面很难做到多AP同步，另一方面在短距离内时间差也很难准确测量。由于AP不停的发送通信信标，当移动设备持续的进行被动扫描（Passive Scanning）检测周围的AP节点时，就能根据接收的信标之间的时间间隔计算出RSS值。RSS值的大小间接反映了移动设备到AP的距离，因此可以从其中包含的位置信息来估计设备的当前位置。

本章首先介绍基于Wi-Fi的定位方法，主要集中在基于学习模型的方法上。针对定位中信号分布差异导致的定位精度下降问题，引入了迁移学习这一机器学习方法，并描述了如何用它来解决自适应定位的问题。最后介绍了一些实际可用的无线定位系统。

2.1　基于Wi-Fi的定位方法

基于Wi-Fi的定位方法可以根据不同的标准进行分类：考虑系统组织结构可以分为基于客户端的和基于无线网络架构的；考虑是否采用概率方法可以分为确定性模型和概率性模型。在本章中，我们主要根据对信号强度的使用方法不同分为两类：基于传播模型（Propagation-based Model）和基于学习模型（Learning-based Model）。

2.1.1　基于传播模型

基于传播模型的方法主要是利用信号传播特性的知识。根据观察到的无线信号在空气中传播时所呈现出的非线性和强噪声的衰减特性，人们试图从底层对其进行分析，找到造成这样一种特性的隐藏因素。典型的影响信号传播的因素包括路径损耗、阴影遮蔽效应、多径衰落特征等，这样的信号传播模型是一

种复杂的多径模型[1, 19, 37]，并且通常来说，还需要加入物理环境和网络分布的一些额外信息以建立更准确的模型。例如，在传播模型中，AP的实际物理位置通常都需要给定；当建筑物的结构或材料已知时，墙衰减因子和走道效应也要考虑加入到模型中。收集这些环境信息通常需要一定的人工标记工作。

在建立起信号传播模型后，可以采用不同的算法来进行位置估计。以常见的三角定位为例，我们首先将在某个位置设备采集的信号强度值转换为该设备到不同的AP的距离信息，然后根据已有的AP 位置信息，就能使用最小二乘（Least Square）拟合估计出该设备当前的可能位置。RADAR系统[3] 提出另一类方法，通过将定位区域离散化为方格并且根据信号传播模型估计每个方格上的信号强度，然后通过最近邻和三角定位的方法来推断设备的位置。

在室内Wi-Fi环境中，有很多类型的传播模型用来描述不同环境情况下的信号衰减特性，其中比较常见的是考虑了墙衰减因子（Wall Attenuation Factor，WAF）的模型[9]。公式(2.1)描述了这种模型的信号强度和距离的关系：

$$P(d)[dBm] = P(d_0)[dBm] - 10nlog\frac{d}{d_0} - \begin{cases} nW * WAF, & nW < C \\ C * WAF, & nW \geq C \end{cases} \tag{2.1}$$

其中n表示随距离增大路径损耗的速率，$P(d_0)$是到AP的距离为d_0处的信号强度而d_0是发射AP和接收设备的分隔距离，C是最多墙数而nW是发射AP和接收设备之间的实际墙数，WAF是墙衰减因子。根据上述关系就能根据每个位置的信号强度计算出到AP的距离。

这种模型的定位精度主要依赖于选择的传播模型是否能够很好的匹配真实室内环境，因此这种方法的关键问题在于如何建立好的室内信号传播模型。如果该模型考虑太多环境的因素例如墙、地面、人员流动，就会导致传播模型过于复杂不可解。但是采用一种简单的传播模型则会导致在复杂的实际环境中不够准确。

2.1.2 基于学习模型

基于学习模型的定位方法主要采用机器学习的方法进行定位。移动设备在一个位置能同时接收到多个AP的信号强度，将这些RSS值组合成为一个信号强度向量，则可以看作是该位置的信号模式（Pattern）在信号空间中的一种特征（Feature），而实际的位置坐标则可以看做是该模式的真实标记（Label）。尽

管由于阴影遮蔽和多径效应，Wi-Fi信号经常呈现出强噪声和非线性的变化特点，当在不同位置采集大量数据进行分析时，这种模式特点也能够被很好的捕获。将信号强度向量及其对应的位置坐标组成有标记（Labeled）数据进行训练，得到的分类器（Classifier）或者回归函数（Regression Function）就能作为一个定位模型，对于任意输入的信号可以输出其对应的位置标记，这样就完成了一次定位。

如果采集了足够的有标记数据，我们就可以使用有监督学习（Supervised Learning）的算法来建立定位模型。常见的有监督学习算法可以分为两大分支：确定性模型和概率性模型。确定性模型主要采用确定性的推断方法来估计用户位置，例如k-近邻（k-Nearest-Neighbor，k-NN）算法，决策树（Decision Tree）等。其中k-近邻算法当$k = 1$时也称为最近邻算法。[3, 2]采用了k-近邻的算法来估计用户位置，他们将实时采集的信号数据与训练数据中的所有信号强度向量进行匹配，选出最相似的来确定当前的位置。[8] 采用了决策树的方法进行定位，他们首先对整体区域进行聚类划分，然后在每一个聚类中根据AP的信息熵（Entropy）大小建立决策树。在实时定位的时候首先确定可能的聚类，然后运用该聚类中的决策树通过数次简单的比较运算就能得到估计的位置。

概率性模型主要是构造环境中每个位置的条件概率分布。[24]采用了概率推断的方法进行定位，首先基于从环境中的 9个AP采集到的信号强度计算每个位置上的信号强度概率分布，然后加入对用户运动轨迹的空间约束条件限制信号发生剧烈变化，这样能够达到较好的定位精度。[37]比较了非概率的最近邻方法和两种概率方法：朴素贝叶斯（Naïve Bayes）和直方图分布，实验结果显示基于概率的方法要好于最近邻的方法。[16] 将粒子滤波（Particle Filter）的概率方法应用到定位上，证明了该方法能得到和确定性模型一样好的效果，并能有效的融合多种传感信息。[22]将隐马尔科夫模型（Hidden Markov Model，HMM）用于人的运动轨迹建模，用状态转移矩阵表示位置间的转移关系，然后利用著名的Viterbi算法[35]推断最可能的运动轨迹序列。

采用机器学习方法的一个重要问题就是特征选择（Feature Selection）。在定位中，比较常用的是直接采用原始信号强度向量作为特征，每一维的数据即为来自某一个AP的信号强度[8, 50]。这种特征比较直观，易于通过直方图等方式观察信号特性，但是在该空间上求解分类器未必能得到最优解。因此很多研究工作提出了先进行空间映射再分类的方法，试图在降维的低维空间[31, 29]或

某些潜在空间[13, 33, 51]或核变换后的再生核希尔伯特空间（RKHS）[28]中寻找最佳的分类器。[31]提出一种基于流形学习的降维方法，同时利用移动设备和AP的有标记数据和无标记数据进行联合定位（Co-localization），该方法能够在低维流形空间回归得到设备和AP的准确位置。[29]将这一方法扩展为纯在线模型，通过数据流的输入实时建立定位模型，避免了离线训练的阶段。[13]采用高斯处理将原始高维数据映射到低维的潜在空间，通过求出潜在的变量使用贝叶斯滤波建立定位模型。[33, 51]都试图在潜在空间中寻找有用的特征建立定位模型。[28]对原始数据进行核变换之后采用典型相关分析（Canonical Correlation Analysis，CCA）求解定位问题，并测试了不同的核变换在定位上效果。另一种方法是在原始信号强度向量上进行特征提取（Feature Extraction），从而找到一些比信号强度值更加有效的特征。[21]提出采用每两个AP的强度值的比值组合成新的特征，并且通过实验证明该方法比直接用原始信号强度值更加稳定有效。

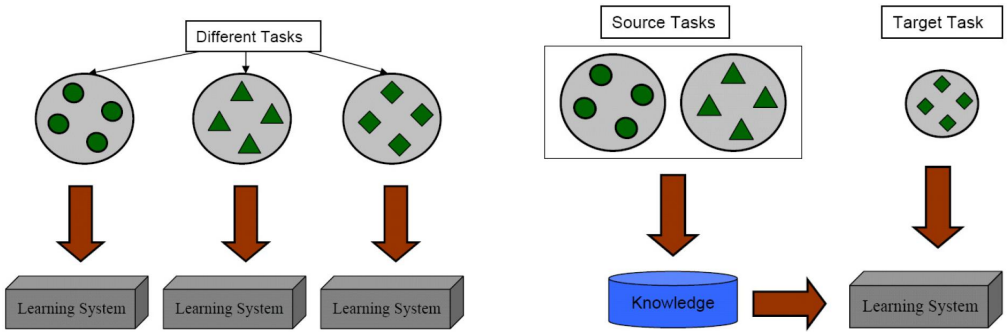
基于学习模型的定位方法是一种模式识别的方法，只需要知道每个位置对应的信号数据特征即可，因此不需要有AP位置的先验知识。但是这种方法的缺点在于需要收集大量的训练数据来学习分类器或者回归函数，因为有限的样本不足以反映出信号分布的特点和规律。这样就使得人工标记的任务变得相当繁重，需要在环境中的每一个位置标记大量数据用于训练，增加了无线定位的成本。为了能够在只有少量标记数据的情况下仍然能训练得到准确的定位模型，机器学习中的半监督学习（Semi-supervised Learning）[53]被引入到定位领域，出现了很多这方面的研究工作。[30]将流形正则化（Manifold Regularization）[5]应用到定位上，利用未标记（Unlabeled）数据在信号空间上的分布扩张有限的有标记数据张成的流形，从而找到信号空间到物理位置的准确映射关系。[29]同时利用有标记和无标记数据，采用流形学习的方法进行降维，利用拉普拉斯图的关系恢复无标记数据的位置坐标。

在Wi-Fi定位问题中，Wi-Fi信号的分布容易受到环境影响发生变化。一旦发生变化，已有的定位模型就不能很好的应用于新的环境下，而重复收集所有新数据会耗费大量的人工标记成本。一种直观的想法就是利用旧数据的知识来帮助解决新的任务，但是现有的有监督和半监督学习算法不能解决这样的问题。

2.2　迁移学习

迁移学习是近年来兴起的一种机器学习技术，旨在将知识从其它相关领域迁移用于感兴趣的目标领域的任务从而避免昂贵的数据标记工作[32]。在实际应用中，训练数据集和测试数据集的概率分布不可能完全一样，而且它们的特征表示方式也不一定相同。例如，我们要解决某个领域的分类问题，但是却只有另一个领域足够的训练数据，而且这些数据的分布不同于待解决领域的数据分布，或者它们的特征空间不一致。传统的机器学习方法如有监督学习、半监督学习等不能处理这样的问题，因此在不同的领域、不同的任务、不同的分布之间进行知识的迁移即迁移学习成为一种必要。

在迁移学习中，我们关注如何将知识从一个或多个源领域（*Source Domain*）转移到一个目标领域（*Target Domain*），而不是分别在源领域和目标领域进行学习。图2.1说明了迁移学习和传统机器学习的不同，不同形状的图案代表了不同领域的任务。传统机器学习是从每个任务各自所在的领域发掘自身知识进行学习，而迁移学习试图通过少量训练数据从以前的任务中将知识迁移用于新的目标任务。



(a) 传统机器学习　(b) 迁移学习

图 2.1: 传统机器学习和迁移学习的学习过程对比

我们首先给出迁移学习的形式化定义，然后对迁移学习进行分类介绍。

2.2.1　形式化定义

在迁移学习中，某个领域$\mathcal{D}$通常包含两个组成部分：特征空间（Feature Space）$\mathcal{X}$ 和边缘概率分布（Marginal Probability Distribution）$\mathcal{P}(X)$，其中$X =$

$\{x_1, ..., x_n\} \in \mathcal{X}$。如果我们说两个领域不同，则它们可能有不同的特征空间或者边缘概率分布。给定某个特定领域$\mathcal{D} = \{\mathcal{X}, \mathcal{P}(X)\}$，该领域中的任务也包含两个组成部分：标记空间$\mathcal{Y}$和函数$f$。函数$f$是从训练数据对$\{x_i, y_i\}, x_i \in X, y_i \in \mathcal{Y}$中学习得到的，并能够用于预测某个样本$x$对应的标记$f(x)$。从概率的角度来看，$f(x)$也可以写作$\mathcal{P}(y|x)$。

为简单起见，我们只考虑从一个源领域$\mathcal{D}_S$到一个目标领域$\mathcal{D}_T$的迁移学习。我们将源领域数据表示为$\mathcal{D}_S = \{(x_{S_1}, y_{S_1}), ..., (x_{S_{n_S}}, y_{S_{n_S}})\}$，其中$x_{S_i} \in \mathcal{X}_S$是输入，$y_{S_i} \in \mathcal{Y}_S$是对应的输出。类似的，目标领域数据可以表示为$\mathcal{D}_T = \{(x_{T_1}, y_{T_1}), ..., (x_{T_{n_T}}, y_{T_{n_T}})\}$，$x_{T_i} \in \mathcal{X}_T$是输入，$y_{T_i} \in \mathcal{Y}_T$是对应的输出。通常情况下有$n_T \ll n_S$。如果两个领域特征空间不同，则$\mathcal{X}_S$和$\mathcal{Y}_S$ 不同于$\mathcal{X}_T$和$\mathcal{Y}_T$；如果边缘概率分布不同，则$\mathcal{P}(X_S) \neq \mathcal{P}(Y_S)$。迁移学习的任务是学习函数$f(x_{T_i})$，能够预测目标领域的输入$x_{T_i}$对应的标记$y_{T_i}$。

2.2.2 迁移学习分类

在迁移学习中有三个主要的研究问题：1. 迁移什么？2. 如何迁移？3. 何时迁移？迁移什么是指什么知识能够在领域或任务间进行迁移。有些知识只适用于某些特定的领域或任务，而有些知识则是在不同领域间通用的，因而可以用于帮助提高目标任务的性能。确定什么知识可以用于迁移之后，需要有相应的算法或流程来转移这些知识，即解决如何迁移的问题。最后，何时迁移是指在什么情况下能够迁移而什么情况下不能。因为在某些情况下，源领域和目标领域可能并不相关，强制进行迁移未必能起到正面效果。目前大部分工作主要集中在前两个问题，即迁移什么和如何迁移，在给定源领域和目标领域肯定相关的前提下进行。但是如何能够避免负面迁移也是一个关键问题并且正在得到越来越多的关注。

