基于图像检索的室内定位系统的设计与实现

装 订 线

摘 要

当代的建筑越来越向天空和地下发展，也越来越向复杂的建筑群发展。在这样一个钢筋水泥的迷宫中，人们必然想在室内也能精确定位自己的位置。因为室内的 GPS 信号微弱而不稳定，以图像等为信号源的室内定位研究为室内高效定位提供了可能。智能手机上的高清摄像装置所拍摄的图像，本身就携带者丰富的空间信息，利用数字图像处理技术来获取图像特征，再结合机器学习来确定场景位置是室内定位研究的重要研究方向。

本文联系了文本分析和分类方法，先用数字图像处理技术提取图像的视觉特征信息，经过聚类处理后，构造 bag-of-features 模型，形成视觉词汇。利用这一思想，本文提出了一种基于图像的高效定位系统。将室内场景的图像进行特征提取，再将特征数据聚类为每张图片构造视觉词汇，即 BOF 模型，最后通过机器学习训练场景模型，用得到的模型来对用户输入的问询图像进行场景判别。论文研究的主要内容为：

（1）设计一个系统，由用户提供问询图像请求，在服务器端利用 SURF 特征提取，词袋模型，场景模型预测等图像检索技术得到最佳匹配位置。

（2）对于图像特征提取，比较了 SIFT 和 SURF 两种算法，然后对图像特征提取流程进行优化。

（3）对于通过聚类构造视觉词袋模型，优化聚类参数，使得聚类结果可以最大保存图像的有效特征信息。

（4）对于训练场景分类模型，通过比较多种分类算法，选取以支持向量机 SVM 为主的分类方法建立模型，得到准确率超过 95%的效果。

关键词：室内定位，计算机视觉，SURF，词袋模型，支持向量机

Design and Implementation of Indoor Location System Based on Image Retrieval

ABSTRACT

Contemporary buildings are increasingly extending their roots into the sky and underground, and obviously having a tendency towards greater complexity. In such a reinforced concrete maze, people really want to accurately locate their own position indoors. Though the indoor GPS signal is weak and unstable, it is possible to get the location by studying image as signal source for indoor location. The image taken by the high-definition camera on the smart phone itself is based on the rich spatial information of the carrier, the use of digital image processing technology to obtain the image characteristics, combined with machine learning to determine the scene location is an important research direction of indoor location research.

In this paper, we link the text analysis and classification method. Firstly, we use the digital image processing technology to extract the visual feature information from the image. After clustering, we construct the bag-of-features model to form the visual vocabulary. Using this idea, this paper presents an image-based efficient positioning system. The feature data is clustered into the visual vocabulary of each picture, that is, the BOF model, and finally through the machine learning and training scene model, the obtained model is used to identify the scene of the query image input by the user. The main contents of the thesis are:

(1) Design a system, by the user to provide inquiry image request, in the server side using SURF feature extraction, bag-of-features model, scene model prediction and other image retrieval technology to get the best match position.

(2) For image feature extraction, we compare SIFT and SURF algorithms, and then optimize the image feature extraction process.

(3) In order to construct the visual bag model by clustering, the clustering parameters are optimized so that the clustering results can save the effective feature information of the image.

(4) For the training scene classification model, by comparing the various classification algorithms, we choose the SVM-based classification method to establish the model, and get the effect of exceeding 95%.

Key words： indoor location, computer vision, SURF, Bag-of-Features, SVM

装 订 线

目 录

装 订 线

装

订

线

1 绪 论

1.1 研究背景

城市和乡镇的基础设施发展日新月异，规模和复杂度都在迅速增加，导致人们对于精确定位和导航的需求日益增加。另外在移动智能设备占有率飞速提升的大背景下，人们对能通过手边的智能设备来满足定位的需求更加强烈。在这种大环境下，室外定位技术已经发展的相当成熟，比如 GPS 卫星定位系统，全球全天候定位，民用相对定位精度在 5m 以内。在国内，除了西藏外，大陆的 GPRS 信号覆盖率达到了 92%以上，基本上有移动信号的地方都可以用 GPS，在个人用户和车载系统上非常普遍。但是常用的 GPS 或 AGPS 定位都是依靠卫星及手机基站，有部分没有 GPS 功能的低端手机也可以借助手机通讯基站的 LBS 地理位置信息服务来进行定位，然而如果是在地下室或者大型室内场所，信号穿透水泥墙后衰减比较大，精度会大大降低，这些定位技术便英雄无用武之地，这时便需要借助其他信号源来进行室内定位。所以如何在视觉相似度比较高的室内环境场景中实现比较精确的定位，成为了室内定位技术研究的难点和热点。除了人类用户，在机器人领域，室内定位的研究同样迫切，因为大量的机器人需要在室内工作。所以我也在机器人室内定位方向查询了很多资料。

精确的室内定位服务应用范围非常广，内容也极为丰富，具体的应用场景如下：

（1）大型建筑物内的室内定位

大型购物商场内的消费者，可以直接被引导到自己期望的或推荐的商品柜台；博物馆、艺术馆等大型展览场所内的参观者可以通过室内定位快速找到想去的展厅；机场、火车站、地铁站内的乘客可以在人流量大的情况下不迷失方向，节省时间。

（2）公共安全

使用室内定位可以很快找到同伴和自己的位置，解决走失问题；发生紧急状况时，可以引导用户快速逃生或者帮住救援人员确定求救者的位置。

（3）大数据收集

室内定位技术的使用能获取用户的位置信息，而位置信息在大数据时代具有很高的商业和社会价值，商家们可以通过消费者停留时间和路径流量的信息筛选出商铺、广告的黄金位置；建筑规划者可以通过用户的位置信息优化室内建筑布局，使其更加人性化，舒适便捷。

1.2 室内定位发展状况

现在比较成熟的室内定位技术的包括 Wi-Fi、ZigBee、RFID、蓝牙、红外线、射频识别等。但是在精确度、穿透性、抗干扰性、布局复杂程度和成本方面都各有优缺点，所以很少有成熟的商业产品投入市场。比如 Wi-Fi 技术，虽然现在大型商场内部都有 WiFi 信号，但是信息收发器很多，在楼层定位上很容易出错。另外 Wi-Fi 收发器的覆盖半径小于 90 米，所以很容易受到其他信号的干扰。

除此之外，另外还有信号来源于计算机视觉、图像、磁场以及信标等等，但是并没有成熟的

商用产品，还在实验室阶段。在机器人视觉领域，定位问题是个经典问题，也是 MIT 教授 John J. Leonard 和原悉尼大学教授 Hugh Durrant-Whyte[7]关于机器人自主导航提出的三个问题里的第一个：Where am I? 机器人的定位方式包括相对定位和绝对定位。相对定位法利用各种传感器获取机器人的运动状态信息，通过递推累计公式或者积分获得位置信息，但是存在有累计误差的问题，随着运行时间和轨迹的增加，误差也累计增大。绝对定位法是机器人通过获得外界一些位置等已知的参照信息，通过计算自己与参照信息之间的相互关系，进而计算出自己的位置，主要采用基于信标的定位、环境地图模型匹配定位、视觉定位等方法。但是信标定位要求标志物明显，容易辨识；环境地图模型匹配只适于一些结构相对简单的环境，条件限制严格。基于视觉的定位中，单目视觉无法直接得到目标的三维信息。只能通过移动获得环境中特征点的深度信息，适用于工作任务比较简单且深度信息要求不高的情况。双目立体视觉三位测量是基于视差原理获取对应点的三维坐标。这两种视觉定位的图像处理过程都很复杂，有很多视觉模型的研究，比如南加州大学 Itti 教授提出的基于显著性的视觉注意模型(Saliency)。

基于图像的室内定位也是利用计算机视觉的室内定位技术的另一发展趋势。它的过程很容易理解，就是通过一张问询图片得到未知的反馈，对于用户来说使用十分便捷。但是以图像为地理位置信号源有个共同的问题，即室内场景的视觉相似度很高，因此容易造成错误匹配，还有就是如果采用图像匹配算法来检索时间成本高，而如果采用机器学习训练数据模型，模型的精度也需要大量数据的支持，数据预处理时间长。如何能将室内定位做到真正高效也是一大挑战。

1.3 场景图像分类技术的发展和研究情况

场景图像分类是机器人学和计算机视觉领域一个非常重要的课题。所谓场景图像分类，是指对给定的图像，通过观察它所包含的内容，进而判断其拍摄场景的类别。在计算机视觉领域，随着互联网多媒体技术的迅速发展，涌现出海量的复杂数据，为了有效的对这些数据进行分析和管理，需要根据图像内容为其贴上语义标签，而场景图像分类恰巧是解决该问题的一种重要途径[1]。近年来，随着图像处理和模式识别的迅速发展，场景图像分类已成为当前的一个热门研究课题。

常见场景可以大致分为 4 类[2]：自然场景、城市场景、室内场景和事件场景。由于不同场景构成元素额差别较大，同一种分类方法在不同的场景数据集上的分类效果经常存在较大的差异，而这种差异在室外场景和室内场景之间尤为显著。室内定位所需要的室内场景分类中，面临着很大的困难，其主要原因有如下几个方面：其一，同类场景的类内变化较大，由于室内场景本身的复杂性和多样性，在同一类场景下拍摄到的场景图像差异很大；其二，拍摄时的外部因素干扰，在同一场景下，不同的拍摄角度会造成场景图像之间的视觉差异，而拍摄时的光线、遮挡、分辨率也常常使得这种差异更加明显。因此，室内场景分类中一个关键环节就是如何为图像建立一种有效的表示，该表示既可以稳定地获取反映场景类别的结构信息，又可以抑制纹理等细节上的不同差异。

在场景分类的历史上，SIFT（scale-invariant feature transform）[3]和 GIST 是两种比较流行的图像描述子，由 Lowe[4]提出的 SIFT 特征，是为了识别在不同图像中出现的同一目标，它对于平移、缩放、旋转、光照甚至遮挡等情况都能保持一定的稳定性，具有强大、突出的辨别能力。后