根据不同的问题设置，现有的迁移学习工作可以大体分为三类：

归纳式迁移学习（Inductive Transfer Learning） 目标领域有一些有标记数据，源领域的有标记数据可有可无。

直推式迁移学习（Transductive Transfer Learning） 源领域有大量有标记数据，目标领域无有标记数据。

无监督迁移学习（Unsupervised Transfer Learning） 源领域和目标领域都没有有标记数据。

归纳式和直推式迁移学习可以解决分类和回归问题，而无监督迁移学习由于没有任何有标记数据，只能用于聚类和降维问题。

解决上述三类迁移学习问题的方法主要有四种：

实例迁移（Instance-transfer） 源领域的有标记数据经过调整权重系数能够用于目标领域。

特征表示迁移（Feature-representation-transfer） 寻找一种好的特征表示方法减小源领域与目标领域的差异以及分类或回归模型的误差，知识迁移蕴含在特征表示中，该特征表示能够显著提升目标任务性能。

参数迁移（Parameter-transfer） 发掘源领域和目标领域共享的参数或先验知识帮助迁移，知识迁移蕴含在共享的参数或先验知识中。

相关知识迁移（Relational-knowledge-transfer） 假定源领域和目标领域的数据相关知识类似，建立相关知识间的映射进行知识迁移。

目前迁移学习已经成功的应用到了很多实际问题中。[10, 36]提出将迁移学习用于跨领域的文本分类问题，[46]将有限的目标领域数据和大量低质量的源领域数据一起进行训练用于图像分类问题，[45]将一个人的行为数据参数迁移到另一个人身上进行行为识别。

2.2.3　迁移学习用于Wi-Fi定位

目前已有部分研究工作将迁移学习成功应用到了Wi-Fi定位问题上[49, 33, 34, 51, 52]，针对信号分布在时间、空间和设备上的差异进行知识迁移。

[49, 33, 52]都致力于解决信号分布随时间而变化的问题。[49] 提出在环境中布设一些参考点收集实时信号作为目标领域的有标记数据，然后在源领域中学习每个非参考点位置的信号强度与参考点的信号强度的线性关系，并且将这种线性关系的知识迁移到目标领域用来估计目标领域每个非参考点位置的信号强度，然后进行定位。该方法的第一步是对源领域即时间t_0的数据用下式训练回归关系：

$$s_j(t_0) = f_{ij}(r_{1j}(t_0), r_{2j}(t_0), ..., r_{mj}(t_0)) \tag{2.2}$$

其中f_{ij}表示在位置l_i时在第j个AP上，m个参考点收到的信号强度$r_{kj}(t_0)$和移动设备收到的信号强度$s_j(t_0)$之间的关系。他们采用了多回归（Multiple Regression）和模型树（Model Tree）两种方法来求得f_{ij}。然后认为该关系在动态环境中是保持不变的，这样在第二步中，对于目标领域即时间t，利用实时采集的参考点数据$r_{kj}(t)$就能估计出每个非参考点位置的信号强度

$$s_j^{est}(t) = f_{ij}(r_{1j}(t), r_{2j}(t), ..., r_{mj}(t)) \tag{2.3}$$

然后利用估计值采用最近邻的方法就能够进行位置分类预测。该方法在1.5米的范围内能达到80% 左右的精度，而在3米范围内能达到90%以上的精度。

[33]将流形正则化的方法扩展为联合正则化（Co-regularization）的形式，利用目标领域的一些有标记数据和源领域的对应项关系，同时对源领域和目标领域进行优化，从而将源领域的知识迁移到目标领域。该方法的优化解的是学出最优的映射函数对，满足：

$$\begin{aligned}(f^{(1)*}, f^{(2)*}) = \underset{f^{(1)}\in H_{K_1}, f^{(2)}\in H_{K_2}}{\operatorname{argmin}} \{ & \\ & \frac{\mu}{l_1}\sum_{i=1}^{l_1} V(x_i^{(1)}, y_i^{(1)}, f^{(1)}) + \gamma_A^{(1)}\|f^{(1)}\|_{H_{K_1}}^2 + \gamma_I^{(1)}\|f^{(1)}\|_I^2 + \\ & \frac{1}{l_2}\sum_{i=1}^{l_2} V(x_i^{(2)}, y_i^{(2)}, f^{(2)}) + \gamma_A^{(2)}\|f^{(2)}\|_{H_{K_2}}^2 + \gamma_I^{(2)}\|f^{(2)}\|_I^2 + \\ & \frac{\gamma_I}{l}\sum_{i=1}^{l}(f^{(1)}(x_i^{(1)}) - f^{(2)}(x_i^{(2)}))^2\}\end{aligned} \tag{2.4}$$

在公式(2.4)中，前三项是保证源领域中的函数$f^{(1)}$能正确分类有标记数据以及在函数空间和流形几何空间的光滑性，之后的三项是对目标领域中的函数$f^{(2)}$做同样的约束，而最后一项是用来约束$f^{(1)}$和$f^{(2)}$在对应项上产生相同的位置标记。该方法在3米范围内能达到80%左右的精度。

[52]的基本思想类似于[49]，即利用非参考点与参考点之间的线性关系在源领域和目标领域基本保持不变进行知识迁移。不同在于，[52]采用隐马模型HMM 作为定位模型，这样就能同时利用信号强度和用户轨迹的序列信息，因此相比之下可以迁移的知识更多，而且实验也证明该方法的定位精度更好。

[34]提出将源领域区域A的Wi-Fi信号数据迁移到对目标领域区域B的定位上，前提条件是两个区域共享相同的AP分布，这样就能作为两个领域之间进行迁移的桥梁。该迁移算法是[33]的联合正则化算法的一个扩展，在第一步中对源领域的数据正则化获得AP在环境中的分布知识，第二步将AP分布知识加入到目标领域的数据中作为约束进行正则化，从而求得目标领域的定位函数。

[51]考虑不同设备上的Wi-Fi信号分布不一致，将该问题看作是一个多任务学习（Multi-task Learning）[6]的问题，通过同时对多个相关领域的任务进行学习而从大量数据中受益，从而将源领域的知识迁移到目标领域。但是由于多任务学习要求相关任务的特征空间或概率分布要比较相似，而不同设备上的信号分布不满足这一条件，他们将多任务学习扩展到潜在特征空间上，这样只要在潜在空间上的特征相似就可以应用多任务学习的方法。

2.3　无线定位系统概述

最早的Wi-Fi定位系统是微软公司（Microsoft）提出的RADAR定位系统[3]。他们在环境中采集每个位置的信号强度并且进行手工标记，然后将这些数据存为一个叫做Radio map的表格。表2.1给出了一个Radio map的示例，在每个位置分四个不同的方向收集来自于不同AP（以MAC地址进行区分）的信号强度，每一行就是一组信号强度向量。在线定位的时候则通过在Radio map中查找最相似的信号强度向量进行匹配确定位置，这种最近邻算法能够在2.94米内达到50%的定位精度。在他们的后续工作[2]中，采用了一种类似Viterbi的方法加入用户轨迹约束，将定位精度提高到2.37米。Horus系统[50]分析了无线信道变化的各种原因，然后提出解决方法以提高定位精度。他们采用将位置进行聚类以减小算法的复杂性，最后能达到平均0.6米的定位精度。

随着环境规模的增大，定位精度也会出现明显的下降。Place Lab[25] 系统能够通过AP信号定位各种类型的笔记本、PDA或者移动电话，在大范围内的平均定位误差在20–30米。UCSD大学的ActiveCampus[14]项目也采用802.11无线信号在室内外环境定位Pocket PC，他们采用公式拟合出到AP的距离和信号强度之间的函数关系代替手工标定，实验证明该系统通过多AP的信号强度能够达到10米左右的精度。麻省理工学院的iFind项目[17]提供校园内的位置感知服务，他们利用Place Lab系统的定位技术以及在校园内采集的数据进行定位，精度也在十几米之内。

目前比较成熟的商用Wi-Fi定位系统主要用于室外定位[11, 39]。Ekahau公司[11] 是最早将Wi-Fi定位技术商业化的公司，他们提供了一系列物品和人员的实时定位跟踪解决方案，通过在定位目标上绑定小型的Wi-Fi标签就能实现精确定位，定位精度能达到2–3米。Skyhook公司[39]的目标是提供全球范围的Wi-Fi定位，他们通过驾驶攻击（Wardriving，也称作接入点映射）[44] 这

表 2.1: Radio map示例

位置	方向	00:B0:C6:00:0B:40 的信号强度[dBm]	00:B0:C6:00:0D:B3 的信号强度[dBm]	00:B0:C6:00:0E:8A 的信号强度[dBm]
p_1	0°	-59	-75	-71
	90°	-54	-73	-67
	180°	-49	-72	-69
	270°	-55	-73	-65
p_2	0°	-35	-64	-50
	90°	-27	-64	-43
	180°	-40	-65	-52
	270°	-30	-60	-46
p_3	0°	-69	-66	-73
	90°	-65	-60	-68
	180°	-63	-66	-70
	270°	-68	-62	-76

种Wi-Fi普查的方式将全世界的Wi-Fi接入点分布信息收集在数据库中，通过匹配查找能够提供10–20米精度的全球范围内的定位。

2.4 本章小结

本章对已有的基于Wi-Fi的定位技术进行了回顾，包括定位方法和定位系统。基于Wi-Fi 的定位方法主要有基于传播模型和基于学习模型的，而由于室内环境的复杂性，大多数室内定位都采用基于学习模型的方法。在这一类方法中，又可以分为确定性模型和概率性模型。根据训练数据的不同，可以采用有监督学习或半监督学习算法处理不同的问题，并且对定位问题中的特征选择与提取问题进行了讨论。

由于环境的影响，Wi-Fi信号分布经常会出现差异而导致已有定位模型精度急剧下降。为了解决这一问题而又能尽量减少人工标记劳动，我们引入了迁移学习的方法。迁移学习是一种将源领域的知识迁移到新的目标领域以解决目标领域有标记数据不足的问题的机器学习方法，已有的解决方法将迁移学习用于定位领域，分别解决了信号在时间、空间和设备上的分布不一致问题。

第三章　统计和规则结合的无线定位方法

在已有的定位方法中，基于概率统计的方法是一种比较准确有效的方法。通过采集大量数据，我们能够获得较为准确的信号分布情况，从而进行定位计算。但是由于Wi-Fi信号的强噪声特性，单纯依靠信号的统计分布进行定位并不能够达到最佳的定位精度。

本章主要在传统的统计方法的基础上，加入规则库作为定位约束条件，提出了一种统计和规则结合的无线定位方法，并在实际采集的Wi-Fi数据集上进行了实验验证。

3.1　定位问题的形式化定义

我们首先给出定位问题的形式化定义，然后给出若干可行的统计方法。

假设待定位的无线环境中总共布设有n个AP，为了便于数据采集，我们将整个区域离散化划分为m个方格。当用户在环境中移动的时候，随身携带的移动设备的无线网卡能够不间断的收到来自各个AP的Wi-Fi信号强度，这样的一组信号强度值可以表示为一个信号强度向量$S_i = (s_{i1}, s_{i2}, ..., s_{in}) \in \mathbb{R}^n$，其中$s_{ij}$表示从第$j$个AP接收到的信号强度值。对于那些接收不到的AP，我们将缺失值设为-100，这是环境中能够探测到的最小信号强度。对于在每一个方格内采集的数据，如果我们标记了该方格位置的坐标标签l_i，那么信号强度向量和坐标标签就组成了一个有标记数据$\{(S_i, l_i)\}$，无标记数据则只有信号强度向量$\{(S_i)\}$。坐标标签l_i根据实际情况可以是物理环境中的实际坐标(x_i, y_i)，也可以是方格的不同ID标号。本文统一将l_i定义为方格的ID标号，每个方格有唯一对应的ID标号，即$l_i \in \{1, 2, ..., m\}$。

给定n_l个有标记数据$\{(S_i, l_i)\}_{i=1}^{n_l}$和$n_u$个无标记数据$\{(S_i)\}_{i=1}^{n_u}$作为训练数据，定位问题要解决的任务是对任意输入的信号强度向量$V = (v_1, v_2, ..., v_n)$，输出当前位置的估计值l^{est}。

3.2　基于统计的定位算法

如前所述，基于统计的定位方法的基本思想是通过在环境中采集不同位置的信号强度特征数据，这些数据可以看作是每个位置独特的“指

纹（Fingerprint）”。在采集了一定量的数据后，我们就能知道每个位置的信号“指纹”特性分布。然后在定位阶段，将实时收集到的信号强度数据与数据库中的“指纹”分布进行比对就能估计出对应的可能位置了。由于基于统计的定位方法需要先获得每个位置的信号“指纹”特性分布，因此这种方法通常都分为两个阶段：离线训练阶段和在线定位阶段。在离线训练阶段，我们将采集的感兴趣区域的信号强度向量特征进行训练得到定位模型，这一阶段可以采用不同的算法得到不同的定位模型；在线定位时，将移动设备实时采集的信号强度向量输入定位模型进行定位，就能得到当前设备的估计位置。

3.2.1 单点定位算法

单点定位算法是指只靠一次采集得到的信号强度向量来估计这一时刻的位置，不考虑用户上一次定位的信号强度和位置。这种定位方式完全依赖于当前采集的瞬态信号强度向量作为输入，比较常见的单点定位算法主要包括以下这些：

最近邻 最近邻算法是最常见的模式识别方法之一，也是Wi-Fi定位领域应用最广泛的方法之一。在训练数据包含了各个位置的信号特征$\{(S_i, l_i)\}$，这些数据可以表示为表2.1中的Radio map的形式。最近邻算法计算输入的信号强度向量V和Radio map中的每个特征S_i之间的欧氏距离差即公式(3.1)得到一组距离量度$\{d(i)\}$，与V距离最小的信号特征对应的坐标标签作为当前的估计位置，即$l^{est} = \text{argmin}_{l_i}\, d(i)$。

$$d(i) = \sqrt{\sum_{j=1}^{N}(v_j - s_{ij})^2} \tag{3.1}$$

朴素贝叶斯 朴素贝叶斯是在该领域另一种广泛使用的方法。训练数据包含了不同位置l_i的概率分布$P(l_i)$，并且对于第j个AP，可以从数据集中计数得到每个位置上的信号分布概率$P(v_j|l_i)$。假设不同AP的信号分布是相互独立的，那么$P(V|l_i) = \prod_{j=1}^{n} P(v_j|l_i)$。利用贝叶斯定理即公式(3.2)就能计算输入的信号强度向量V 在每个位置的后验概率（Posterior）$P(l_i|V)$，其中似然概率（Likelihood）$P(V|l_i)$和先验概率（Prior）$P(l_i)$是从训练集中训练得到。得到的概率最大的位置被认为是当前的估计位置，即$l^{est} =$

$\text{argmax}_{l_i} P(l_i|V)$。

$$P(l_i|V) = \frac{P(V|l_i)P(l_i)}{P(V)} \propto P(V|l_i)P(l_i) \tag{3.2}$$

核方法 基于核的方法也是利用了贝叶斯定理求输入的信号强度向量V 在每个位置的后验概率$P(l_i|V)$，但是它的似然概率不是简单通过朴素贝叶斯中计数的方法来求得，而是通过一个核函数K估计信号密度分布，即

$$P(V|l_i) = \frac{1}{n_{l_i}} \sum_{l_i} K(V; S_i) \tag{3.3}$$

其中$K(\cdot; S_i)$即为核函数。高斯核就是一种常见的核函数，

$$K_{Gauss}(V; S_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{\|V - S_i\|^2}{2\sigma^2}) \tag{3.4}$$

其中的σ是高斯核的宽度参数，控制径向作用范围。这种方法的位置估计过程与朴素贝叶斯的方法类似。

直方图分布 直方图分布的方法也可以看作是另一种贝叶斯定理的应用。它将整个信号分布空间离散化为k个不重叠的组，而处于同一个组内的所有信号都用于计算该组的似然概率。组数k是可调的，控制了对空间划分的疏密程度。除了确定组数之外，我们还需要确定每个组的组距，即每个组的宽度。最简单的划分方式就是等宽组$[min + iw, min + (i+1)w], 0 \leq i < k$，其中$w = (max - min)/k$，$min$和$max$是采集到的信号中的最小值和最大值。这样每个位置上来自每个AP的信号强度直方图分布就能用k个组分布概率来描述，对于输入的信号强度向量V，其似然概率$P(V|l_i) = \prod_{j=1}^{n} P(v_j|l_i)$，而$P(v_j|l_i)$则通过$v_j$所在的组分布概率来表示。位置估计过程与朴素贝叶斯相同。

这些统计定位算法有各自的优缺点：最近邻算法训练过程简单，只需要将采集的训练数据罗列出来进行比对即可，在数据分布良好的情况下精度不错，但是对于噪声非常敏感；朴素贝叶斯依赖于大量数据构造信号分布概率，能够一定程度上降低噪声的影响，通常情况下精度能够好于最近邻算法，但是对训练数据量要求很大；核方法利用核函数的特性估计信号密度分布，有效的核函数能够很好的拟合信号分布特性，获得较好的分布效果，但是核函数的选取没

有有效的判别方法，通常都是靠经验选取；直方图分布能够很有效的降低噪声的影响，但是对组数和组距的选取很重要，太小的组距就退化为朴素贝叶斯的方法了，而太大的组距会降低信号分布的区分度。

3.2.2 轨迹跟踪算法

以上的统计算法都只针对瞬态的信号强度进行定位，如果输入的是信号强度向量序列，那么就需要将一连串的位置状态都考虑进来。HMM是一种常用来处理信号序列的方法，也可以用于解决定位问题。一个用于建模用户轨迹的HMM可以表示为一个五元组(L, O, λ, A, π)，每个元素解释如下：

- 隐藏状态L：将用户的实际位置作为隐藏状态，则L实际是位置状态空间；
- 观察值状态O：将观察到的原始Wi-Fi信号作为观察值，则O实际是信号向量空间；
- 隐藏状态转移矩阵A：隐藏状态为位置状态，则A实际是位置状态转移矩阵，a_{ij}表示从位置状态l_i转移到l_j的概率$P(l_j|l_i)$；
- 给定某个隐藏状态，对应的观察值的概率λ：即给定某个位置状态l_j，观察值o_k出现的概率为$b_{jk} = P(o_k|l_j)$；
- 隐藏状态的初始概率π：即每个位置状态的初始概率，可以简单设为等概率分布。

这样的HMM可以表示为如图3.1所示。在离线训练阶段，给定一系列标定的轨迹$T = \{(tr_i, q_i) : i = 1, ..., N\}$作为训练数据，每条用户轨迹为$tr_i = (o^1, o^2, ..., o^{|t|})$，对应的位置序列为$q_i = (l^1, l^2, ..., l^{|t|})$，其中$o^j \in O$，$l^j \in L$。HMM的三个主要参数$\theta = (\lambda, A, \pi)$可以用EM（Expectation-maximization）算法估计得到。将训练得到的HMM用于在线定位时，某个观察到的信号序列$tr = (o^1, o^2, ..., o^{|tr|})$对应的最可能的隐藏状态序列即位置序列$q$可以用Viterbi算法[35]求出。

3.3 基于规则的定位算法

不同于统计定位算法着重采集每个位置的信号“指纹”分布的统计特性，基于规则的方法主要是针对每个位置的信号特点建立起判别规则，然后利用判别规则对输入的信号强度向量进行判断，从而找到最符合判别规则的位置。

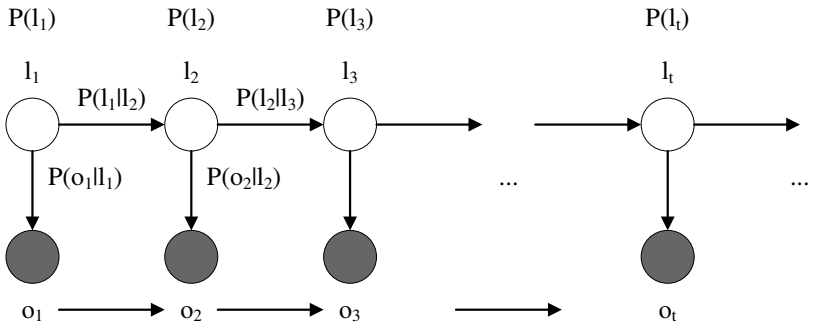


图 3.1: HMM用于用户轨迹建模

3.3.1　决策树

最常见的基于规则的方法就是决策树，它广泛应用于模式分类的问题中。决策树的基本思想是很自然和直观的：通过对输入的测试样本进行一系列问题的判断对其进行分类。举例来说，我们给出针对8个方格（$G_1 \sim G_8$）建立的一颗决策树如图3.2所示。每个内部节点都对应了针对某个AP信号强度的问题，从内部节点上又延伸出若干子树，每个分支对应了信号强度的不同分布区间。从根节点出发，测试样本通过若干次问题判断可以到达某个叶节点，该叶节点给出了该测试样本最可能的位置。例如通过图3.2中的决策树对测试样本$\widetilde{S} = (-91, -84, -65, -78)$进行判断，首先在根节点处判断$AP_2$的值，由于$AP_2$的值为$-84$，因此选择中间的分支进而判断$AP_4$的值，依次判断下去直到到达叶节点。该测试样本最后达到G_3节点，因此它的估计位置为G_3。

决策树是一种适合于在小范围内进行判断的方法，当要分类的位置逐渐增多时，决策树的高度也会随着增加，从而导致树的建立过程越来越复杂。所以通常决策树都会和聚类方法结合使用。首先通过大范围的聚类得到低精度的位置判断，然后在每个聚类之内再建立决策树进行高精度的分类，这样能够提高决策树的实用性。

3.3.2　相对信号强度关系

以上的方法都是基于具体的信号强度值来进行定位，而在实际过程中由于信号强度的绝对值不稳定，我们提出一种利用信号强度间的相对关系来建立判别规则的方法RSSR（Relative Signal Strength Relation）。该方法的基本思想很直观，即在某个位置，如果从AP_i接收到的信号强度大于从AP_j接收到的信号强

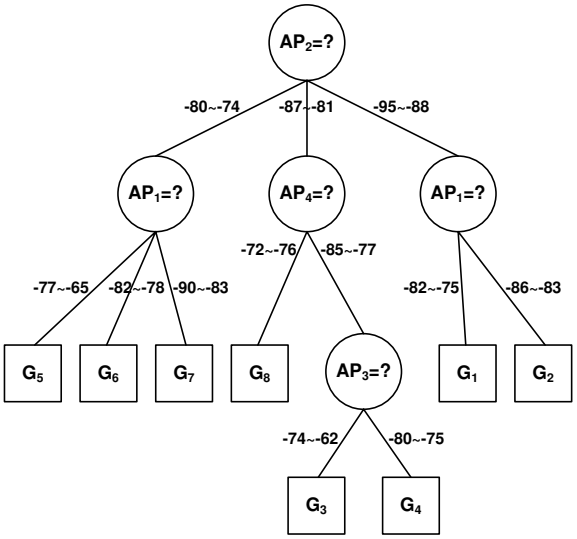


图 3.2: 用于定位的决策树示例

度，记为$RSS_i > RSS_j$，即使信号强度由于噪声发生跳变，这一关系在环境中基本是保持不变的。这样我们可以将这一关系提取出来作为对该位置进行判别的一项规则。

我们采用启发式（Heuristic）的方法来创建规则，最简单的规则方式就是比较AP_i和AP_j的信号强度大小。如果一个位置到AP_i的距离小于到AP_j的，那么就有$RSS_i > RSS_j$。这样可以创建三种类型的规则$rule_1 = (RSS_i > RSS_j)$、$rule_2 = (RSS_i < RSS_j)$和$rule_3 = (RSS_i = RSS_j)$。另外，尽管我们考虑的是相对信号强度，如果某些位置的信号强度特别稳定，我们同样可以利用绝对信号强度值建立规则，如$rule_4 = (-72 \leq RSS_i \leq -70)$。

举例来说，如果两个AP在实际环境中的分布情况如图3.3所示，对方格G_1和G_2我们可以建立规则$rule_1 = (RSS_1 > RSS_2)$，对方格G_4和G_5建立规则$rule_2 = (RSS_1 < RSS_2)$，而对G_3可以建立规则$rule_3 = (RSS_1 = RSS_2)$。显然，仅仅靠单一的规则是没法区分某些位置的，例如G_1和G_2、G_4和G_5，因此要靠其它AP的信号强弱关系建立尽可能多的规则来提高位置的区分度。

在我们的定位问题中，我们默认不知道AP的位置，这样就不能直接通过环境分布情况来建立启发式规则。我们沿袭决策树的建立方法，利用训练数据建立规则库，然后利用规则库来定位。由于环境中的AP总数为n，因此所有可能的AP对的数目为$C_n^2 = \frac{n(n-1)}{2}$。对某个位置上任意一对(RSS_i, RSS_j)，我们可

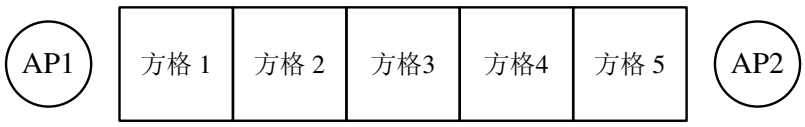


图 3.3: 环境平面布置图样例

以建立$rule1$、$rule2$和$rule3$三种规则中的某一种。这样每个位置上的规则数目为C_n^2，所有可能的规则组合数目为$3^{C_n^2}$。对于输入的测试样本，我们将其转换为规则表示，匹配到的最符合的规则对应的位置即可认为是估计得到的位置。

3.4　统计和规则结合的定位方法

基于统计的方法和基于规则的方法在定位上有各自的优缺点：前者对于整体的信号分布特性进行统计，能够在大范围内很快的确定一个准确的小区域，但是由于小区域内的邻近点的信号分布类似，它在小区域内的定位效果有限；后者根据小区域内每个位置的自身信号特点建立规则，易于利用这些规则进行准确划分，但是如果在大范围内进行规则判断，整个定位过程将相当繁琐。因此，我们提出了将统计和规则结合的定位方法，利用统计方法确定大致定位范围，规则库作为约束项对结果进行筛选，从而找到合适的位置。

针对单点定位算法，我们首先采用统计定位算法获取若干候选位置，然后采用规则库对其进行进一步判断。例如，采用最近邻算法我们可以找到与输入的测试样本距离最近的k个近邻，然后提取这k个近邻对应的规则项对测试样本进行验证，选出最符合规则匹配的位置。或者通过贝叶斯的方法找到测试样本的后验概率最大的k个位置，然后通过规则匹配选择最符合的。

轨迹跟踪的算法同样可以先找出k个备选的轨迹，然后利用人的行走轨迹的平滑性，在这些轨迹中选择轨迹衔接最为平滑和连续的作为估计结果，这种方法在RADAR系统的改进版[2] 中就提出过，他们采用类Viterbi算法用于用户连续跟踪和消除歧义点。我们这里提出一种基于规则的HMM定位方法RBHMM（Rule-based HMM），将规则库和HMM相结合进行定位。

RBHMM同样可以用一个五元组(L, O, λ, A, π)来描述，不同于前面介绍的HMM， RBHMM用信号间强弱规则替代原始信号向量作为观察值，这样RBHMM的五个元素可以描述如下：

- 隐藏状态L：将用户的实际位置作为隐藏状态，则L实际是位置状态空间；

- 观察值状态O：将观察到的信号强弱关系（rule1,2,3）作为观察值，则O实际是关系类别空间；

- 隐藏状态转移矩阵A：隐藏状态为位置状态，则A实际是位置状态转移矩阵，a_{ij}表示从位置状态l_i转移到l_j的概率$P(l_j|l_i)$；

- 给定某个隐藏状态，对应的观察值的概率λ：即给定某个位置状态l_j，观察到关系组$o_k = (rule_{k1}, rule_{k2}, ...)$出现的概率为$b_{jk} = P(o_k|l_j)$；

- 隐藏状态的初始概率π：即每个位置状态的初始概率，可以简单设为等概率分布。

RBHMM的离线训练过程与HMM一样，给定一系列标定的轨迹$T = \{(tr_i, q_i) : i = 1, ..., N\}$作为训练数据，采用EM算法估计出三个主要参数$\theta = (\lambda, A, \pi)$。不同之处在于RBHMM的观察值不是信号强度向量，而是从相对信号强度关系中提取出的规则。在线定位时，也要将观察到的信号序列转换为关系组序列$tr = (o^1, o^2, ..., o^{|tr|})$，通过Viterbi算法求对应的位置序列$q$。这种方法也可以看作是对原始信号数据进行特征提取之后进行定位，从而将统计方法和规则约束结合起来。

3.5 实验验证

为了测试不同算法的实验效果，我们在实际室内楼层中搭建了无线网络环境以及定位系统，并且采集了Wi-Fi数据来验证我们的算法性能。

3.5.1 数据采集

我们在金白领科研楼三层布设无线网络环境。实验环境覆盖的面积大概为30米×15米，包括五个房间和一条走廊。我们在环境中布设了7个TENDA AP来组建无线网络架构，如图3.4中红色三角形所示。整个环境被划分为161个方格用来进行信号采集，每个方格的大小为1米×1米。

数据采集过程中，用户手持移动设备在每个方格中心停留一段时间采集当前位置的信号并进行标记，移动设备上安装有我们开发的数据采集软件。我们使用内置Intel® Pro/Wireless 3945ABG无线网卡的IBM® R60笔记本电脑在