来在 Lazebnik 等[1]基于 SIFT 特征提出了空间金字塔匹配模型 SPM（spatial pyramid matching），使得 SIFT 在场景分类中表现出了很好的性能。在 2006 年的 ECCV 大会上，SURF（Speeded Up Robust Features，加速稳健体征）被首次发表[6]，这是一种稳健的图像识别和描述算法，部分灵感来自于 SIFT 算法。SURF 标准的版本比 SIFT 要快数倍，并且其作者声称在不同图像变换方面比 SIFT 更加稳健。SURF 算法的概念及步骤均建立在 SIFT 之上，但详细的流程略有不同。由 Oliva 和 Torralba[5]提出的 GIST 特征是为了捕获图像中反映场景类别的空间结构特性，忽略其中所包含的物体或背景的细微纹理信息。这种设计理念使得它在室外场景中能获得较好的性能，但在室内场景中的表现却不尽人意。由 Dalal 和 Triggs[8]提出的用于人脸检测的 HOG（histograms of oriented gradients）特征，一经提出就在目标检测上得到了广泛应用。而 HOG 特征最成功的的应用是由 Felzenszwalb 等[9]提出的 DPM（deformable part model）模型，后来经过 Pandey 和 Lazebnik[10]改进，被应用到了场景分类之中。这种方法利用了 DPM 精确检测特定目标的优势，使得它能够在整体分类精度上有一定的提高。但是这种方法稳定性较差，从而导致不同类别之间的分类精度相差很大。Wu 等[11]提出的 CENTRIST（census transform histogram）特征，能够通过对图像局部信息建模，获得一种图像整体表示。并且能够捕捉到概要性质的简单几何位置。这种表示方法在主流场景分类数据集上取得了很好的效果，但是由于其不具有缩放和旋转不变性，使得它的应用范围受到约束。由 Li 等[12]提出的 Object Bank 方法则是场景分类研究中的又一标志性成果，他们认为一类场景由一系列的目标组成。即一组特定目标的集合可以确定一类场景。这种想法显然更加符合实际情况，与前述方法不同的是该方法试图从高层语义的角度入手，以一系列目标检测子的多尺度响应图为基础构建其特征向量。该特征由于能够捕获图像中蕴含的高层语义属性，在各类场景数据集上都有比较理想的表现。另外，由 Sadeghi 等[13]提出的 LPR（latent pyramidal regions）方法及由 Juneja 等[14]提出的 Bag of Parts 方法也都是基于高层语义的有效场景分类方法，它们的核心思想都是利用多个区域过滤器在图像金字塔上的多尺度响应图来捕捉图像蕴含的结构信息，进而在理解场景语义的基础上对场景类别做出正确的估计。

当场景种类达到千类以上且数据库容量突破百万张时，传统的基于底层特征和高层语义的方法经常难以处理这些海量数据，而基于深度学习的方法则在这种大数据上有着很好的表现。尤其是深度卷积神经网络在场景分类任务中已经取得了全新的突破。这种卷积神经网络能从大量的图像数据中学习到图像的一些共有深层属性，基于该网络的响应特征已经慢慢成为了图像识别方面的一种通用表达。但是在数据集不是很大的时候，深度学习的效果在测试集上却不佳，因此更多的学者开始采用支持向量机、Boosting、最近邻等分类器来构建浅层机器学习模型。基于词袋模型的图像识别，把每幅图像描述成为一个局部区域/关键点特征的无序集合，将图像中的局部特征聚类成为视觉词汇，相当于文本检索中的词，所有的视觉词汇作为码字又构成码书，在某种监督学习的策略下，对码书中的特征向量进行训练，获得对象或场景的分类模型。所以这是一种仿照文本检索领域 Bag-of-Words 的方法，称为 Bag-of-Features。Bag-of-Words 在计算机视觉中的应用首先出现在 Andrew Zisserman[15]中为解决对视频场景的搜索，其提出了使用 Bag-of-Words 关键点投影的方法来表示图像信息。后续更多的研究者归结此方法为 Bag-of-Features，并用于图像分类、目标识别和图像检索。这种方法主要实现了将任意图片的所有特征用一个固定维数的向量表示，

并且这个维数并不因图片特征点数量不同而变化。但是同时也会将图像表示成一个无序局部特征集的特征包，丢掉了所有的关于空间特征布局的信息，在描述上有一定的局限性。

1.4 主要工作

在本篇论文中，提出了一种基于视觉的高效室内定位系统，利用词袋模型算法，将请求定位的图像信息提取特征，利用已经训练好的场景分类模型，得到该图像的所属场景类别，实现定位。该系统结合了 SURF 特征提取，词袋模型，机器学习分类算法，位置匹配针对室内场景图像进行查找和检索。具体来说主要做了以下工作：

（1）采集数据，数据来源为学校图书馆和网上数据库。

（2）数据预处理，标记场景类型。

（3）基于 SURF 特征建立词袋模型，比较 SIFT 特征，获得室内场景图像的视觉词汇，构成字典。

（4）训练场景分类模型，分别使用支持向量机，逻辑斯蒂和决策树。

（5）测试分类效果。

（6）编写服务器和客户端，对结果进行展示。

1.5 本文章节安排

本文一共分为五个章节。

第一章为绪论部分。大致介绍了室内定位大发展现状和研究背景，以及场景分类的发展现状，并大致说明了本文的主要工作。

第二章为后续的实现部分做了理论基础准备，首先介绍了图像特征提取算法，然后介绍了核心词袋模型的构建，最后讲述了几个分类算法。

第三章是室内定位系统的实现部分，及具体如何用第二章所讲的理论实现一个完整的有室内定位功能的系统。然后分别介绍了每个实现的细节。

第四章是实验部分。以同济大学图书馆为室内场景取景地，按照第三章的实现，实现了整个系统，然后对图像特征提取算法的比较，分类算法的比较进行了实验，并分析比对了是按结果。

第五章是总结和展望部分。总结了本课题的研究结果和意义，并对未来的发展提出了建议。

2 相关理论基础简介

本文提出的基于视觉的室内定位研究是根据已知室内场景图像，进行特征提取和量化后，输入已经训练好的场景分类模型中，得到最相似的场景，实现定位。所以之前必须有一个场景数据库，重复提取特征，聚类构造字典的过程，然后训练一个多类分类器，将每张图片的词袋模型作为特征向量，将该张图片所属类别作为标签。除了用分类器，也可以用相关性排序的方法来检索图像，但是训练样本多的时候，这种方法就比较耗时。

场景分类过程是一系列技术的组合，主要技术包括了特征提取、特征分类、训练模型和结果预测。具体来说，场景分类的一般流程如图 2.1：

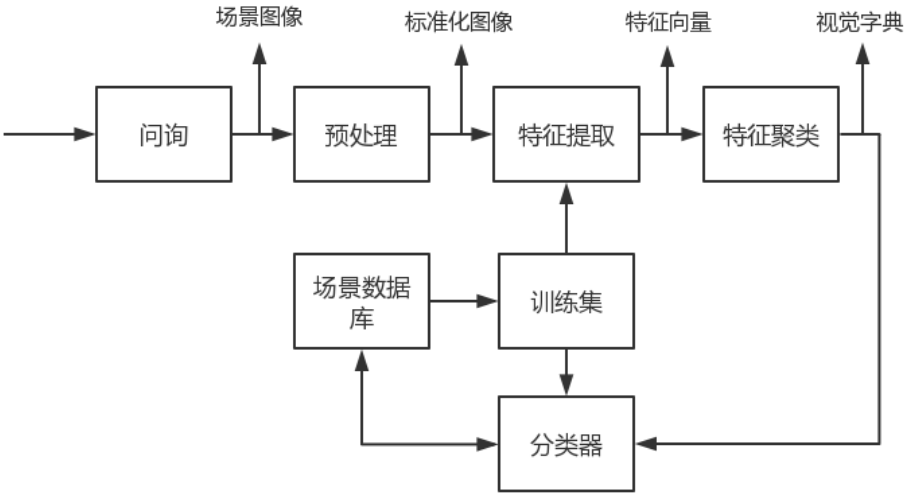

图 2.1 场景分类流程

2.1 SURF 特征

SURF（Speeded Up Robust Features）是一种稳健的图像识别和描述算法，首先于 2006 年发表在 ECCV 大会上。这个算法可被用于计算机视觉任务，如物体识别和 3D 重构。它部分的灵感来自于 SIFT 算法[16]。SURF 的标准版本比 SIFT 要快数倍，并且其作者声称在不同图像变换方面比 SIFT 更加稳健。

SURF 使用海森矩阵的行列式值作特征点检测并用积分图加速运算。SURF 的描述子基于 2D 离散小波变换响应并且有效利用了积分图。

SURF 算法可以说是 SIFT 算法的加速版，其快速的基础实际上只有一个，即积分图像 haar 求导。SURF 特征提取算法主要有 5 个步骤，下面我们一一详细叙述。

2.1.1 构建海森（Hessian）矩阵

SIFT 算法建立一幅图像的金字塔，在每一层进行高斯滤波并求取图相差（DOG）进行特征点的提取，而 SURF 则用的是海森矩阵（Hessian Matrix）[17]进行特征点的提取，所以海森矩阵是 SURF 算法的核心。

假设函数$f(x,y)$，Hessian 矩阵$\mathcal{H}$是由函数偏导数组成。图像中某个像素点的 Hessian Matrix 的定义为

$$\mathcal{H}\big(f(x,y)\big)=\begin{bmatrix}\dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x\partial y}\\ \dfrac{\partial^2 f}{\partial x\partial y} & \dfrac{\partial^2 f}{\partial y^2}\end{bmatrix} \tag{2.1}$$

从而每一个像素点都可以求出一个 Hessian Matrix。Hessian Matrix 的判别式为：

$$\det(\mathcal{H})=\frac{\partial^2 \mathrm{f}}{\partial \mathrm{x}^2}\frac{\partial^2 \mathrm{f}}{\partial \mathrm{y}^2}-\left(\frac{\partial^2 \mathrm{f}}{\partial \mathrm{x}\,\partial \mathrm{y}}\right)^2 \tag{2.2}$$