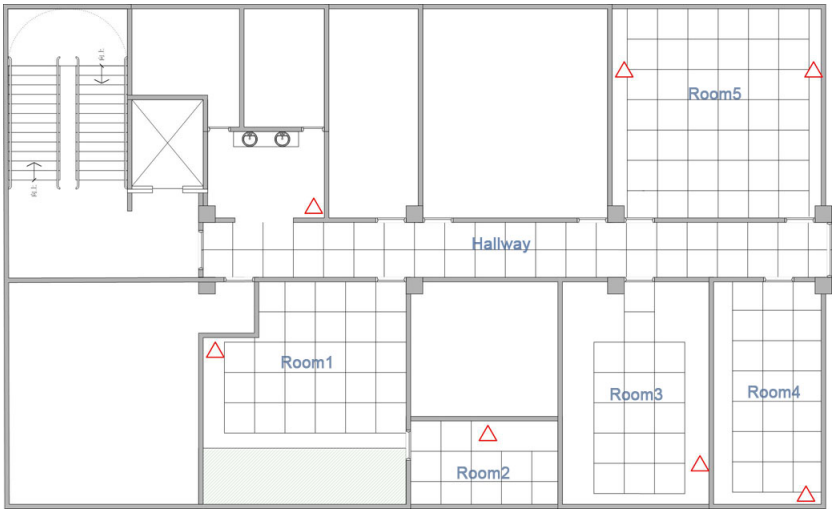


图 3.4: 无线定位实验环境的布局图

每个方格以10Hz的频率采集200个样本，总共得到200 × 161个样本，我们就用这些数据测试不同的单点定位算法的性能。

为了比较轨迹跟踪算法的性能，我们采用了香港科技大学采集的Wi-Fi数据集[47]，里面包含40条训练轨迹数据和43条测试轨迹数据。我们用这些数据来测试RBHMM的性能。

3.5.2　性能分析

为了减少噪声的影响，我们将每10个样本求平均值，这样总共剩下20 × 161个样本。我们对每个方格的数据采用“留一法（Leave-one-out）”划分训练样本和测试样本，即每次在每个方格的数据中取1个出来作为测试样本，剩下的19个作为训练样本，将19 × 161个数据一起进行训练，然后在剩余的1 × 161个测试样本上进行测试，重复这个过程20次得到平均定位精度和标准差。

我们首先测试了传统的统计算法的性能表现，包括最近邻算法、朴素贝叶斯、高斯核方法以及直方图分布四种方法。其中直方图分布我们测试了不同的组数和组距，在组距为3时性能表现最好，这里采用了这个结果。所有实验结果如表3.1所示，这里的误差距离定义为估计的位置与真实位置之间的距离，在误差距离之内的估计结果都认为是正确的。从表中看出，高斯核方法的精度是最好的，直方图分布在组数和组距选择适当的时候结果也不错，最近邻算法的精度一般，但是相对比较稳定，而朴素贝叶斯由于训练数据量不够大，效果是最差

的。

表 3.1: 各种统计方法在不同误差距离下的精度（%），括号内为标准差

误差距离（米）	0.5	1	2	3	4
最近邻算法	72.1(9.31)	76.3(8.37)	83.4(5.74)	90.0(2.65)	93.9(1.41)
朴素贝叶斯	68.6(9.83)	69.0(9.68)	69.7(9.46)	71.1(9.16)	73.1(8.60)
高斯核方法	84.7(10.1)	87.1(8.56)	90.6(5.40)	94.6(2.65)	97.3(1.31)
直方图分布	80.8(7.73)	83.4(6.65)	87.0(5.75)	91.3(4.22)	93.6(3.08)

我们利用训练数据建立规则库之后与统计方法一起进行测试，从统计方法中产生 5个候选结果后再利用规则库判断，得到的实验结果如表3.2所示。很明显加入规则库的约束后，所有算法的精度有一定程度的提升，其中朴素贝叶斯提升的程度最大，超过了10%，最近邻算法也提升了大约8%，而另两种统计方法本身的定位精度已经较高，相对提升的幅度较小。

表 3.2: 带约束的统计方法在不同误差距离下的精度（%），括号内为标准差

误差距离（米）	0.5	1	2	3	4
最近邻算法	80.3(9.20)	84.2(7.92)	89.1(3.42)	94.4(1.65)	97.0(1.41)
朴素贝叶斯	78.3(9.62)	83.1(8.18)	86.9(5.75)	90.2(3.22)	94.0(1.78)
高斯核方法	88.2(9.11)	90.7(7.32)	94.2(2.56)	97.1(1.32)	98.9(1.00)
直方图分布	84.8(6.83)	88.4(6.12)	91.5(4.75)	94.1(3.26)	95.6(3.08)

在轨迹算法测试中，我们用40条训练轨迹分别建立RBHMM和普通的HMM，然后在测试轨迹上比较了两者的性能，实验结果如图3.5所示。RBHMM的定位精度要高于普通的HMM，并且标准差较小，相比之下更加稳定。

3.6 本章小结

本章首先对定位问题进行形式化定义，并详述了在定位中广泛采用的统计方法，包括最近邻、朴素贝叶斯、核方法和直方图分布等单点定位算法，以及用于轨迹跟踪的HMM算法。由于统计方法在一些局部区域的判断不够精确，也有人采用基于规则的定位方法，例如决策树等。我们提出了一种基于相对信号强度关系的规则比较法，利用AP间的相对信号强弱关系建立判别规则，也能用于

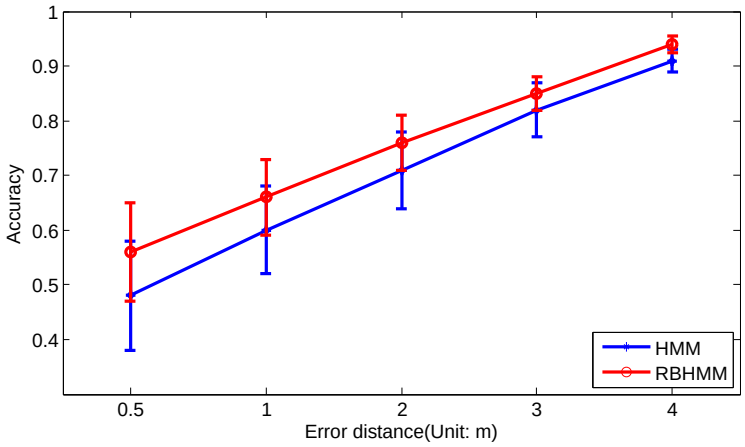


图 3.5: 基于规则的HMM和传统HMM的性能对比

分类。在这之上，我们提出了统计和规则结合的无线定位方法，在传统的统计方法基础上加入规则库作为约束项，能够进一步提高统计方法的定位精度。另外，我们提出的基于规则的HMM将统计方法和规则相结合，能够改进HMM在轨迹跟踪问题的效果。

第四章　基于迁移学习的自适应无线定位方法

前面所讨论的定位方法都有一个默认的前提条件 —— Wi-Fi信号的分布在环境中是固定不变的，即建立的静态定位模型总是有效的。然而真实环境不可能是一成不变的静态系统，在动态环境中，Wi-Fi信号的分布差异势必导致定位模型的失效或性能下降。因此，如何设计动态定位模型以自适应的应用于不断变化的动态环境，是Wi-Fi定位系统面临的一个实际问题，也是一个重要问题。已有的研究工作虽然解决了一些定位中的迁移学习问题，但是这些方法都只考虑一种因素对于信号分布的影响，而在实际系统中，信号分布可能会同时受到多种因素的影响，比如采用不同于训练数据集的设备在不同的时间段进行定位，这时的信号分布会同时受到时间和设备两方面的影响，而已有的工作都没有考虑这样的情况。

本章针对信号分布同时受到时间和设备影响的问题，提出一种流形对齐的降维方法，通过在低维流形空间中进行知识迁移，实现自适应定位的目标，解决了目标领域有标记数据不足的问题。

4.1　定位中的迁移学习问题

4.1.1　问题描述

已有的Wi-Fi定位系统通常都有一个默认的前提条件，即Wi-Fi信号的分布在环境中是基本不变的，即使考虑了噪声的因素。这样才能保证训练得到的定位模型在在线阶段是适用的。事实上，由于Wi-Fi信号的特性，其分布很容易受到环境变化的影响而改变。具体来说，温度、湿度、室内人员的活动、大型物件的移动都会导致Wi-Fi信号的分布出现差异，使得已有的定位模型失效而降低定位精度。我们将这个因素归结为时间因素，即信号在不同的时间段会呈现不同的分布。图4.1显示了同一个设备在同一地点的不同时间段采集的信号的均值分布，很明显在AP1和AP5上信号强度有着很大的差异。

另一方面，为减少人工标记成本，定位模型通常都是基于某一种移动设备（例如笔记本电脑）采集的数据而建立。但在实际定位系统中，等待定位的目标移动设备可能是多种多样的（例如PDA或者智能手机等）。这些不同的设

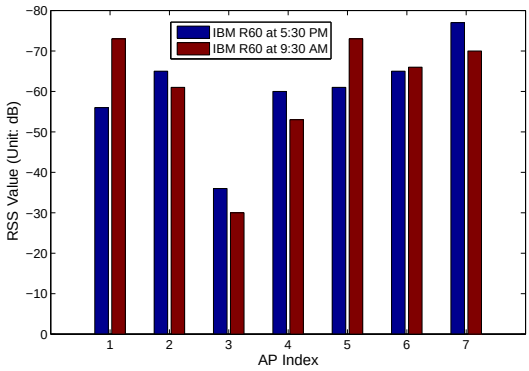


图 4.1: IBM R60笔记本在同一地点不同时间采集的数据

备所配备的无线网卡有着迥异的信号敏感度，即使是同一型号的设备，可能也会由于硬件设备上的差异使得信号敏感度不同。因此不同设备上的Wi-Fi信号分布也存在差异，我们将这个因素归结为设备因素。图4.2显示了两个不同的设备（笔记本和手机）在同一时间同一地点采集的信号分布，可以看到不同设备的信号也有着迥异的分布。

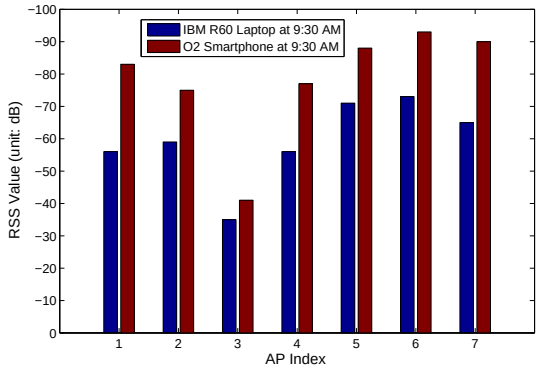


图 4.2: IBM R60笔记本和O2智能手机在同一地点同一时间采集的数据

当采用A时间（或A设备）的数据训练得到的定位模型用于B时间（或B设备）的定位时，由于信号分布迥异，该定位模型会严重失效而导致精度急剧下降。我们在本章后面的实验部分测试了将A时间（或A设备）的数据训练得到的定位模型用于B时间（或B设备）的效果，证明了信号分布差异对于定位模型的影响相当严重，因此有必要保持定位模型的实时更新以提供准确的定位服务。

一种可能的方法是重新采集所有的数据用于训练新的模型，这样可以保证模型的准确有效。但是在大规模的环境中采集所有的数据会耗费极大的人工标记工作量，增加人工成本。而且，一旦这些新采集的数据也失效，势必又要再一次重新采集所有数据。因此这种方法不是一种有效的解决方法。

考虑到已有的旧数据包含了很多信号与信号以及信号与位置之间的关系，如果能利用这些知识帮助建立新的定位模型，就可能大大减少所需的标记数据，从而减少人工标记的工作，实现有效的自适应定位。

4.1.2　形式化定义

延续第三章中的形式化定义，即无线环境中布设有n个AP，整个环境面积被划分为m个方格。采集到的有标记数据为$\{(S_i, l_i)\}$，无标记数据则只有信号强度向量$\{(S_i)\}$。

我们考虑在两个不同的时间段T_a和T_b采集的数据集的差异问题。在T_a时刻，收集了足够的有标记数据$\{(S_i^{(a)}, l_i)\}_{i=1}^{n_{la}}$，而在$T_b$ 时刻，只有少量有标记数据$\{(S_i^{(b)}, l_i)\}_{i=1}^{n_{lb}}, n_{lb} \ll n_{la}$，并有一些很容易得到的无标记数据$\{(S_i^{(b)}\}_{i=1}^{n_{ub}}$供选择。这两个时刻的数据有相同的特征空间，但是边缘概率分布不同，即$\mathcal{P}(S^{(a)}) \neq \mathcal{P}(S^{(b)})$。我们的目标是能够在$T_b$时刻进行准确的定位，但是该时刻没有足够的训练数据，因此有必要利用T_a时刻获得的有标记数据的知识。

类似的，两种不同的设备V_a和V_b采集的数据集也存在这样的情况。V_a收集了足够的有标记数据$\{(S_i^{(a)}, l_i)\}_{i=1}^{n_{la}}$，但是目标设备$V_b$只有有限的有标记数据$\{(S_i^{(b)}, l_i)\}_{i=1}^{n_{lb}}, n_{lb} \ll n_{la}$ 和一些无标记的数据$\{(S_i^{(b)})\}_{i=1}^{n_{ub}}$。同样，两种设备的数据特征空间相同，但是边缘概率分布不同$\mathcal{P}(S^{(a)}) \neq \mathcal{P}(S^{(b)})$。因此对$V_b$的定位也需要$V_a$收集的数据的帮助。

进一步考虑不同设备在不同时间的问题，也是类似的定义方式。因此这些问题能够统一为一个迁移学习的问题，即给定某个源领域$\mathcal{D}_S$ 足够的有标记数据$\mathcal{D}_S^l$，我们要对边缘概率分布不同于$\mathcal{D}_S$的目标领域$\mathcal{D}_T$进行分类预测，而该领域只有有限的有标记数据$\mathcal{D}_T^l$和一些可选的无标记数据$\mathcal{D}_T^u$。

4.2　信号特性分析

如前所述，我们在定位问题上应用迁移学习，就需要先解决迁移学习的三个主要问题：迁移什么、如何迁移、何时迁移。在本节中我们通过对信号特性的

分析首先回答迁移什么的问题。

4.2.1 流形假设

在无线定位问题上，主要的对象是无线信号，因此要迁移的知识必然与信号特性相关。首先我们看看信号在空间中传播时的衰减特性，图4.3 给出了移动设备接收到某AP的信号强度与到该AP的距离的变化关系。很明显信号的衰减不是简单的线性变化，而是高度的非线性关系，并且带有强噪声。即使在某一个固定的位置，信号也不会保持一个稳定的值。

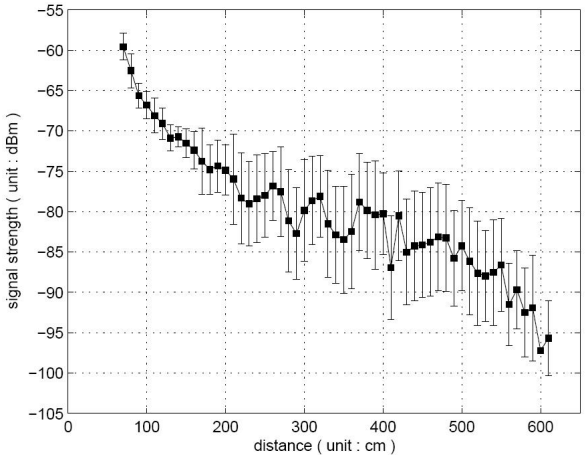


图 4.3: 信号的衰减特性

尽管如此，Wi-Fi信号还是存在一些可以观察到的特性：

1. 离AP越近，收到的信号强度越大；离AP越远，收到的信号强度越小。这个特性从图4.3中也能得到，虽然信号衰减并非线性，但是基本趋势是随着距离增大信号强度减小。

2. 邻近的位置比相距较远的位置收到的信号强度向量更相似。这一点可以从第1点推出，邻近位置的信号衰减情况类似，因而收到的信号强度值也会相似一些。

这种数据分布关系满足了流形学习的假设，即两个数据点如果在内在几何空间中的边缘概率相似，则他们的条件概率也相似。这就等于如果两个信号强度向量在某种流形结构上比较接近的话，它们对应的物理位置也是邻近的。而

这一假设与信号的实际观察特性相吻合。因此，将信号空间看作是某种流形结构进行学习是合理的。事实上，[29]实现了三维信号空间的可视化，如图4.4所示。1、2、3是布设的三个AP节点，整个信号采样区域是以1、2、3为顶点的三角形，其中A、B、C是物理空间上两两相邻的三个位置，而它们对应的信号强度向量在信号空间中的相邻关系仍保持不变。

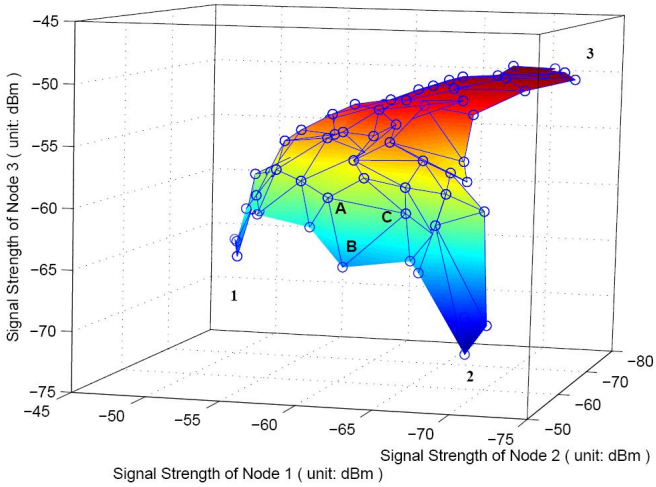


图 4.4: 信号的流形分布

4.2.2　迁移可行性

采用迁移学习解决该问题的可行性同样是基于对Wi-Fi信号数据的观察。尽管不同时间或不同设备上信号的不同分布形成了不同的信号空间，但是它们都是在同一个室内环境中采集的，即基于共同的物理空间，这样可以认为两个信号空间的流形结构能够通过共同的物理空间对应起来，从而使得两个信号空间上的迁移变得可行。图4.5显示了这种对应关系。两个不同的信号空间分别对应于源领域和目标领域，其中相同颜色的点可以通过对应的坐标标签即它们共同的物理位置对应起来，如图中虚线所示。由于目标领域的有标记数据有限，我们的目标是将源领域中信号与位置的对应知识迁移到目标领域获得目标领域的信号与位置的对应关系。这样我们要解决的问题就是，给定一部分这样的对应关系，例如图中有颜色的点，如何学出剩下那些无颜色的点的对应关系？

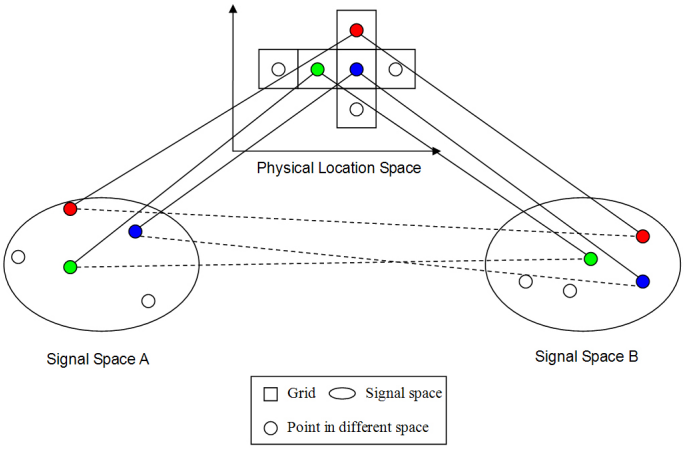


图 4.5: 不同信号空间和实际物理空间的对应关系

4.3 基于流形对齐的自适应定位算法

流形学习的方法和一般的降维分析一样，旨在把一组高维空间中的数据在低维空间中重新表示，不同于线性降维的方法如主成分分析（Principal Components Analysis，PCA）[18]，流形学习属于非线性降维的分支，代表方法主要包括局部线性嵌入（Locally Linear Embedding，LLE）[38]、等度规映射（Isometric Feature Mapping，Isomap）[43]、拉普拉斯特征映射（Laplacian Eigenmaps，LE）[4]等。但是这些方法都只是针对一个数据集进行降维，而不考虑不同数据集之间的关系。事实上，不同的数据集可能在某些低维空间上共享一些潜在有用或者有实际意义的特征。因此，我们引入了流形对齐（Manifold Alignment）[15] 这一降维方法，它是一种带约束的流形学习方法。具体来说，通过不同的数据集之间给定的若干对应关系，流形对齐能够将不同数据集的低维嵌入（即该数据的流形）对齐，从而找到其它数据间的对应关系，实现不同数据集上知识迁移的目的。

4.3.1 流形构造

我们首先介绍LE的流形学习方法是如何对高维数据集构造流形的。

假设$X = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^{d_h}$是在高维空间的一个数据集，我们可以通过构造一个无向有权图来描述该数据集中的局部邻接关系。将X中的每一个向量$\mathbf{x}_i, i \in \{1, 2, ..., n\}$看作是高维空间中的一个节点，每两个节点之间存在一条边e。如果节点i和j是邻居关系（邻居关系可以通过k个最近邻或ϵ-邻域定义），记

为$i \sim j$，边e_{ij}的权重$W_{ij} \neq 0$；否则$W_{ij} = 0$。得到的权重矩阵$W_{n\times n}$通常是实对称矩阵，并且没有非负项，即$W_{ij} = W_{ji} \geq 0$。权重的构造方法可以有不同的选择，我们在实验中会测试不同的权重方法。

基于权重矩阵W，定义对角阵D，其中$D_{ii} = \sum_j W_{ji}$。这样我们可以得到图的拉普拉斯矩阵$L = D - W$，其中：

$$L_{ij} := \begin{cases} \sum_{j\sim i} W_{ij} & \text{if } i = j \\ -W_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

如果L是一个连通图，那么会存在一个唯一的0特征值并且有相对应的特征向量$\mathbf{e} = [1, 1, ..., 1]^T$。

X的低维（d_l）流形嵌入能够从拉普拉斯矩阵L中这样获得：定义一个图顶点上的实值映射函数$\boldsymbol{f} : d_h \to \mathbb{R}$，$X$的最优化的流形嵌入$\boldsymbol{f}^*$等价于公式(4.2)的优化问题：

$$\begin{aligned} \underset{\boldsymbol{f}}{\text{argmin}} \quad & \boldsymbol{f}^T L \boldsymbol{f} = \tfrac{1}{2}\sum_{i,j}(f_i - f_j)^2 W_{ij} \\ \text{s.t.} \quad & \boldsymbol{f}^T \boldsymbol{f} = 1, \boldsymbol{f}^T \mathbf{e} = 0 \end{aligned} \tag{4.2}$$

其中的约束条件是对$\boldsymbol{f}$的缩放和平移进行限制。该优化问题的最优解$\boldsymbol{f}^*$是拉普拉斯矩阵L的最小的d_l 个非零特征值对应的特征向量，这些特征向量组成了一个与原始高维数据结构类似的低维流形嵌入。图4.6是两个合成数据 —— S型曲面和波浪曲面的三维可视图，图4.7 是它们各自单独降维得到的二维流形嵌入，由于原始数据的结构差异，降维后的两个流形差异很大。

图 4.6: S型曲面和波浪曲面的原始高维分布

4.3.2　流形对齐

现在考虑同时构造两个相关的数据集的低维流形。假设X, Y是高维空间的两个数据集，它们分别有子集X_p, Y_p存在着一一对应关系（例如两个数据点有相

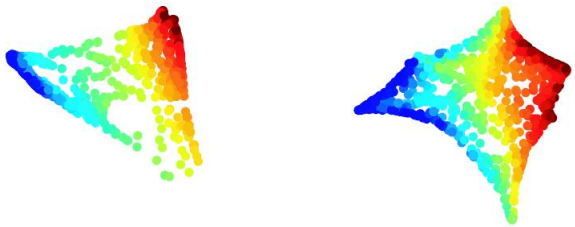


图 4.7: 单独降维得到的二维流形嵌入

同的标记），其它剩余的数据表示为X_s, Y_s，流形对齐的目标是将X_s和Y_s一一对应起来。

与单个数据集的流形构造类似，令$\boldsymbol{f}$和$\boldsymbol{g}$是分别定义在X和Y的图上的实值映射函数$\boldsymbol{f}, \boldsymbol{g} : d_h \to \mathbb{R}$，$X, Y$之间的对应关系可以表示为$x_i \leftrightarrow y_i, i \in p$。根据公式(4.2)，$\boldsymbol{f}$和$\boldsymbol{g}$分别表示了每个数据集中提取出的低维流形的坐标。由于两个数据集的对应点是一一对应的，那么这些点在低维空间$\boldsymbol{f}$和$\boldsymbol{g}$上也应该是对应的，即它们的坐标应该很相近。对单一流形构造的算法进行扩展，这种包含对应关系的最优化流形可以表示为公式(4.3)的优化问题：

$$C(\boldsymbol{f}, \boldsymbol{g}) = \mu \sum_{i \in p} |f_i - g_i|^2 + \lambda_1 \boldsymbol{f}^T L^x \boldsymbol{f} + \lambda_2 \boldsymbol{g}^T L^y \boldsymbol{g} \tag{4.3}$$

其中L^x和L^y分别是X和Y的拉普拉斯矩阵，μ、λ_1和λ_2是每一项的权重系数。该式子的第一项衡量$\boldsymbol{f}$和$\boldsymbol{g}$ 在对应点上的差异，后两项保证流形在低维空间上的光滑性。

但是公式(4.3)的优化是没有定义清楚的，因为同时对$\boldsymbol{f}$和$\boldsymbol{g}$ 的变换是不等的。因此我们定义$\boldsymbol{h} = [\boldsymbol{f}^T, \boldsymbol{g}^T]^T$，这样最小化公式(4.3)等价于最小化Rayleigh商：

$$\underset{\boldsymbol{h}}{\text{argmin}} \quad \widetilde{C}(\boldsymbol{h}) = \frac{\boldsymbol{h}^T L^z \boldsymbol{h}}{\boldsymbol{h}^T \boldsymbol{h}}, \quad \text{s.t.} \quad \boldsymbol{h}^T \mathbf{e} = 0 \tag{4.4}$$

其中L^z定义为

$$L^z = \begin{bmatrix} \lambda_1 L^x + U^x & -U^{xy} \\ -U^{yx} & \lambda_2 L^y + U^y \end{bmatrix} \tag{4.5}$$

U^x、U^y、U^{xy}和U^{yx}是只在对角线上有非零元素的矩阵

$$U_{ij} = \begin{cases} \mu, & i = j \in p \\ 0, & \text{otherwise} \end{cases}$$

求解L^z的最小的d_l个非零特征值对应的特征向量就可以得到一个d_l维的流形嵌入$E^z{}_{(x+y)\times d_l} = [E^{xT}_{x\times d_l}, E^{yT}_{y\times d_l}]^T$，其中$\boldsymbol{f}^* = E^x$，$\boldsymbol{g}^* = E^y$。

进一步，由于公式(4.3)中的μ是用来调整对应项和光滑项的权重系数，当$\mu \to \infty$时，该公式等价于强制约束对于任意$i \in p$，有$f_i = g_i$。这样该优化问题可以变成另一个特征值求解问题：

$$\underset{\boldsymbol{h}}{\operatorname{argmin}} \quad \widetilde{C}(\boldsymbol{h}) = \frac{\boldsymbol{h}^T L^z \boldsymbol{h}}{\boldsymbol{h}^T \boldsymbol{h}}, \quad \text{s.t.} \quad \boldsymbol{h}^T \mathbf{e} = 0 \tag{4.6}$$

其中的$\boldsymbol{h}$和L^z定义为

$$\boldsymbol{h} = \begin{bmatrix} \boldsymbol{f}_p = \boldsymbol{g}_p \\ \boldsymbol{f}_s \\ \boldsymbol{g}_s \end{bmatrix} \tag{4.7}$$

$$L^z = \begin{bmatrix} \lambda_1 L^x_{pp} + \lambda_2 L^y_{pp} & \lambda_1 L^x_{ps} & \lambda_2 L^y_{ps} \\ \lambda_1 L^x_{sp} & \lambda_1 L^x_{ss} & 0 \\ \lambda_2 L^y_{sp} & 0 & \lambda_2 L^y_{ss} \end{bmatrix} \tag{4.8}$$

该优化问题同样可以通过寻找L^z的最小的d_l个非零特征值来解决，而不需要设定参数μ。对应的d_l个特征向量构成了一个d_l维的联合低维流形$E^z{}_{(p+fs+gs)\times d_l} = [E^T_{p\times d_l}, E^T_{fs\times d_l}, E^T_{gs\times d_l}]^T$，这样$\boldsymbol{f}^* = [E^T_{p\times d_l}, E^T_{fs\times d_l}]^T$，$\boldsymbol{g}^* = [E^T_{p\times d_l}, E^T_{gs\times d_l}]^T$。

公式(4.5)和(4.8)都是在构造一个新的拉普拉斯矩阵L^z，这种方法可以看作是将两个不同的图连接成一个新的图。其中根据公式(4.5)的定义是在两个图的对应点之间加上权重为μ的边，而公式(4.8)直接将对应点重合成为一个点。不论采用哪种方法，连接成的图的低维流形都能将两个组成图自动对齐。图4.8给出了S型曲面和波浪曲面的原始流形和对齐后的流形对比。