判别式的值是$\mathcal{H}$矩阵的特征值，可以利用判定结果的符号将所有点分类，根据判别式取值正负，从而判别该点是或不是极点的值。在SURF算法中，通常用图像像素I(x,y)取代函数值$f(x,y)$，然后选用二阶标准高斯函数作为滤波器。通过特定核间的卷积计算二阶偏导数，这样便能计算出海森矩阵的三个矩阵元素L_{xx}，L_{xy}，L_{yy}，从而计算出海森矩阵公式为：

$$\mathcal{H}(x,\sigma)=\begin{bmatrix}L_{xx}(x.\sigma) & L_{xy}(x,\sigma)\\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma)\end{bmatrix} \quad (2.3)$$

其中$L_{xx}(x,\sigma)$是二阶标准高斯函数$\frac{\partial^2}{\partial x^2}g(\sigma)$与图像 I 在点 x 处的卷积，$L_{xy}(x,\sigma)$和$L(x,\sigma)$同理。详细的说，由于我们的特征点需要尺度无关性，所以在进行海森矩阵构造前，需要对其进行高斯滤波。这样，经过滤波后再进行海森矩阵的计算，其公式为

$$L(x,t)=G(t)\cdot I(x,t) \quad (2.4)$$

$L(x,t)$是一幅图像在不同解析度下的表示，可以利用高斯核$G(t)$与图像$I(x)$在点 x 的卷积来实现。通过这种方法可以为图像中每个像素计算出其海森矩阵的决定值，并用这个值来判别特征点。据[18]得知，高斯函数对尺度空间的分析来说是最优的选择。然而在实践中，高斯函数需要被离散化和裁剪。[19]因为高斯滤波器在很多情况下都不是理想的，并且在 Lowe 在 LoG 中的成功背景下，Herbert Bay 等用盒式滤波器进一步推进用近似现代值代替$L(x,t)$。这些近似二阶高斯导数，在用积分图计算时非常快。图 2.2 的结果显示了用离散化和裁剪的高斯核和近似高斯值的性能比较。

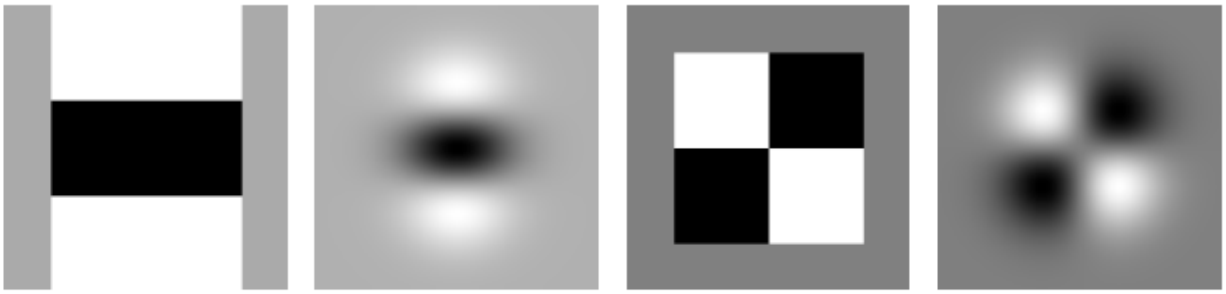

图 2.2 从左至右依次表示的就是在 y 方向用近似值的盒式滤波器和用高斯平滑后求二阶导数的比较以及在 xy 混合方向上用近似值的盒式滤波器和用高斯平滑后求偏导数的比较。

用近似值代替的一大优点就是通过运用积分图，与盒式滤波器做卷积是非常容易运算的。图 2.3 中的盒式滤波器是对$\sigma=1.2$的高斯二阶导数的近似，并且代表最低尺寸。

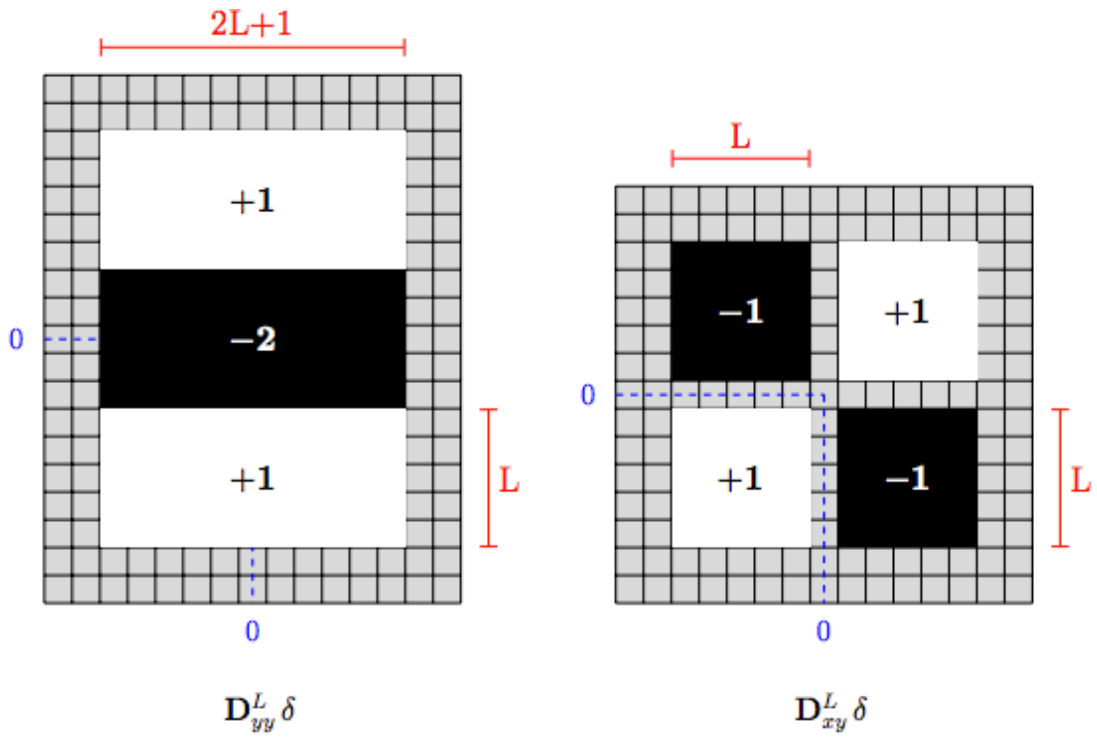

图 2.3 参数 L=5 的二阶盒式滤波器

用D_{xx}，D_{yy}和D_{xy}来表示近似值，为了计算的高效性，保持每个矩形区域的权重简单。但是为了平衡海森行列式表达式中的相对权重，用 Frobenius 规范得

$$\frac{\left|L_{xy}(1.2)\right|_F\left|D_{xx}(9)\right|_F}{\left|L_{xx}(1.2)\right|_F\left|D_{xy}(9)\right|_F}=0.912\cdots\simeq 0.9 \qquad (2.5)$$

则海森矩阵判别式可表示为：

$$det(\mathcal{H}_{approx})=D_{xx}D_{yy}-\left(0.9D_{xy}\right)^2 \qquad (2.6)$$

2.1.2 生成尺度空间

尺度空间理论是早期视觉操作的框架，由计算机视觉界开发，以处理图像数据的多尺度性质。上面所讲得到的海森行列式图类似于 SIFT 中的 DOG 图，但是在金字塔中图像分为很多层，每一层叫做一个 octave，每一个 octave 中又有几张尺度不同的图片，如图 2.4 所示。

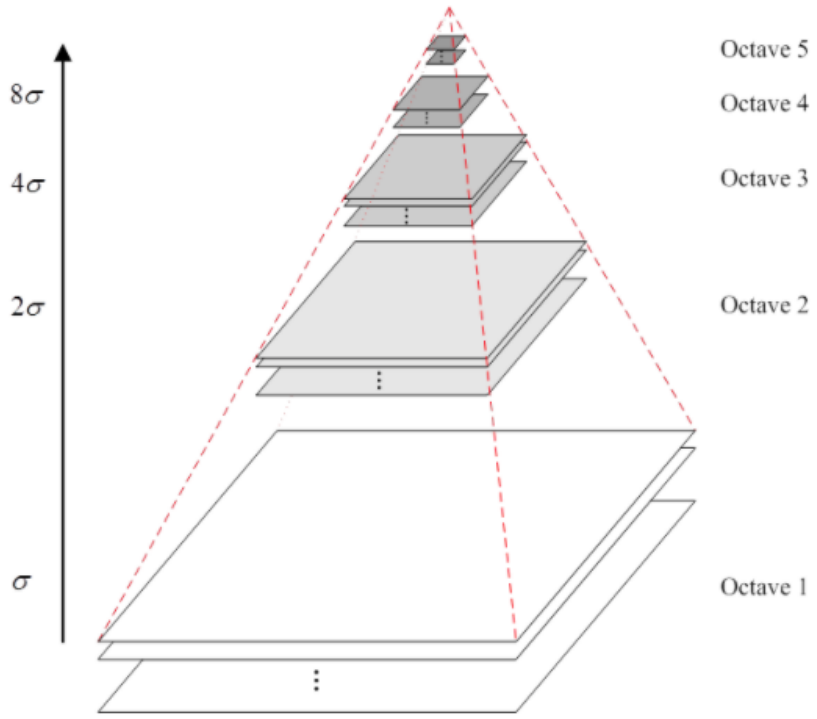

图 2.4 图像金字塔

在 SIFT 算法中，同一个 octave 层中的图片尺寸(即大小)相同，但是尺度(即模糊程度)不同，而不同的 octave 层中的图片尺寸大小也不相同，因为它是由上一层图片降采样得到的。在进行高斯模糊时，SIFT 的高斯模板大小是始终不变的，只是在不同的 octave 之间改变图片的大小。而在 SURF 中，图片的大小是一直不变的，不同的 octave 层得到的待检测图片是改变高斯模糊尺寸大小得到的，当然，同一个 octave 中个的图片用到的高斯模板尺度也不同，如图 2.5 所示。算法允许尺度空间多层图像同时被处理，不需对图像进行二次抽样，从而提高算法性能。