基于对齐后的流形，一个数据集上的点可以直接在另一个数据集上寻找最近邻建立对应关系，从而避免在高维空间推导变换关系。

4.3.3　自适应定位算法

在定位问题上，给定某个源领域$\mathcal{D}_S$足够的有标记数据$\mathcal{D}_S^l$，我们要对目标领域$\mathcal{D}_T$进行预测或分类，而该领域只有有限的有标记数据$\mathcal{D}_T^l$ 和一些可选的无标记数据$\mathcal{D}_T^u$。我们提出基于流形对齐的定位算法LuMA（Localization using Manifold Alignment）来解决由于时间和设备造成的信号分布差异问题。

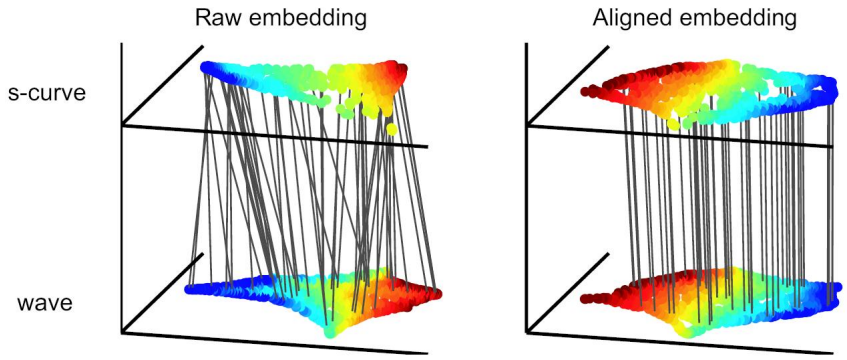


图 4.8: 原始流形和对齐后的流形对比

这样就回答了迁移学习的第二个问题 —— 如何迁移。LuMA的基本思想是利用$\mathcal{D}_S^l$和$\mathcal{D}_T^l$的对应关系，找出$\mathcal{D}_T^u$和$\mathcal{D}_S^l$之间的对应关系，这样就能扩大目标领域的有标记数据从而训练出有效的定位模型。

具体来说，LuMA算法的输入输出以及步骤如下：

输入：源领域的有标记数据$\mathcal{D}_S^l = \{(S_i^{(src)}, l_i)\}_{i=1}^{n_{ls}}$，目标领域的有标记数据$\mathcal{D}_T^l = \{(S_i^{(tar)}, l_i)\}_{i=1}^{n_{lt}}$和未标记数据$\mathcal{D}_T^u = \{(S_i^{(tar)})\}_{i=1}^{n_{ut}}$；

输出：目标领域的定位模型。

步骤：

1. 在$\mathcal{D}_S^l$中找出n_{lt}个数据点对应于$\mathcal{D}_T^l$，分别作为X_p和Y_p。要确保两列数据处于一一对应的顺序排列。$\mathcal{D}_S^l$中剩余的$n_{ls} - n_{lt}$个数据作为X_s，$\mathcal{D}_T^u$作为Y_s。

2. 将X_p和X_s组合为X并根据公式(4.1)计算拉普拉斯图矩阵L^x。类似可求得L^y。

3. 根据公式(4.8)计算连接图的拉普拉斯矩阵L^z。

4. 计算L^z的特征值，取最小的d_l个非零特征值对应的特征向量构造对齐的流形。

5. 对$\boldsymbol{g}_s$中每一个数据，其类标号可以通过在$\boldsymbol{f}_s$中找到的最近邻来赋予。这样可以得到一个自动标记的$\mathcal{D}_T^u$。

6. 用$\mathcal{D}_T^l$和自动标记的$\mathcal{D}_T^u$一起，可以通过有监督学习算法如支持向量机（SVM）、k-近邻、贝叶斯推断等建立目标领域的定位模型。

4.4　实验验证

为了验证所设计的LuMA算法的实际效果，即回答迁移学习的第三个问题何时迁移，我们在实际室内楼层中搭建了无线网络环境以及定位系统，并且采集了不同时间段不同设备的信号数据用于实验性能分析。

4.4.1　数据采集

我们采用第三章中描述的无线实验环境进行数据采集。由于实验涉及到不同时间段多设备的定位精度，因此需要采集多种情况下的Wi-Fi数据。我们采用两种类型的设备进行数据采集：一种是IBM® R60笔记本电脑，配有内置Intel® Pro/Wireless 3945ABG无线网卡；另一种是多普达O2智能手机，自带Wi-Fi模块。然后选择差异明显的两个时间段（上午 9:30上班后和下午 5:30下班后）各采集两套数据，为便于区别，我们将这四套数据分别命名为“笔记本上午”、“手机上午”、“笔记本下午”和“手机下午”。在下面的部分，我们将采用这些数据集来测试我们的算法性能。

4.4.2　参数设置

在LuMA算法中，参数μ不用设定，λ_1和λ_2可以根据数据集的大小来决定。我们发现当$\frac{\lambda_1}{\lambda_2} = \frac{n_{lt}+n_{ut}}{n_{ls}}$时流形对齐的效果最好。另外，在我们的方法中，维数$d_l$的选择也是一个问题，由于原始数据是7维，我们实验发现在 3维时的对齐效果最好，因此在实验中采用这样的参数设置。

关于定位模型的训练算法，我们发现简单的k-近邻算法也可以得到不错的结果，因此在算法的步骤6中采用k-近邻作为有监督学习算法。另外由于信号的强噪声，我们采用曼哈顿距离取代常用的欧式距离来计算距离差，这样能将更多的重点放在信号的缺失而不是强度的变化上，从而增加准确度。

在构造权重矩阵W时可以采用不同的算法。例如可以采用高斯处理$W_{ij} = e^{-|x_i-x_j|^2/2\sigma^2}$，这样得到的权重矩阵$W$是非负对称矩阵$W_{ij} = W_{ji} \geq 0$。而通过$k$个最近邻确定邻居关系得到的$W$不是对称矩阵，并且也可能存在负系数。例如采用最小二乘的方法构造权重系数：

$$W_{ij} = \underset{W}{\operatorname{argmin}} |x_i - \sum_{j\sim i} W_{ij}x_j|^2 \tag{4.9}$$

这样得到的W_{ij}是能最好逼近x_i的近邻的系数，并且通常是非对称的。LLE采用公式(4.9) 求权重矩阵，约束条件是$\sum_j W_{ij} = 1$，这样产生的系数可能出现负数。或者可以对该公式加上$W_{ij} \geq 0$ 的非负约束得到一组非负的凸系数进行凸近似。

对含有负系数的拉普拉斯矩阵L，我们可以通过构造半正定矩阵$L^T L$来优化下式：

$$\boldsymbol{f}^T L^T L \boldsymbol{f} = \sum_i |f_i - \sum_{j \sim i} W_{ij} f_j|^2 \tag{4.10}$$

由于$L^T L$是半正定矩阵并且满足$L^T L\mathbf{e} = 0$，它的特征向量也能用于构造低维流形用于最小化公式(4.10)。

我们在实验中比较了高斯、LLE和凸近似三种权重构造方法，其中LLE的权重构造方法得到的流形对齐效果最好，因此在后面的实验中都采用LLE的权重构造方法。

4.4.3 性能比较

为测试LuMA算法的迁移学习有效性，我们做了三组不同的实验：首先是对不同时间段的迁移，然后是在不同设备上的迁移，最后是不同时间不同设备的迁移。我们选取了三个不同的模型作为对比：

- 只采用源领域数据建立的模型（Source Data Model，SDM），用来给出旧定位模型造成的定位精度下降的程度。
- 只采用目标领域数据建立的模型（Target Data Model，TDM），用来给出目标领域可以达到的最好精度，是迁移学习算法能达到的上限。
- 采用线性迁移的模型（Linear Shift Model，LSM），即源领域和目标领域的对应位置上每个AP的信号强度都存在$s_i^{(S)} = \lambda s_i^{(T)} + \Delta s$，其中的$\lambda$和$\Delta s$可以通过训练得到。

首先测试在不同时间段的迁移，我们选取“笔记本上午”数据集作为源领域，目标领域是“笔记本下午”数据集，实验结果如图4.9所示。纵坐标轴表示测试集数据在3米误差距离内的精度，横坐标轴表示我们所选取的目标领域中的有标记数据数量即n_{lt} 占所有位置的百分比，即$\frac{n_{lt}}{m}$（以下各图同）。SDM的效果

与目标领域数据无关，但是最差，说明迁移学习对于定位精度的提升是有帮助的。LSM假定两个数据集之间存在线性关系进行迁移，但是信号随时间的变化并不是单纯的线性关系，因此LSM的结果不如LuMA。TDM开始的结果很不好，随着有标记数据的增加精度增长很快。但事实上，我们不可能得到超过50%的有标记数据作为参考点，这样在只有少量有标记数据时，我们的方法LuMA要有效得多。

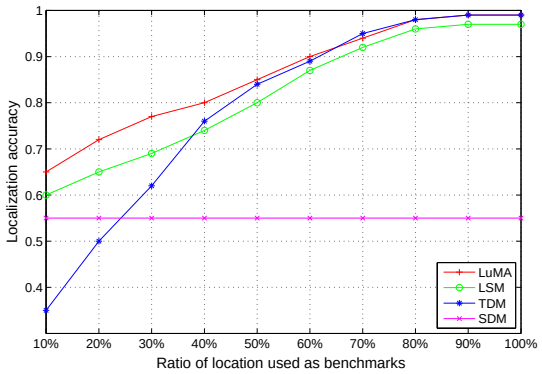


图 4.9: 信号分布受不同时间段影响的结果图

为了测试在不同设备上的迁移，我们选择“笔记本上午”数据集作为源领域，目标领域则是“手机上午”数据集，实验结果如图4.10所示。由于源领域和目标领域差异更大，SDM的效果比受时间影响更差。LuMA的效果也很不错略逊于LSM，可能的原因是不同设备感受信号的能力确实存在某种线性关系。

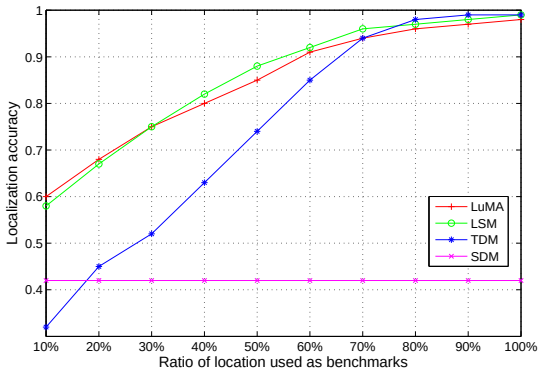
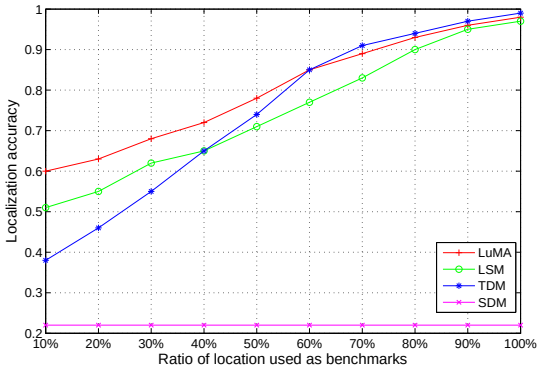
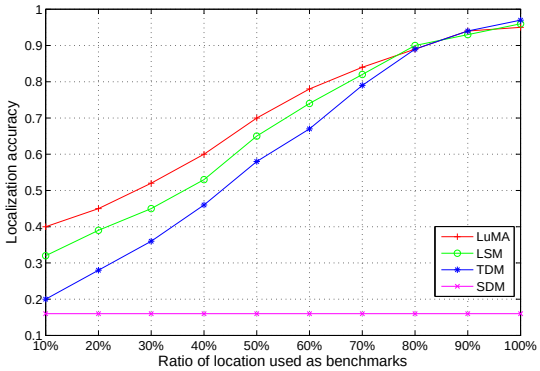


图 4.10: 信号分布受不同设备影响的结果图

最后考虑在不同时间不同设备上的迁移。这是一种最为常见的情况，因此我们分别选取不同的源领域和目标领域进行测试。首先还是以“笔记本上午”数据集为源领域，而目标领域则是“手机下午”数据集。实验结果如图4.11(a)所示，SDM的效果只有大约20%，因为受到时间和设备的双重影响旧模型失效很严重。当加入时间因素后，不同设备间的线性关系被打破，因此LuMA的性能表现远远好于LSM。然后我们以“手机上午”数据为源领域，目标领域为“笔记本下午”数据集，实验结果如图4.11(b)。由于笔记本的信号浮动比较大，定位的精度非常低。但是LuMA的结果相比之下也是最好的。



(a) 笔记本作为源领域，手机作为目标领域



(b) 手机作为源领域，笔记本作为目标领域

图 4.11: 信号分布同时受不同时间和设备影响的结果图

总体来说，迁移学习与只采用源领域数据建立的旧定位模型相比能够大大的提高定位的精度，并且在只有少部分目标领域数据的情况下确实能够从源领

域数据中迁移知识帮助提高目标领域的定位精度，而相对于简单的线性迁移方法来说，我们的基于流形对齐的迁移学习方法更加稳定有效。这一实验结果验证了我们所提出的迁移学习方法是有效的而不是会导致结果变差的负面迁移。

4.5　本章小结

本章首先分析了造成信号分布差异的两大因素 —— 时间和设备对定位模型的影响，描述了要解决的问题。接着形式化的定义了定位中的迁移学习问题，即要利用源领域大量的有标记数据帮助目标领域有限的有标记数据建立准确的定位模型。然后我们对信号特性进行了分析，主要是信号的非线性衰减特性和信号空间的流形假设，这样我们就能够通过流形学习的方法来进行信号空间之间的知识迁移。我们提出了一种流形对齐的降维方法，通过两个数据集之间部分的对应关系，我们在某个低维空间上找到这两个数据集对齐了的流形嵌入，通过该对齐的流形，能够建立两个数据集上的一一对应关系。基于该降维方法，我们提出了自适应定位算法LuMA，并在实际环境中采集的数据上验证了我们的算法的有效性。

第五章　基于位置的多媒体服务系统

随着移动计算技术的发展，为用户提供基于位置的多媒体服务是一种有实际意义的应用，譬如提供智能导游[12]、商场定制导购[20]以及交互式社交等。一方面的刺激因素是由于当今移动设备的多媒体支持能力不断增强，几乎现如今所有的手机都能支持视频、音频、图片以及其它媒体；另一方面，由于无线网络带宽的持续增长，多媒体内容的大数据量传输不再成为瓶颈，使得这一应用变得实际可行。