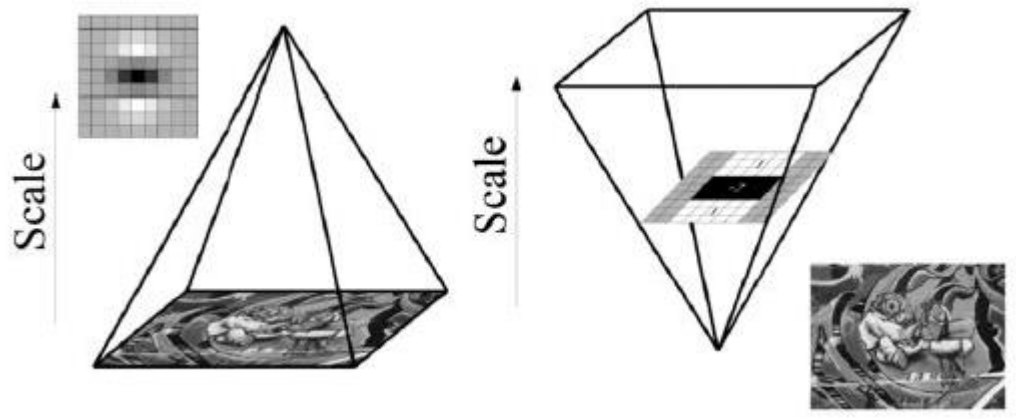

图 2.5 从左至右：传统描述子（SIFT）尺度空间保持 filter 不变，依赖上层结果，改变图像大小；SURF 保持图像大小不变，改变 filter 的尺度大小，提高速度和精度。

2.1.3 精确定位特征点

将经过 hessian 矩阵处理过的每个像素点与其 3 维邻域的 26 个点进行大小比较，如果它是这 26 个点中的最大值或者最小值，则保留下来，当做初步的特征点。检测过程中使用与该尺度层图像解析度相对应大小的滤波器进行检测，以 3×3 的滤波器为例，该尺度层图像中 9 个像素点之一.如下图中检测特征点与自身尺度层中其余 8 个点和在其之上及之下的两个尺度层 9 个点进行比较，共 26 个点，图 2.6 中标记‘x’的像素点的特征值若大于周围像素则可确定该点为该区域的特征点。

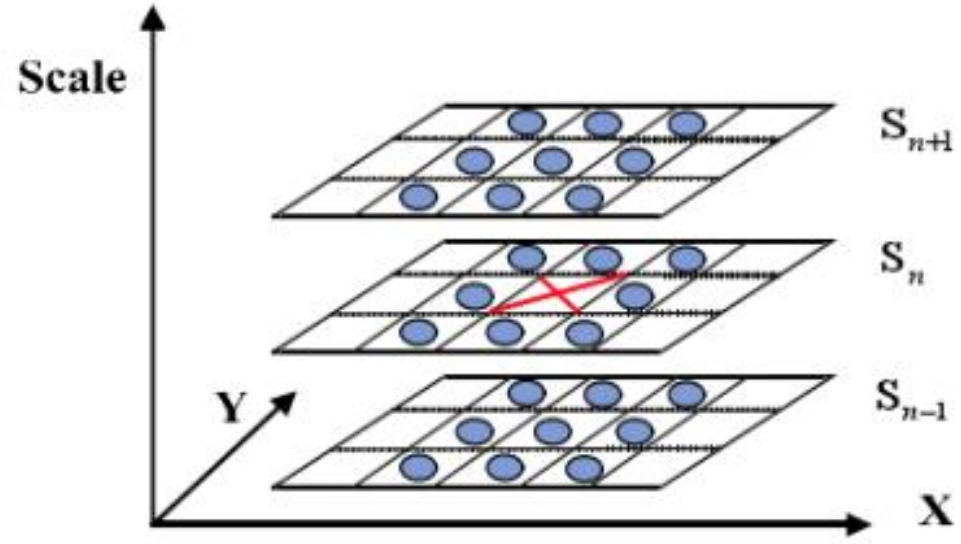

图 2.6 图像金字塔的某一层

然后，采用三维线性插值法得到亚像素级的特征点，同时也去掉那些值小于一定阈值的点，增加极值使检测到的特征点数量减少，最终只有几个特征最强点会被检测出来。

2.1.4 确定主方向

为了保证旋转不变性，在 SURF 中，不统计其梯度直方图，而是统计特征点邻域内的 Harr 小波特征。即以特征点为中心，计算半径为 6s(S 为特征点所在的尺度值)的邻域内，统计 60 度扇形内所有点在 x(水平)和 y(垂直)方向的 Haar 小波[20]响应总和(Haar 小波边长取 4s)，并给这些响应值赋高斯权重系数，使得靠近特征点的响应贡献大，而远离特征点的响应贡献小，然后 60 度范围内的响应相加以形成新的矢量，遍历整个圆形区域，选择最长矢量的方向为该特征点的主方向。这样，通过特征点逐个进行计算，得到每一个特征点的主方向。该过程的示意图如图 2.7：

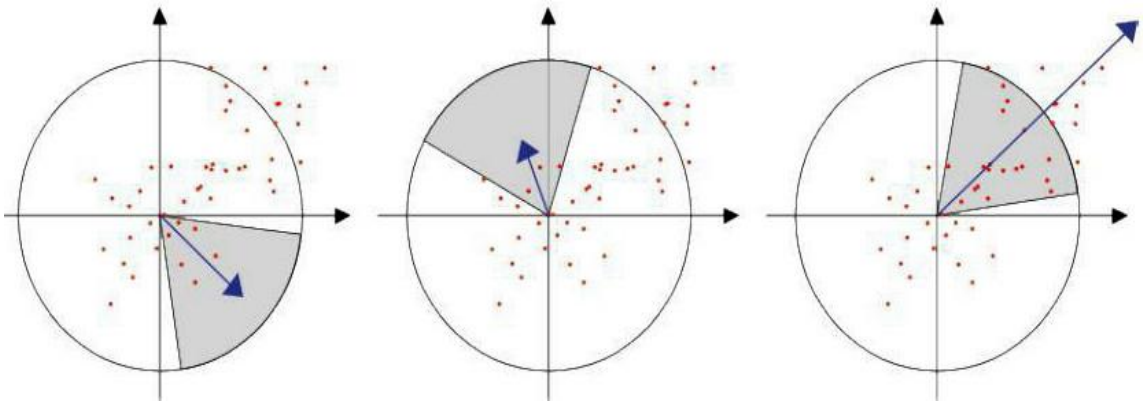

图 2.7 利用 Harr 小波确定主方向过程

2.1.5 构造 SURF 特征点描述子

在 SURF 中，也是在特征点周围取一个正方形框，框的边长为 20s(s 是所检测到该特征点所在的尺度)。该框带方向，然后把该框分为 16 个子区域，每个子区域统计 25 个像素的水平方向和垂直方向的 haar 小波特征，这里的 x(水平)和 y(垂直)方向都是相对主方向而言的。该 haar 小波特征为 x(水平)方向值之和，水平方向绝对值之和，垂直方向之和，垂直方向绝对值之和。该过程的示意图如图 2.8 所示：

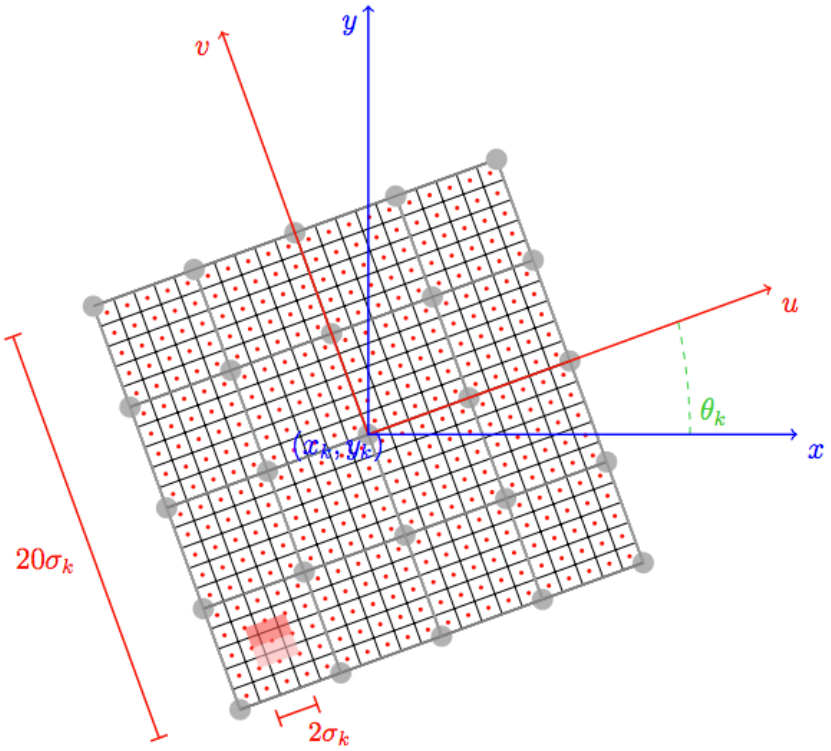

图 2.8 SURF 描述算子示意图。该格栅区域$\mathcal{R}$被分为 16 个子区域，用来建立以兴趣点$X_k:(x_k, y_k, L_k)$为中心，以θ_k为主方向的邻域空间。

与 SIFT 类似，SURF 描述符通过子区域聚合局部梯度信息。然而，SIFT 建立梯度方向的小直方图（通过梯度大小加权），SURF 计算垂直和水平的梯度响应的一阶统计量。 SURF [21]的作者声称使用总和和绝对值之和是紧凑性和效率之间的最佳折中。描述区域$\mathcal{R}_{i,j}$的统计向量为

$$\forall (i,j) \in [\![1,4]\!]^2, \mu_k(i,j) = \begin{pmatrix} \sum_{(u,v)\in\mathcal{R}_{i,j}} d_x(u,v) \\ \sum_{(u,v)\in\mathcal{R}_{i,j}} d_y(u,v) \\ \sum_{(u,v)\in\mathcal{R}_{i,j}} |d_x(u,v)| \\ \sum_{(u,v)\in\mathcal{R}_{i,j}} |d_y(u,v)| \end{pmatrix} \tag{2.7}$$

以兴趣点$X_k: (x_k, y_k, L_k, \theta_k)$为中心的描述子可以简单地通过联结每个子区域计算得到的16个向量$\mu_k(i,j)$得到

$$\mu_k = (\mu_k(i,j))_{1\le i,j\le 4} \tag{2.8}$$

2.2 图像词袋模型（Bag-of-Features）

具有局部特征的图像的描述已被成功应用于几种复杂的图像分析问题，如场景识别和对象分类。这种方法被称为 Bag-of-Features（BoF），它的灵感来自已知的文本检索算法 Bag-of-Words。文本词袋模型需要三步来建立文档。第一步包括解析文件将数据集转换为单词，即将文档分割成较小的组件。图片也同样可以采样到较小的区域（斑块）。通常采用两种抽样策略用于图像词袋模型：稀疏采样或密集采样。稀疏采样检测一组信息性关键点（例如角落）及其各自的支持区域。对于密集采样，则是在图像上建立网格，再在网格中均匀采样。

当然，图像与文本有很大的区别，基于文本的词袋模型已经有了既存的词典不需要通过学习获得，而图像词袋模型中的视觉词典需要通过监督或非监督的学习来获得。因为每个视觉单词都是独立的，不考虑其在图像中的位置，所以需要对图像进行局部特征提取，局部区域特征独立性和稳定性较强，前文提到的 SURF 特征就是一种局部特征提取算法。利用词典中的视觉词典来表示图像中的特征，使得图像更加紧凑统一。

词袋模型主要有两个缺点：一是因为视觉单词独立，所以提取的特征之间没有空间位置联系，会影响检索的准确度；二是从图像的特征向量量化成为视觉词 汇的过程中，降低了视觉词汇间的辨析力，从而产生错误的匹配。

2.2.1 基本原理

Bag-of-Features 模型仿照文本检索领域的 Bag-of-Words 方法，把每幅图像描述为一个局部区域/关键点(Patches/Key Points)特征的无序集合。使用某种聚类算法(如 K-means)将局部特征进行聚类，每个聚类中心被看作是词典中的一个视觉词汇(Visual Word)，相当于文本检索中的词，视觉词汇由聚类中心对应特征形成的码字(code word)来表示（可看当为一种特征量化过程）。所有视觉词汇形成一个视觉词典(Visual Vocabulary)，对应一个码书(code book)，即码字的集合，词典中所含词的个数反映了词典的大小。图像中的每个特征都将被映射到视觉词典的某个词上，这种映射可以通过计算特征间的距离去实现，然后统计每个视觉词的出现与否或次数，图像可描述为一个维数相同的直方图向量，即 Bag-of-Features。

BoF 主要流程如图 2.9：

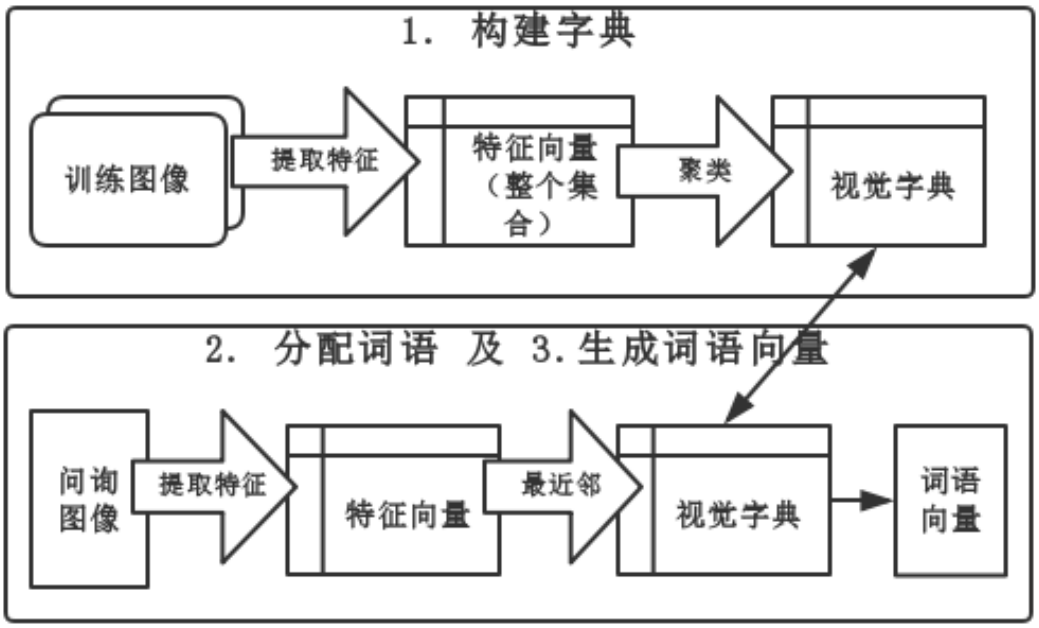

图 2.9 BoF 模型框架

对某张图片提取出的特征可以量化到词典中的某个视觉词汇上，从而该图像可以用词典中的视觉词汇的直方统计图来表示。如图 2.10 所示，我用 3 张图片构建了一个视觉字典，字典中包含 50 个视觉词汇，每张图的特征映射到该字典后得到的直方图就代表了这张图的 bag of features。

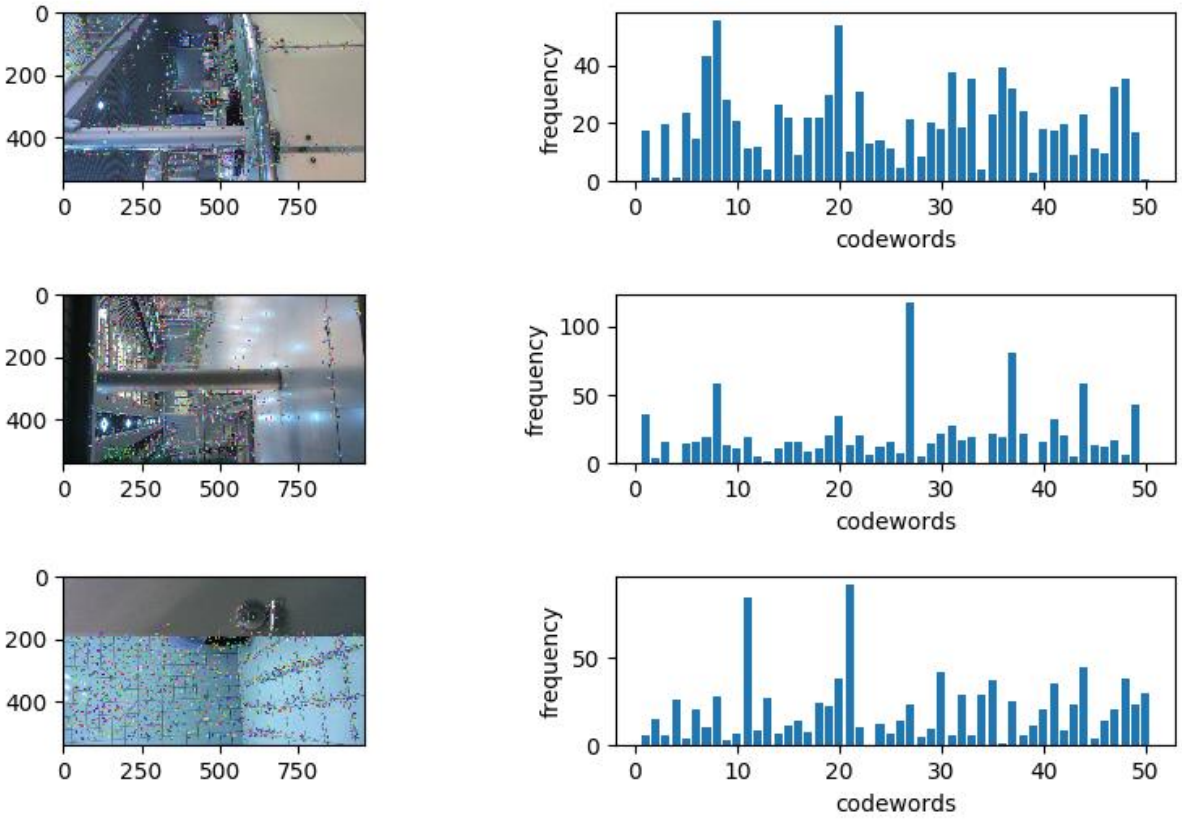

图 2.10 图像与对应的词袋模型直方图（图像上的彩色点为 SURF 特征点）

2.3 分类算法

解决分类问题的方法很多，单一的分类方法主要包括：决策树、贝叶斯、人工神经网络、K-近邻、支持向量机和基于关联规则的分类等；另外还有用于组合单一分类方法的集成学习算法，如 Bagging 和 Boosting 等。本文主要用了支持向量机、逻辑斯蒂回归和贝叶斯，所以着重介绍这三个算法。

2.3.1 支持向量机

支持向量机（SVM，Support Vector Machine）是一种二类分类模型，它最早是由是 Cortes 和 Vapnik 根据[23]统计学习理论提出的。SVM 的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。它的学习策略是根据结构风险最小化准则，使得间隔最大化，可形式化为一个求解凸二次规划的问题。对于分类问题，支持向量机算法根据区域中的样本计算该区域的决策曲面，由此确定该区域中未知样本的类别。

支持向量机学习方法包含由繁至简的模型：线性可分支持向量机、线性支持向量机及非线性支持向量机，简单模型是复杂模型的基础。当训练数据线性可分时，通过硬间隔最大化，学习一

个线性的分类器；当训练数据近似线性可分时，通过软间隔最大化也学习一个线性的分类器；当训练数据线性不可分时，通过使用核技巧及软间隔最大化，学习非线性支持向量机。

2.3.2 逻辑回归(Logistic Regression)

逻辑回归[24]，虽然这个算法从名字上来看，是回归算法，但其实际上是一个分类算法，学术界也叫它 logit regression, maximum-entropy classification (MaxEnt)或者是 the log-linear classifier。它是统计学习中的经典分类方法。