本章在无线定位的基础上，提出一种基于位置的多媒体服务原型系统，并在实际环境中实现了一个应用系统LMSS，能够根据移动用户的位置提供定制化的视频。

5.1　原型系统设计

我们首先描述两种基于位置的多媒体服务应用作为范例，然后提出适用于它们的原型系统。

5.1.1　应用场景

首先，游览者在博物馆参观的场景。简单的站在那里看陈列的展品肯定是比较枯燥乏味的方式，而跟随导游讲解只能得到模糊的概述。游览者更想看到的应该是精心准备的描述、丰富多彩的背景资料视频以及具有吸引力的讲解以便更好的了解展品，而不是传统的单调的参观方式。如果博物馆能够根据不同的游览者当前所处的位置为其递送丰富而且个性化的多媒体资源，游览者就能自动获取离他最近的展品的多媒体信息，这样的一种参观方式的体验完全不同于传统的参观活动。

第二个场景：消费者在超市购物。当消费者在超市中徘徊时，附近的打折商品和新品上市的信息对他们是有吸引力的并且应该及时的推荐给他们。从超市商家的角度看，根据消费者所在的位置来发放广告信息也会是有帮助的。例如消费者在饮料区域的时候，对其发放饮品广告是比较合适的。这样的广告既不会显得讨厌，又可以作为一种个人化的定制服务。另外，利用这种多媒体服务可以为休息区的顾客播放有趣的电影短片作为娱乐。

5.1.2 原型系统架构

为了满足以上描述的场景需求，我们提出了如图5.1所示的基于位置的多媒体服务系统。该原型系统主要包含两大任务：技术研究和内容管理。技术研究部分主要关注该原型系统中的技术层面，例如定位技术、网络管理和数据库技术等；内容管理部分关注服务层面，注重于多媒体资源的添加管理以及与用户的交互行为。

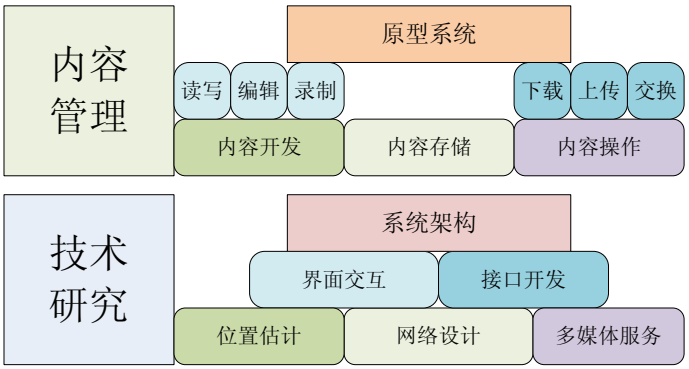


图 5.1: 基于位置的多媒体服务原型系统

具体来说，技术研究部分可以分为三个模块：位置估计、网络设计和多媒体服务。位置估计模块利用各种技术提供准确的物理位置信息；多媒体服务模块解决提供服务的流媒体方式和编解码等问题；随着客户端数目增加，网络设计模块为同时高效支持多客户端也变得很重要。

内容管理部分主要涉及和多媒体资源操作相关的模块。要提供丰富多彩的多媒体服务必须有大量的资源，另一个重要方面是用户与多媒体资源的交互过程。用户可能希望上载视频到服务器公开给大众，或者只想和朋友小范围的分享。因此，内容管理部分包含了所有对要递送的内容的操作行为。

5.2 LMSS系统

基于以上提出的原型系统，我们设计了一个基于位置的多媒体服务验证系统LMSS（Location-based Multimedia Service System），利用Wi-Fi定位技术为上层应用提供可靠的位置信息。下面对整体架构和每个部分进行具体描述。

5.2.1　系统架构

图5.2给出了我们所设计的LMSS的架构，采用了客户端-服务器结构模式。该系统能够同时为多个客户端提供多媒体服务，而客户端可以是任意具备Wi-Fi功能的设备。客户端与服务器间的通信都通过无线网络架构完成。

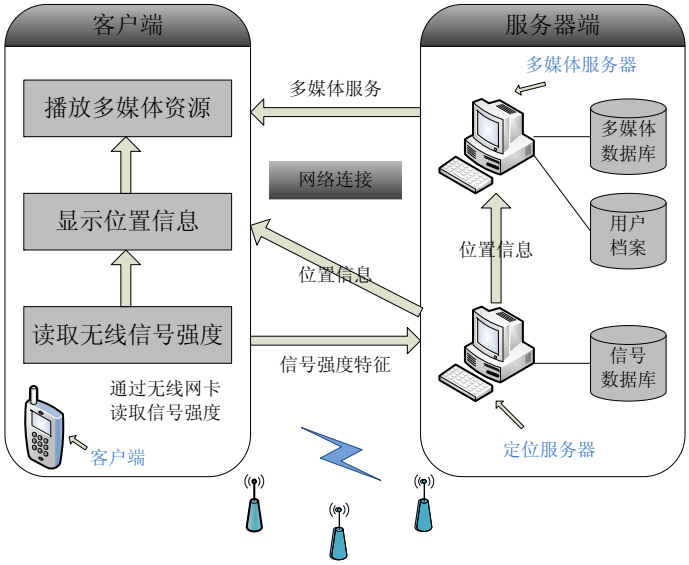


图 5.2: LMSS的系统架构和工作流程

图5.2也给出了LMSS的整个工作流程：当客户端请求多媒体服务时，它自动从当前无线网络环境中收集信号强度向量发送到定位服务器。定位服务器将该信号特征输入定位模型得到估计的位置信息，并且同时发送给客户端和多媒体服务器。然后多媒体服务器根据用户档案中所存储的位置和多媒体关联信息，从多媒体数据库中查询对应的多媒体资源并发送给客户端。最后客户端获得多媒体信息并进行播放。

下面我们对每一个部分的功能进行具体介绍。

5.2.2　客户端

LMSS的客户端是在移动设备上运行、用于和服务器交互的程序。我们开发了一套公有接口与服务器进行通信，并且实现了两种不同的移动客户端：Windows XP系统的笔记本电脑和Windows Mobile系统的智能手机。客户端主要功能包括：

1. 与服务器连接注册；

2. 收集并发送Wi-Fi信号特征；

3. 播放接收的多媒体资源。

为便于实现，我们开发了一套API，包括从无线驱动中获取信号强度，与服务器通信交换数据等。

5.2.3 服务器

服务器包含两类服务器：定位服务器和多媒体服务器，它们的具体功能列在表5.1中。

表 5.1: LMSS服务器的功能列表

服务器	功能
定位服务器	身份验证 位置估计 连接通信
多媒体服务器	位置语义提取 多媒体查询 连接通信

在定位服务器中，身份验证主要用于确保用户有权获得该服务，连接通信是与其它部分的应答响应，而核心功能就是位置估计。我们可以采用第三章中提出的统计与规则的定位方法来进行定位。然而由于信号分布易受时间和设备影响发生很大的变化，我们可以利用第四章提出的迁移学习算法如流形对齐[41]等来增强系统的自适应性。

定位服务器得到的位置信息通常都表示为(x, y, z)的坐标形式，对于应用层来说，采用语义形式的表示更加适用，例如在哪一楼层哪一个房间。因此多媒体服务器首先根据环境地图信息将位置翻译为语义表示。多媒体服务器的第二步是利用语义表示从多媒体数据库中查询对应的资源并发送给客户端，这个过程可以通过流媒体或者直接下载的方式完成。

多媒体资源数据库也是以语义表示的形式进行存储，如前所述，对该数据库的内容管理也是基于位置的多媒体服务的一大任务。目前LMSS只关注技术研究部分，而这部分我们留作将来的工作。

5.3　实验验证

我们在实际Wi-Fi环境中实现了LMSS并对其进行了评测，其中性能评测包括两部分：定位精度和功能表现。在前两章中我们针对定位精度做了仔细的评测，因此这里只对多媒体服务的性能表现予以评测。

5.3.1　实验环境

LMSS的实验环境和第三章中的实验环境相同，也在环境中布设了7个AP。我们利用两台外接TP-Link无线网卡的DELL® PC机分别作为定位服务器和多媒体服务器，移动设备则采用第四章中用到的笔记本和智能手机。在实验中我们简单设置服务器和客户端通过点对点Socket连接进行通信传输，而暂时不考虑复杂的网络拓扑结构。

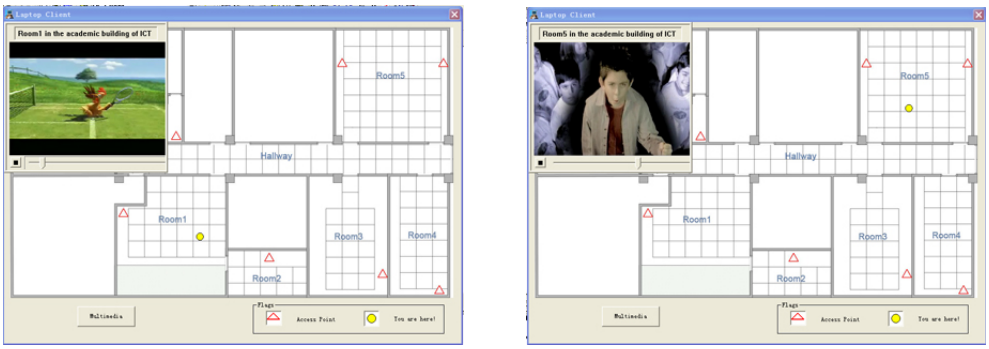
5.3.2　多媒体服务

为了实现之前所描述的应用场景，LMSS的主要功能就是根据移动设备的不同位置智能的提供不同的视频服务。例如，当移动用户进入某个房间时，他/她可能想得到一些关于该房间的描述信息，LMSS能够自动通过定位功能发送相关的视频给用户。我们在笔记本和智能手机上都测试了该功能，部分客户端截图如图5.3和5.4所示。

尽管这些应用都是提供相同的功能，但它们的交互界面由于不同的设备属性而略有差异。笔记本拥有足够大的显示屏，因此我们为用户提供尽可能多的信息：既显示了地图的全貌并且将用户的当前位置用小圆圈指示出来，又能在主窗口中嵌入一个小窗口播放接收到的视频。而对于手机有限的屏显来说，我们隐去了对用户位置的指示而直接全屏为用户播放视频。这样一种交互界面的设计也使得LMSS更加具有实用性。

5.4　本章小结

本章利用无线定位技术，提出了一种基于位置的多媒体服务系统，该系统根据移动用户的位置提供定制化的多媒体服务。我们首先提出了一种原型系统，



(a) 在房间1中的视频播放　　(b) 在房间5中的视频播放

图 5.3: LMSS的笔记本客户端效果图

(a) 在房间1中的视频播放　　(b) 在房间5中的视频播放

图 5.4: LMSS的手机客户端效果图

详述了各个模块的功能和目的。基于该原型系统，实现了一个基于位置的多媒体服务系统LMSS，通过定位服务器和多媒体服务器的协同为客户端提供基于位置的视频服务。我们在实际无线网络环境中对系统进行了测试，并且对用户交互界面进行了讨论。

第六章　结束语

本论文的主题是研究无线位置感知技术中的迁移学习问题。本章对本文的工作进行了总结，并指出了下一步可能的研究方向。

6.1　本文工作总结

在未来普适计算环境中，无线位置感知技术必将成为不可或缺的技术之一。获得目标的位置是各种计算服务的基础，也是必须的信息。随着无线位置感知技术的发展和各种因素促进大规模的无线设施布设，大范围的位置感知会成为主要的需求。一方面，在大规模环境中进行准确的定位本身就是一个难点问题；另一方面，在这样的环境中进行无线数据采集会花费大量的人力物力，而一旦环境变化，在已有数据失效导致定位精度下降的情况下，重复进行数据采集又会耗费人工标记成本。为了尽量减少数据的采集标记工作，对定位算法的自适应要求就变得越来越高。本文采用机器学习的方法对无线定位技术展开研究，首先提出了统计和规则结合的无线定位方法，能够在传统方法基础上进一步提高定位精度；其次针对环境变化对无线信号的影响，提出了一种基于迁移学习的无线定位方法，旨在解决信号分布变化的问题，实现自适应定位。

本文首先回顾了基于Wi-Fi的定位技术，包括不同的定位方法和定位系统。基于Wi-Fi的定位主要分为两大类：基于传播模型和基于学习模型。由于室内环境的复杂性，传播模型在室内环境中的建模是一个很大的问题，因而大部分室内Wi-Fi定位方法都采用了基于学习的方法。在这一类型的定位方法中，又出现了确定性模型和概率性模型的不同分支，并且针对信号特征提取、有标记训练数据有限等情况提出了基于核的方法以及半监督学习等学习算法。由于无线信号受到环境的影响会发生剧烈变化，而已有的定位模型在新领域的定位精度急剧下降，为了能够以较少的新领域的有标记数据建立准确有效的新定位模型，迁移学习这一机器学习方法被引入。迁移学习的目的在于将源领域的知识迁移到新的目标领域以解决目标领域训练数据不足的问题。我们对迁移学习进行了详细的介绍，并对已有的Wi-Fi定位的迁移学习解决方法的应用及其局限性进行了分析。

接着我们对定位问题进行形式化定义，并详述了在定位中广泛采用的统计方法，包括最近邻、朴素贝叶斯、核方法和直方图分布等单点定位算法，以及用于轨迹跟踪的HMM算法。由于统计方法在一些局部区域的判断不够精确，也有人采用基于规则的定位方法，例如决策树等。我们提出了一种基于相对信号强度关系的规则比较法，利用AP间的相对信号强弱关系建立判别规则，也能用于分类。在这之上，提出了统计和规则结合的无线定位方法，在传统的统计方法基础上加入规则库作为约束项，能够进一步提高统计方法的定位精度。另外，我们提出的基于规则的HMM将统计方法和规则相结合，能够改进HMM在轨迹跟踪问题上的效果。

接着我们论述了在实际无线定位系统中遇到的迁移学习问题，主要是多种因素同时对信号分布造成的影响，而已有的工作都没有考虑这样的情况。我们首先对信号特性进行分析，包括信号空间的流形假设和迁移可行性。然后提出了一种流形对齐的降维方法，根据两个数据集之间的部分数据的对应关系，将两个数据集的流形在低维嵌入上进行对齐，从而能够确定两个数据集上数据间的一一对应关系，实现数据集之间的知识迁移。基于该流形对齐的降维方法提出了一种自适应的定位算法LuMA，通过已有数据解决目标领域下训练数据不足的问题，并在实验环境中验证了算法的有效性。

最后基于定位原型系统，提出了一种基于位置的多媒体服务系统LMSS，根据移动用户的位置提供定制化的视频，作为基于位置的服务的一项实际应用。

本文的研究工作比较全面的研究了当前的基于Wi-Fi的定位技术，提出了统计和规则结合的无线定位方法用于提升单一定位方法的精度，提出了基于迁移学习的自适应定位方法，针对不同时间不同设备情况下信号分布差异导致定位模型精度下降的问题，大大减少数据的标记工作而保证有效的定位精度。不过由于时间仓促，本文的研究成果也存在一些不足的地方：

首先在统计和规则结合的定位方法中，本文只考虑了比较简单的信号间相对强弱关系，而没有建立更为细致的规则，这样就忽略了一些细节的信息，而这些细节信息能够帮助进行更准确的判别。

其次本文用到的自适应定位算法的基本思想是利用源领域的有标记数据来标记目标领域的无标记数据从而扩大目标领域的有标记数据，这一方法要求有足够的无标记数据用于训练。尽管无标记数据在实际中很容易获得，但对这部分数据的需求也是该方法的局限性之一。

另外本文所采用的流形对齐的降维方法是特征值求解问题，求解实对称矩阵的特征值问题比较有效的方法是Jacobi方法。而在数据量很大的情况下，要求解的矩阵是一个大的稀疏矩阵，Jacobi方法在处理这种矩阵时运算量会很大，需要耗费大量的时间。不过由于特征值求解主要用于知识迁移建立定位模型即训练过程，因而不会对实时定位阶段产生影响。

6.2　下一步研究方向

在本文工作的基础之上，下一步的研究方向可以有以下几点：

- **三维空间定位技术**。本文提出的定位中的迁移学习是基于二维平面进行的，而在实际生活中，对于三维空间的定位技术也有着迫切的需求。由于同一座大楼的不同楼层结构有着很大的相似性，如何利用当前楼层的数据建立其它楼层的定位模型，实现三维空间的准确定位，仍然是一个亟待解决的问题。

- **迁移学习的融合技术**。已有的迁移学习的研究工作大部分集中在利用一个源领域的知识帮助解决一个目标领域的一个任务，而如何能将多个源领域的知识融合在一起解决一个或多个目标领域的任务也是一个值得研究的问题。

参考文献

[1] Robert Akl, Dinesh Tummala, and Xinrong Li. Indoor propagation modeling at 2.4 ghz for ieee 802.11 networks. In *Wireless and Optical Communications*, 2006.