最大熵是概率模型学习的一个准则，将这个准则推广到分类问题得到最大熵模型。逻辑回归模型与最大熵模型都属于对数线性模型。

逻辑回归属于概率性判别式模型，之所谓是概率性模型，是因为 LR 模型是有概率意义的;之所以是判别式模型，是因为 LR 回归并没有对数据的分布进行建模，也就是说，LR 模型并不知道数据的具体分布，而是直接将判别函数，或者说是分类超平面求解了出来。

总而言之，分类算法的手段都是在都是求解 $p(C_k|x)$，即计算一个样本的条件概率 $p(C_k|x)$。逻辑回归是求样本的后验概率，计算结果可以通过贝叶斯公式得到：$p(C_k|x)=p(x|C_k)p(C_k)$，其中 $p(x|C_k)$是类条件概率密度，$p(C_k)$是类的概率先验。使用这种方法的模型，称为是生成模型，即：$p(C_k|x)$是由 $p(x|C_k)$和 $p(C_k)$生成的。分类算法所得到的 $p(C_k|x)$可以将输入空间划分成许多不相交的区域，这些区域之间的分隔面被称为判别函数(也称为分类面)，有了判别函数，就可以进行分类了，上面生成模型，最终也是为了得到判别函数。如果直接对判别函数进行求解，得到判别面，这种方法，就称为判别式法。逻辑回归就属于这种方法。

2.3.3 贝叶斯

贝叶斯（Bayes）分类[25]算法是基于贝叶斯定理与特征条件独立假设的分类方法，如朴素贝叶斯（Naive Bayes）算法。该类算法利用贝叶斯定理来预测样本属于某个类别的概率，选择其中概率最大的一个类别作为该样本的最终类别。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。

由于贝叶斯定理的成立本身需要一个很强的条件独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。为此就出现了许多降低独立性假设的贝叶斯分类算法，如 TAN（Tree Augmented Naive Bayes）算法，它是在贝叶斯网络结构的基础上增加属性对之间的关联来实现的。

3 基于图像的室内定位的实现

室内定位可以为我们的生活带来极大的便利，所以有着非常大的研究潜力。通过上面的叙述，对于室内定位即室内场景分类的研究的重点主要集中在四个方面：

（1）场景图像的特征提取。

（2）视觉字典的聚类生成，并构建通过字典映射得到训练向量。

（3）训练场景分类模型，为最后的预测做准备。

（4）对问询图像进行预测，判断属于哪个场景，进而实现定位。

对于图像特征提取，本文采取了 SIFT 和 SURF 两种算子来提取，然后衡量效果。对于场景图像分类建模，本文也采取了三中分类算法来分别建模，并对结果进行比对衡量。分别是 SVM 支持向量机算法、逻辑回归和贝叶斯。

3.1 室内定位框架概览

室内定位一般有两种方法，一种是提取特征后利用索引技术来匹配问询图像和数据库的图像检索方法，但是这种方法在数据量大的时候效率比较低。另一种就是在提取特征后用监督或非监督学习或深度学习来建立场景分类模型，继而通过预测问询图像的所属场景来实现室内定位。因为采集的数据库图片量较大，有 5000 多张，所以采用第二种方法。室内场景分类由一下子模块构成：

（1）图像特征提取模块。该阶段主要提取图像特征，为之后构建视觉词典和训练模型做准备。

视觉字典构建模块。该阶段的任务是将提取的图像特征聚类生成视觉词典，为之后训练向量的生成做准备。

（2）场景模型训练模块。该模块利用视觉字典得到每张图片的视觉词袋，即训练向量，进而训练场景分类模型。

（3）问询图像预测模块。该模块利用训练好的场景分类模型预测输入的测试图像，实现室内定位。

（4）结果评估模块。选取合适的指标对分类结果进行评估。

具体框架如图 3.1：

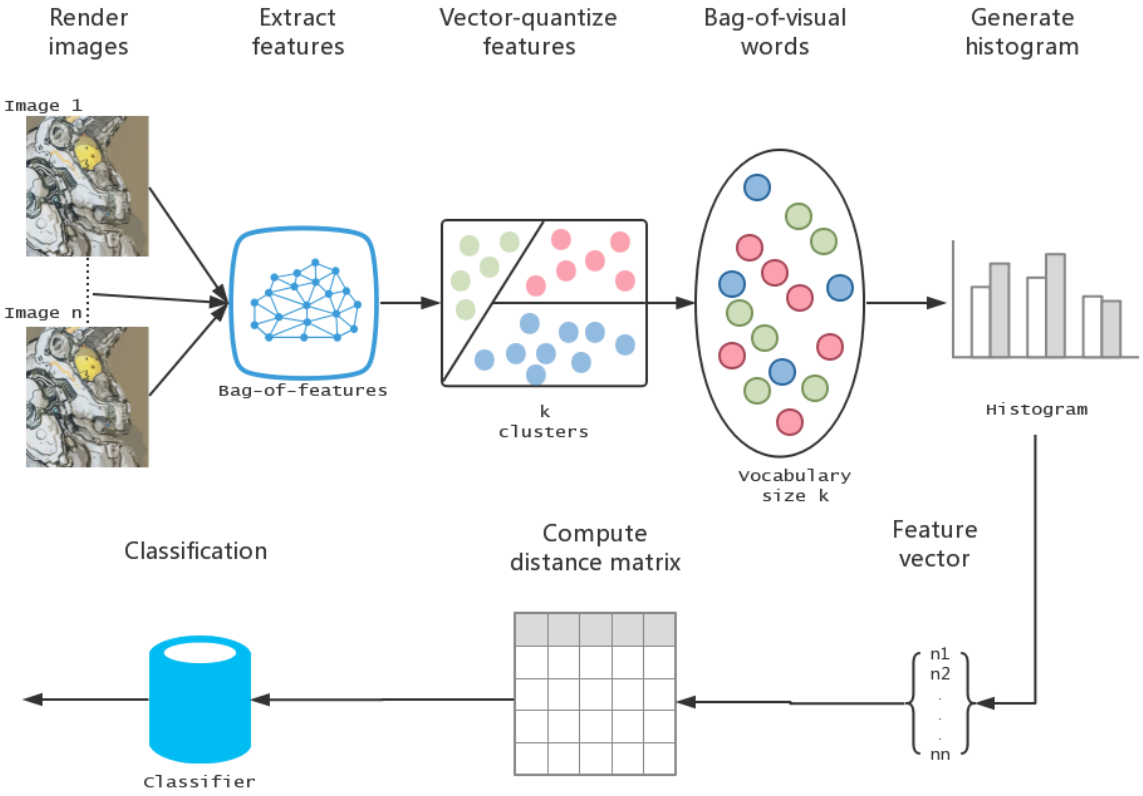

图 3.1 室内定位框架

3.2 场景图像数据整理

3.2.1 采集数据

我以同济大学图书馆一楼整体为数据采集点，在每个独立的空间拍摄了几段视频，将每个独立房间的细节特征务必拍全。同济大学图书馆一楼主要分为以下 12 个场景：

（1） 前门厅
（2） 前门厅二楼
（3） 中央咨询台
（4） 电子阅览室
（5） 二楼天桥
（6） 前门南走廊
（7） 新书阅览室
（8） 书库
（9） 卫生间
（10） 新阅览楼
（11） 阅览楼走廊

装 订 线

（12） 阅览楼二楼

然后将采集的视频进行帧提取。这里运用了 Python opencv 的 VideoCapture：

```
stream=cv2.VideoCapture('lib-1.MOV')
ret, frame=stream.read()
```

代码 3.1 视频提取帧图像（节选）

共提取到 1510 张图片。

3.2.2 整理数据

将得到的图片以所属场景标签命名。先用 SURF 和 SIFT 算子遍历提取所有图片的特征点，如果某张图片没有可以提取的特征，就把这张图片剔除，比如有的图片只有一面白墙，或有的图片因为提取自视频流非常模糊。

清晰数据后，一共得到了 1317 张有效图片，这就是我们整个数据库。然后按照简单交叉验证的模型选择方法，以大概 6：4 的比例构建训练集和测试集，即训练集有 799 张图片，测试集有 518 张图片，

3.3 图像特征提取

本文选取 SURF 算法和 SIFT 算法共同提取图像特征。

SURF 特征描述子用不同大小的盒式滤波器与原始积分图像做卷积，利用积分图，快捷而易于并行。对全部数据集做 SURF 运算，用时平均 150.18s，一共得到 1298759 个特征点，每个特征点有 64 维。

SIFT 算子同样具有尺度不变性，而且非常通用经典，对室内图像的描述表现也很良好。对全部数据做 SIFT 运算，用时平均 250.12s ，一共得到 573947 个特征点，每个特征点有 128 维。

3.3 视觉词汇

视觉词汇是相对于文本词汇的，一篇文章可以由很多单词来构成，同理，一张图片经过处理后，也可以通过视觉词汇来表示。将图像用词袋模型的方式来表示有很大的优越性，提高了目标识别、场景拼接等方面的运算效率。这种方法可以被称为，经过图像特征提取后，通过相应的聚类手段，从提取的特征点里抽取出区别性高、可重复的一部分点构成视觉词汇集合，给学习模型提供高效训练向量。

3.3.1 图像特征聚类

对图像进行 SURF 特征提取后，需要通过聚类的手段将提取到的 SURF 特征进行聚类，再从中提取出显著的视觉词汇。聚类最常用的算法是 K-means 聚类算法[26]，原理是聚类中心m_k和其点x_i之间的欧式距离最短:

$$D(X,M)=\sum_{cluster\ k}\sum_{\substack{point\ i\ in\\ cluster\ k}}(x_i-m_k)^2 \qquad (3.1)$$

K-means 聚类算法的步骤如下:

（1）输入 n 个样本对象，并从中随机选择 k 个样本对象初始化为聚类中心；

（2）循环 3 和 4 直到每个聚类收敛；

（3）根据每个聚类对象的均值(聚类中心)，计算每个对象与所有聚类中心的距离,并将对象重新划分为聚类中心距离最小的类中；

（4）若聚类发生变化，则计算聚类中含有的对象聚酯,更新聚类中心；

另外一种很流行的算法是 Gaussian Mixture Model(GMM)[27]，事实上，GMM 和 k-means 很像，不过 GMM 是学习出一些概率密度函数来，k-means 计算的是将每个数据点被归类到其中某一个聚类中心，而 GMM 计算的是每个数据点属于某个类的概率，又称作 soft assignment。

GMM 的假设非常简单，它的名字高斯混合模型（Gaussian Mixture Model），表示的就是将数据看做服从混合高斯分布。从中心极限定义出发，假设数据服从高斯分布是比较科学合理的。另外，混合模型可以通过增加模型数量来无限逼近某个连续的概率分布，这就增加了算法的灵活性和适用范围。

每个 GMM 由 K 个 Gaussian 分布组成，每个 Gaussian 称为一个“组件”，这些组件线性加成在一起就组成了 GMM 的概率密度函数：

$$p(x)=\sum_{k=1}^{K}p(k)p(x|k)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(x;\mu_k,\sum_k 1\right) \qquad (3.2)$$

其中

$$\mathcal{N}(x;\mu_k,\sum_k 1)=(2\pi)^{\left(-\frac{d}{2}\right)}\left|\sum 1\right|^{-\frac{1}{2}}exp(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\sum 1^{-1}(x-\mu)). \qquad (3.3)$$

相对于之前 K-means 的聚类，因为 GMM 是更偏向密度的，在这里的结果更好一些。所以本文选择 GMM 来对图像特征进行聚类。

在构造视觉词典的过程中，字典大小的选择也是问题，就是聚类数的选择，字典过大，单词缺乏一般性，对噪声敏感，计算量大，关键是图象投影后的维数高；字典太小，单词区分性能差，对相似的目标特征无法表示。

本文对字典维数的选取没有数学方法来计算，只能通过穷举法，本文从 800 到 20000，以 800，1000，1100，1200，1300，1400，1500，1600，1700，1800，1900，10000，15000，20000 的量级来分别进行聚类，再衡量哪一个大小最合适。

3.3.2 特征向量量化

视觉词典构建完后，只是该阶段的第一步，接下来我们就可以将它作为评判一张图像的基础。第一步我们已经提取了每张图片的 SURF 特征向量，如果直接用这些向量去训练模型那么实在是太消耗时间和硬件了，有的图像的特征向量有几千几万个，TLD 数据库的所有 SURF 特征向量超过 53 万条。所以，我们需要提取更加能够代表这个图像的向量和权重，这就需要将每张图像的特征向量映射到视觉词典上，用视觉词汇形成的“词袋”来代表这张图像。

特征的量化一般依据硬量化(Hard quantization)或软量化(Soft quantization)方法[28]。硬量化的

特点是只将特征向量映射到最近邻的聚类中心，这就导致该特征向量将与其他非最近邻中心毫无关系，这就会损失特征描述，进而降低准确率。软量化则是将全部视觉词汇作为基底，将图像的 SURF 特征向量映射到所有的视觉词汇上，只是每个视觉词汇所占的比例不同。故此，本文选择了软量化的方式来将特征向量量化。假设有一个包含 k 个视觉单词的视觉词典，定义一个 k 维向量$\omega = [\omega_1, \cdots, \omega_t, \cdots, \omega_k]$，其中$\omega_t$代表某个视觉词汇 t 在一图片中所占权重。$\omega_t$的公式为：

$$\omega_t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j,t) \tag{3.4}$$

其中，M_i表示第 i 个最近邻视觉词汇的特征点数量，方法$sim(j,t)$为特征 j，t 之间的余弦相似度。

余弦相似度表示的是两个向量之间余弦值的大小比较来近似描述向量的相似度。0°角的余弦值是 1，而其他任何角度的余弦都不大于 1；并且其最小的值是-1。从而两个向量之间的角的余弦值确定两个向量是否大致指向相同的方向。无论一个向量的的模如何变化，他们之间的夹角都是不变的。所以，在应用余弦相似度之前要将向量的每一个维度归一化。其计算公式如下:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{3.5}$$

经过量化后，就可以将一张图像中所有的特征向量映射到多个相近的视觉词汇上，这样，这张图像就可以用一个 k 个柱的统计直方图来表示，每个柱代表一个视觉词汇在该图像中的统计数量，其中 k 是视觉词典中的视觉词汇数量。

3.4 训练模型

SVM 模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

除了进行线性分类之外，SVM 还可以使用所谓的核技巧有效地进行非线性分类，将其输入隐式映射到高维特征空间中。

3.4.1 选取合适的核模型

运用非线性 SVM 分类器是因为量化后的图像特征是无法用一维超平面分类的，所以可以用更高维的特征空间来分割训练集。不用直接计算特征空间变换方程$\varphi(x)$，我们可以定义一个核函数 K:

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j). \tag{3.6}$$

该函数为原始特征空间提供了非线性决策的边界：

$$\sum_{i} \alpha_i y_i K(x_i, x) + b. \tag{3.7}$$

接下来，因为在 BOF 模型中我们用了直方图来表示图像的特征向量，所以这里用直方图交叉核(Histogram intersection kernel)[29]

$$I(h_1, h_2) = \sum_{i=1}^{N} \min\big(h_1(i), h_2(i)\big), \tag{3.8}$$

用来对特征构成的直方图进行相似度匹配，下面介绍下原理。

假设图像或其他数据的特征可以构成直方图，根据直方图间距的不同可以得到多种类型的直方图：

$$\Psi(x) = [H_{-1}(x), H_0(x), \cdots, H_L(x)], \tag{3.9}$$

假设 H0(x)里每个直方图宽度为 a，那么 H1(x)为 2a，以此类推。

两个数据集的相似度可以用下式来匹配：

$$K_\Delta(\Psi(y), \Psi(z)) = \sum_{i=0}^{L} \omega_i N_i, \tag{3.10}$$

y 和 z 分别代表不同的数据集。其中 w 代表权重，将 wi 设置为 1/(2^i)，N 代表每两层之间的新匹配的数目，可以通过下式计算：

$$N_i = I(H_i(y), H_i(z)) - I(H_{i-1}(y), H_{i-1}(z)), \tag{3.11}$$

上式里面的 I 可以通过下式计算：

$$I(h_1, h_2) = \sum_{i=1}^{N} \min\left(h_1(i), h_2(i)\right). \tag{3.12}$$

即核函数。

原理如图 3.2：

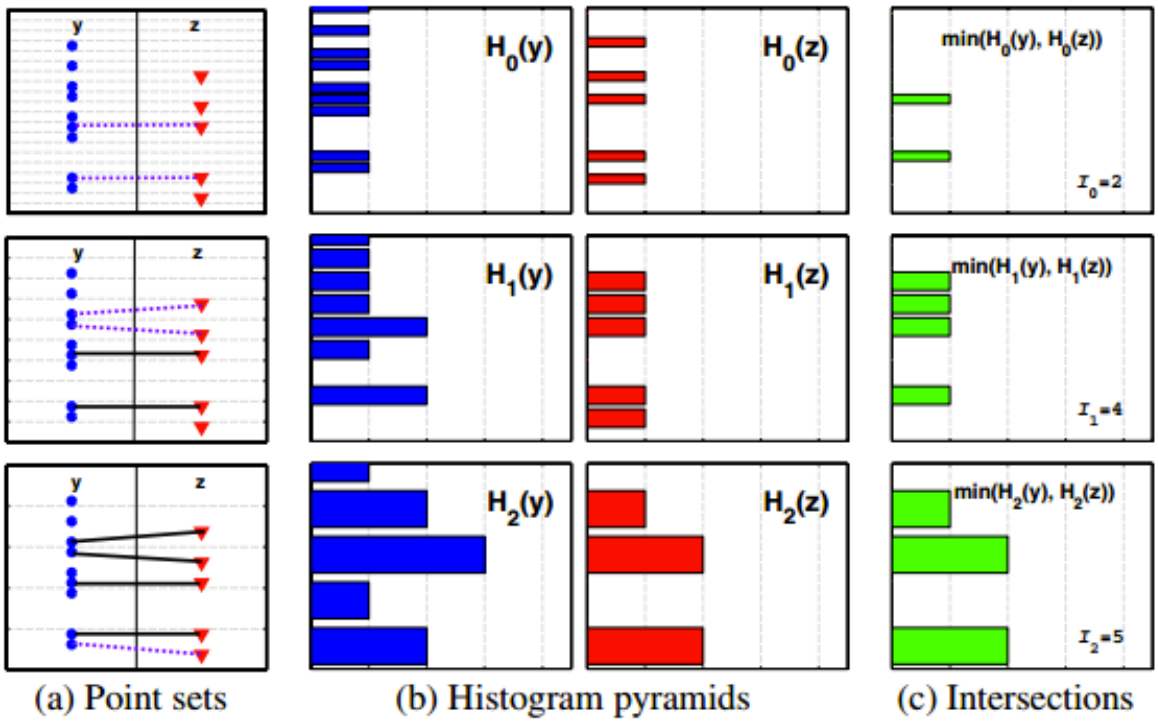

图 3.2：(a)里的 y 和 z 代表两种数据分布，三幅图代表三层金字塔，每一层里有间距相等的虚线。可以看到红点蓝点的位置是固定的，但是根据直方图宽度的不同可以划到不同的直方图里，如(b)所示。(c)图就是 I 的计算结果，

是通过(b)里两种直方图取交集得来的，不过直方图的高度忽略不计，只计算交集后的数目，(c)图每个图的下方都给出了交集数目，比如$x_0 = 2, x_2 = 4, x_3 = 3$。

I 得到了,通过 Ni=Li-Li-1 得到 N。由于 wi 之前设置为 1/(2^i)了，所以

$$K_\Delta(\Psi(y),\Psi(z)) = \sum_{i=0}^{L} \frac{1}{2^i}\Big(I\big(H_i(y),H_i(z)\big) - I\big(H_{i-1}(y),H_{i-1}(z)\big)\Big). \quad (3.13)$$