[2] Paramvir Bahl and Venkata N. Padmanabhan. Enhancements to the radar user location and tracking system. Technical report, Microsoft Research, 2000.

[3] Paramvir Bahl and Venkata N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *INFOCOM*, pages 775–784, 2000.

[4] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[5] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, November 2006.

[6] Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 1997.

[7] Yiqiang Chen, Juan Qi, and Zhuo Sun. Lmss: A location-based multimedia service system using wireless networks. *Intelligent Systems Design and Applications, International Conference on*, 3:33–38, 2008.

[8] Yiqiang Chen, Qiang Yang, Jie Yin, and Xiaoyong Chai. Power-efficient access-point selection for indoor location estimation. *IEEE Trans. Knowl. Data Eng.*, 18(7):877–888, 2006.

[9] K. W. Cheung, J. H. M. Sau, and R. D. Murch. A new empirical model for indoor propagation prediction. *IEEE Transactions on Vehicular Technology*, 1998.

[10] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Transferring naive bayes classifiers for text classification. In *AAAI*, pages 540–545, 2007.

[11] www.ekahau.com.

[12] Michael Epstein and Silvia Vergani. History unwired: mobile narrative in historic cities. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, pages 302–305, New York, NY, USA, 2006. ACM.

[13] Brian Ferris, Dieter Fox, and Neil Lawrence. Wifi-slam using gaussian process latent variable models. In *In Proceedings of IJCAI 2007*, pages 2480–2485, 2007.

[14] William G. Griswold, Patricia Shanahan, Steven W. Brown, Robert T. Boyer, Matt Ratto, R. Benjamin Shapiro, and Tan Minh Truong. Activecampus: Experiments in community-oriented ubiquitous computing. *IEEE Computer*, 37(10):73–81, 2004.

[15] J. Ham, D. Lee, and L. Saul. Semisupervised alignment of manifolds. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics 2004*, 2004.

[16] Jeffrey Hightower and Gaetano Borriello. Particle filters for location estimation in ubiquitous computing: A case study. In *In Proceedings of International Conference on Ubiquitous Computing (UbiComp*, pages 88–106, 2004.

[17] S. Huang, F. Proulx, and C. Ratti. ifind: a peer-to-peer application for real-time location monitoring on the mit campus. July 2007.

[18] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1989.

[19] K. Kaemarungsi and P. Krishnamurthy. Modeling of indoor positioning systems based on location fingerprinting. volume 2, pages 1012–1022 vol.2, March 2004.

[20] Takayuki Kanda, Dylan F. Glas, Masahiro Shiomi, Hiroshi Ishiguro, and Norihiro Hagita. Who will be the customer?: a social robot that anticipates people's behavior from their trajectories. In *UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing*, pages 380–389, New York, NY, USA, 2008. ACM.

[21] Mikkel Baun Kjærgaard and Carsten Valdemar Munk. Hyperbolic location fingerprinting: A calibration-free solution for handling differences in signal strength (concise contribution). In *PerCom*, pages 110–116, 2008.

[22] J. Krumm and E. Horvitz. Locadio: inferring motion and location from wi-fi signal strengths. pages 4–13, 2004.

[23] Axel Küpper. *Location-Based Services: Fundamentals and Operation*. Wiley, October 2005.

[24] Andrew M. Ladd, Kostas E. Bekris, Algis Rudys, Guillaume Marceau, Lydia E. Kavraki, and Dan S. Wallach. Robotics-based location sensing using wireless ethernet. In *MobiCom '02: Proceedings of the 8th annual international conference on Mobile computing and networking*, pages 227–238, New York, NY, USA, 2002. ACM.

[25] Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian E. Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, Fred Potter, Jason Tabert, Pauline Powledge, Gaetano Borriello, and Bill N. Schilit. Place lab: Device positioning using radio beacons in the wild. In *Pervasive*, pages 116–133, 2005.

[26] Lin Liao. *Location-based activity recognition*. PhD thesis, Seattle, WA, USA, 2006. Adviser-Fox,, Dieter and Adviser-Kautz,, Henry.

[27] Jeffrey J. Pan. *Learning-based Localization in Wireless and Sensor Networks*. PhD thesis, Hong Kong, China, 2007. Adviser-Yang,, Qiang.

[28] Jeffrey Junfeng Pan, James T. Kwok, Qiang Yang, and Yiqiang Chen. Accurate and low-cost location estimation using kernels. In *IJCAI*, pages 1366–1371, 2005.

[29] Jeffrey Junfeng Pan and Qiang Yang. Co-localization from labeled and unlabeled data using graph laplacian. In *IJCAI*, pages 2166–2171, 2007.

[30] Jeffrey Junfeng Pan, Qiang Yang, Hong Chang, and Dit-Yan Yeung. A manifold regularization approach to calibration reduction for sensor-network based tracking. In *AAAI*, 2006.

[31] Jeffrey Junfeng Pan, Qiang Yang, and Sinno Jialin Pan. Online co-localization in indoor wireless networks by dimension reduction. In *AAAI*, pages 1102–1107, 2007.

[32] Sinno J. Pan and Qiang Yang. A survey on transfer learning. Technical report, 2008.

[33] Sinno Jialin Pan, James T. Kwok, Qiang Yang, and Jeffrey Junfeng Pan. Adaptive localization in a dynamic wifi environment through multi-view learning. In *AAAI*, pages 1108–1113, 2007.

[34] Sinno Jialin Pan, Dou Shen, Qiang Yang, and James T. Kwok. Transferring localization models across space. In *AAAI*, pages 1383–1388, 2008.

[35] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.

[36] Rajat Raina, Andrew Y. Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *ICML*, pages 713–720, 2006.

[37] Teemu Roos, Petri Myllymäki, Henry Tirri, Pauli Misikangas, and Juha Sievänen. A probabilistic approach to wlan user location estimation. *In-*

ternational Journal of Wireless Information Networks, 9(3):155–164, July 2002.

[38] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[39] www.skyhookwireless.com.

[40] David C. Steere, Antonio Baptista, Dylan McNamee, Calton Pu, and Jonathan Walpole. Research challenges in environmental observation and forecasting systems. In *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 292–299, New York, NY, USA, 2000. ACM.

[41] Zhuo Sun, Yiqiang Chen, Juan Qi, and Junfa Liu. Adaptive localization through transfer learning in indoor wi-fi environment. *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, pages 331–336, Dec. 2008.

[42] Ping Tao, Algis Rudys, Andrew M. Ladd, and Dan S. Wallach. Wireless lan location-sensing for security applications. In *Proceedings of the Second ACM Workshop on Wireless Security (WiSe)*, San Diego, CA, September 2003.

[43] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[44] Arvin Wen Tsui, Vincent Lin, and Haohua Chu. Analysis and comparison between war driving and war walking in metropolitan wifi radio maps. September 2008.

[45] T.L.M. van Kasteren, G. Englebienne, and B.J.A. Kröse. Recognizing activities in multiple contexts using transfer learning. In *AAAI Fall 2008 Symposium: AI in Eldercare*, 2008.

[46] Pengcheng Wu and Thomas G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *ICML*, 2004.

[47] Qiang Yang, Sinno Jialin Pan, and Vincent Wenchen Zheng. Estimating location using wi-fi. 23(1):8–13, 2008.

[48] Jie Yin, Xiaoyong Chai, and Qiang Yang. High-level goal recognition in a wireless lan. In *AAAI*, pages 578–584, 2004.

[49] Jie Yin, Qiang Yang, and Lionel M. Ni. Adaptive temporal radio maps for indoor location estimation. In *PerCom*, pages 85–94, 2005.

[50] Moustafa Youssef and Ashok K. Agrawala. The horus wlan location determination system. In *MobiSys*, pages 205–218, 2005.

[51] Vincent Wenchen Zheng, Sinno Jialin Pan, Qiang Yang, and Jeffrey Junfeng Pan. Transferring multi-device localization models using latent multi-task learning. In *AAAI*, pages 1427–1432, 2008.

[52] Vincent Wenchen Zheng, Evan Wei Xiang, Qiang Yang, and Dou Shen. Transferring localization models over time. In *AAAI*, pages 1421–1426, 2008.

[53] Xiaojin Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.

致　　谢

倏然间，三年硕士生活行将结束，自己的求学生涯也将暂告一段落。回首在计算所攻读硕士学位的三年历程，尽管汗水与泪水洒满了这条学术之路，但在各方面的关心、支持、鼓励和帮助下，我得以完成这篇饱含感激之情的学位论文。谨以此文向多年来对我给予关怀的师长、同学、朋友和亲人表示诚挚的感谢！

首先，我要衷心感谢我的导师陈益强副研究员对我生活上、学习上无微不至的关心。他以前瞻而开阔的学术视野，在最初的论文选题上就为我指明了方向，无线位置感知技术作为机器学习和普适计算的交叉领域，既能在学习算法的深度上进行研究，也能在普适应用上做出成果。在课题组内陈老师给予了我充分的发挥空间，很多事情都放手让我去做，支持我到香港科技大学进行访问学习，不仅锻炼了我独当一面的能力，也让我在我所感兴趣的领域不断学习一直前进，取得了今天的成绩。对于我在论文研究上遇到的难题，他给予我师长般的悉心指导；而对于我在出国就业等人生道路上的选择，他又如朋友般的与我谈心。再次对这三年中陈老师耳提面命的指导帮助致以敬意和感谢！

其次我要感谢课题组并肩作战的同事和同学。刘军发博士、胡明清博士和纪雯博士，以及其他同事和同学：齐娟、朴松梅、董芳芳、陈振宇、张亚东、宁琼、谢金晶、潘伟、颜庆聪、唐晓庆、高兴宇、陈磊、刑云冰等，也对我的研究方向给予了很大的指导和帮助。

我在香港科技大学计算机科学与工程系访问期间，指导老师杨强教授在学术研究上也给予我很大的帮助。杨老师严谨求真的学术态度、一丝不苟的工作精神给我留下了深刻的影响，引导我在学术道路上一步一个脚印的踏实前进。也要感谢香港科技大学的同学对我的帮助，感谢Vincent、Nathan、徐倩、Junfeng、Derek、Evan、施潇潇、李斌、Hankz等。这段经历会深刻影响我日后的工作和生活。

一直以来，我的父母也无时无刻不在关注我的生活和学习状况，时常询问我的学习进展和遇到的困难，在生活上嘘寒问暖，并一直鼓励和支持我的学术选择。对父母我永远抱有感激之情！

在完成这篇短短的硕士论文的过程中，我学到的不仅仅是领域相关的知识，更重要的是学到的做人和做事的方法和态度。一种合理的方法，一份认真的态度，将会陪伴我一生，令我受益无穷。

谨把本文献给为我辛勤操劳半生的父母!

简　　历

基本情况

孙卓，男，湖北省潜江市人，1985年8月出生。

教育状况

2006.9–2009.7，中国科学院计算技术研究所，硕士，专业：计算机应用。

2002.9–2006.7，清华大学计算机科学与技术系，本科，专业：计算机科学。

研究兴趣

机器学习：半监督学习，迁移学习，流形学习等；

普适计算：上下文相关计算，基于位置的服务，基于传感器的行为识别等。

攻读硕士学位期间发表的论文

Z. Sun, Y. Chen, J. Qi and J. Liu. Adaptive Localization Through Transfer Learning in Indoor Wi-Fi Environment. In *Proceedings of the Seventh International Conference on Machine Learning and Applications (ICMLA'08)*. San Diego, California, USA. December 11-13, 2008.

Y. Chen, **Z. Sun**, J. Qi, D. H. Hu and Q. Yang. LoSeCo: Localization-based Search Computing for Pervasive Device Augmentation. In *Proceedings of PerCom-09 Workshop on Intelligent Pervasive Devices (PerDev09)*. Galveston, Texas, USA. March 9-13, 2009.

Z. Sun, Y. Chen, and J. Qi. Learn Traffic State Based on Cooperative Localization. In *UbiComp'08 Workshop - Devices that Alter Perception (DAP 2008)*. COEX, Seoul, South Korea. September 21st, 2008.

攻读硕士学位期间申请的专利

孙卓，陈益强，齐娟，刘军发，一种基于迁移学习的室内Wi-Fi定位方法（发明），审理中。

攻读硕士学位期间参加的科研项目

国家高技术研究发展计划（863计划）课题“支持无线标准的低功耗智能跟踪定位终端及系统研发”（2007AA01Z305）

攻读硕士学位期间的获奖情况

2009年获中科院计算所北纬通信硕士奖学金

2008年获中科院计算所联想硕士奖学金

2007年IEEE ICDM 数据挖掘竞赛第二名

2009年被评为中国科学院研究生院“三好学生”

2008年被评为中国科学院研究生院“三好学生”

2007年被评为中国科学院研究生院“优秀学生干部”