其它的常用核还有 Generalized Gaussian kernel：

$$K(h_1,h_2) = exp(-\frac{1}{A}D(h_1,h_2)^2), \quad (3.14)$$

其中 D 可以是欧氏距离，也可以是χ^2距离

$$D(h_1,h_2) = \sum_{i=1}^{N} \frac{\big(h_1(i)-h_2(i)\big)^2}{h_1(i)+h_2(i)}。 \quad (3.15)$$

3.4.2 多类 SVM

众所周知，SVM 是二类分类器，但是我们的场景分类模型需要分 12 个类，这就需要将 SVM 分类器改造。有很多种方法和公式来实现 SVM 多类分类器，但是实际最常用的还是通过多种方法组合多个二类 SVM 分类器，常见的方法有 one-against-one 和 one-against-all 两种[30]。

（1）一对多法（one-versus-rest,简称 OVR SVMs）

训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类，这样 k 个类别的样本就构造出了 k 个 SVM。分类时将未知样本分类为具有最大分类函数值的那类。

本文有 12 类要划分（也就是 12 个 Label），他们是 A、B、C、D、E、F、G、H、I、J、K、L。于是在抽取训练集的时候，分别抽取

① A 所对应的向量作为正集，B，C，D，…，L 所对应的向量作为负集；

② B 所对应的向量作为正集，A，C，D，…，L 所对应的向量作为负集；

③ ．．．

④ L 所对应的向量作为正集，A，B，C，…，K 所对应的向量作为负集；

使用这 12 个训练集分别进行训练，然后的得到 12 个训练结果文件。

在测试的时候，把对应的测试向量分别利用这 12 个训练结果文件进行测试。

最后每个测试都有一个结果 f1(x),f2(x),…,f12(x)。于是最终的结果便是这 12 个值中最大的一个作为分类结果。

评价：这种方法有种缺陷,因为训练集是 1:M，这种情况下存在 biased。因而不是很实用。可以在抽取数据集的时候，从完整的负集中再抽取三分之一作为训练负集。

（2）一对一法（one-versus-one,简称 OVO SVMs 或者 pairwise）

其做法是在任意两类样本之间设计一个 SVM，因此 k 个类别的样本就需要设计 k(k-1)/2 个 SVM。

当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。

Libsvm 中的多类分类就是根据这个方法实现的。

本文有 12 类。同上，在训练的时候我选择 A,B; A,C; …;A,L; B,C; B,D;…;B,L;…;K,L 所对应的向量作为训练集，然后得到六 66 个训练结果.

在测试的时候，把对应的向量分别对六个结果进行测试，然后采取投票形式，最后得到一组结果。

投票是这样的：

① A=B=C=…=L=0;

② (A,B)-classifier 如果是 A win,则 A=A+1; otherwise, B=B+1;

③ (A,C)-classifier 如果是 A win,则 A=A+1; otherwise, C=C+1;

④ •••

⑤ (K,L)-classifier 如果是 K win,则 K=K+1; otherwise, L=L+1;

⑥ The decision is the Max(A,B,C,…,L)

本文选取这种方法。

评价：这种方法虽然好,但是当类别很多的时候，model 的个数是 n*(n-1)/2,代价还是相当大的。

3.5 使用模型预测

在 SVM 训练得到场景分类模型后，可以说数据的预处理阶段就结束了，然后就是应用部分。利用得到的场景分类模型，当用户输入新的问询图像后，后台通过 SURF 特征提取算法提取图像特征，然后用之前得到的视觉词典进行特征量化，得到场景分类模型的输入向量，然后模型就可以预测出该图像属于哪个场景，或者说属于哪个场景的概率最大。

总的来说，就是重复了之前的步骤，只是去掉了模型训练部分，直接利用模型进行预测。

4 仿真与实现

4.1 实验环境

4.1.1 硬件环境

表 4.1

处理器	2.9 GHz Intel Core i7
内存	8 GB 1600 MHz DDR3
显卡	Intel HD Graphics 4000 1024 MB

4.1.2 软件环境

表 4.2

操作系统	OS X Yosemite/Windows 8
Python 环境	Python 3.6.1
Opencv 环境	Opencv3.4
开发平台	Sublime/Eclipse

4.2 实验数据集

试验中用到 2 种实验数据集来测试整个场景分类模型效果及各个实验性能。

（1）ImageCLEF/LifeCLEF Dataser（简称 ICD）[31]：这是来自于 ImageCLEF 图像标注 2014 年一个比赛的数据集。该比赛是用在机器人视觉方面的。图像序列采集自 COLD-Stockholm database。该序列通过 MobileRobots PowerBot 机器人平台上安装的两台 Prosilica GC1380C 相机捕获。一共 955 张图片，训练集 810 张图片，测试集 145 张图片，一共有 9 个类别：

训练集：

① Corridor

② Kitchen

③ LargeOffice

④ MeetingRoom

⑤ LargeMeetingRoom

⑥ PrinterArea

⑦ RecycleArea

⑧ SmallOffice

⑨ Toilet

测试集：

① Corridor

② Kitchen

③ LargeOffice

④ MeetingRoom

⑤ LargeMeetingRoom

⑥ PrinterArea

⑦ RecycleArea

⑧ SmallOffice

⑨ Toilet

（2）Tongji Library Datasets（简称 TLD）：本文自己建立的同济大学图书馆一楼室内场景图库。该数据库包含了 1317 张图片，训练集 899 张，测试集 518 张，共 12 个场景：

① 前门厅
② 前门厅二楼
③ 中央咨询台
④ 电子阅览室
⑤ 二楼天桥
⑥ 前门南走廊
⑦ 新书阅览室
⑧ 书库
⑨ 卫生间
⑩ 新阅览楼
⑪ 阅览楼走廊
⑫ 阅览楼二楼

4.3 图像特征提取算法及性能评估实验

4.3.1 实验方法

为了验证 SURF 描述子算法在本文的应用优于 SIFT 算法，通过提取时间和模型分类效果两个指标 来衡量两个算子的性能表现。相同数量的图片，提取时间平均最短以及分类准确率最高的算法最优。

主要工作：

（1）用两种算子提取图像特征。

（2）记录每种算子完成提取的时间。

（3）通过建立好的模型得到每种算子的分类效果。

4.3.2 实验过程和结果

在这一阶段，针对不同的图像变换，使用前文提到的 ICD 和 TLD 数据库进行测试。

（1）图像特征点提取效果实验：为了检比对图像在两种特征点描述子的作用下特征点提取的情况，对 TLD 数据库中的数据分别进行两种算子的运算，特征点提取效果如图 4.1：

[a] [b]

图 4.1： [a]是 SURF 算子对图像的提取效果，[b]是 SIFT 算子对图像的提取效果。图中彩色圆圈表示提取的特征点。

可以明显看出，SURF 特征点描述子提取的特征点数量明显多于 SIFT 特征点描述子提取的特征点数量。那么相应的说明 SURF 对图像空间的描述比 SIFT 算法更丰富全面。

（2）图像特征提取效率实验：为测试 SURF 特征点描述子的计算速度快于 SIFT 特征点描述子，本文用两种特征点描述子分别对 ICD 数据库可 TLD 数据库的图像进行特征提取，记录运行时间，如图 4.2 和图 4.3 所示，在两个数据集的测试下，SURF 特征点描述子的运算时间明显少于 SIFT 特征点描述子：

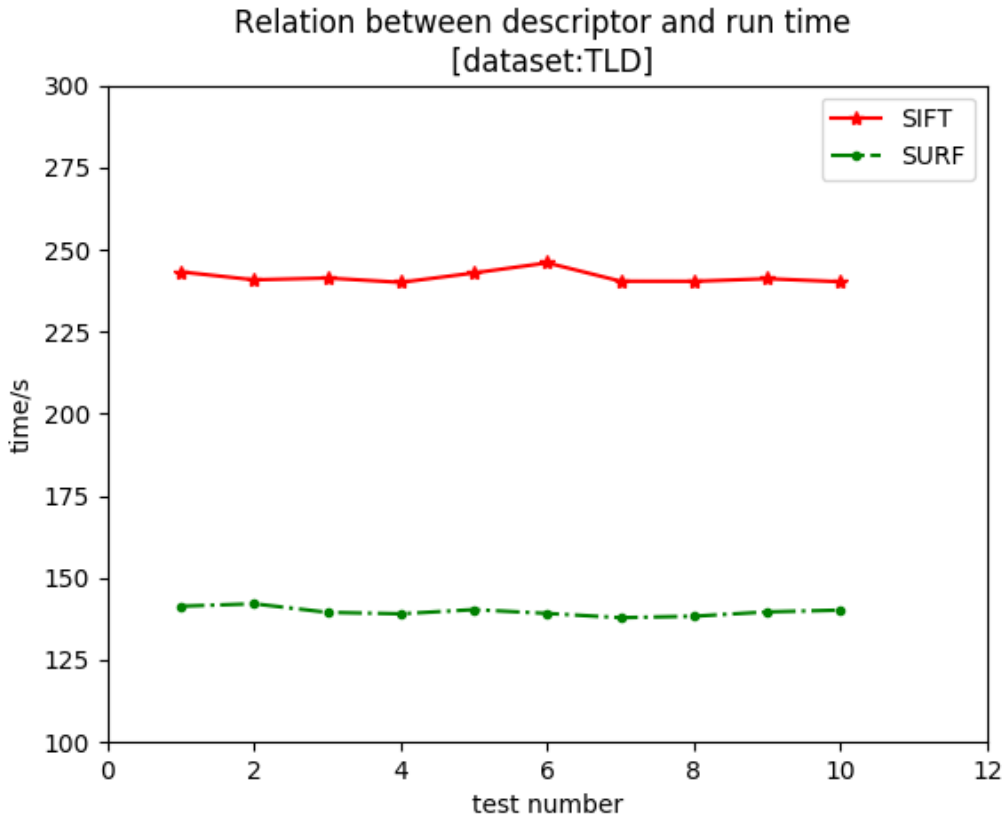

图 4.2　SIFT 与 SURF 算子在 10 次试验下分别运行的时间（TLD 数据库）

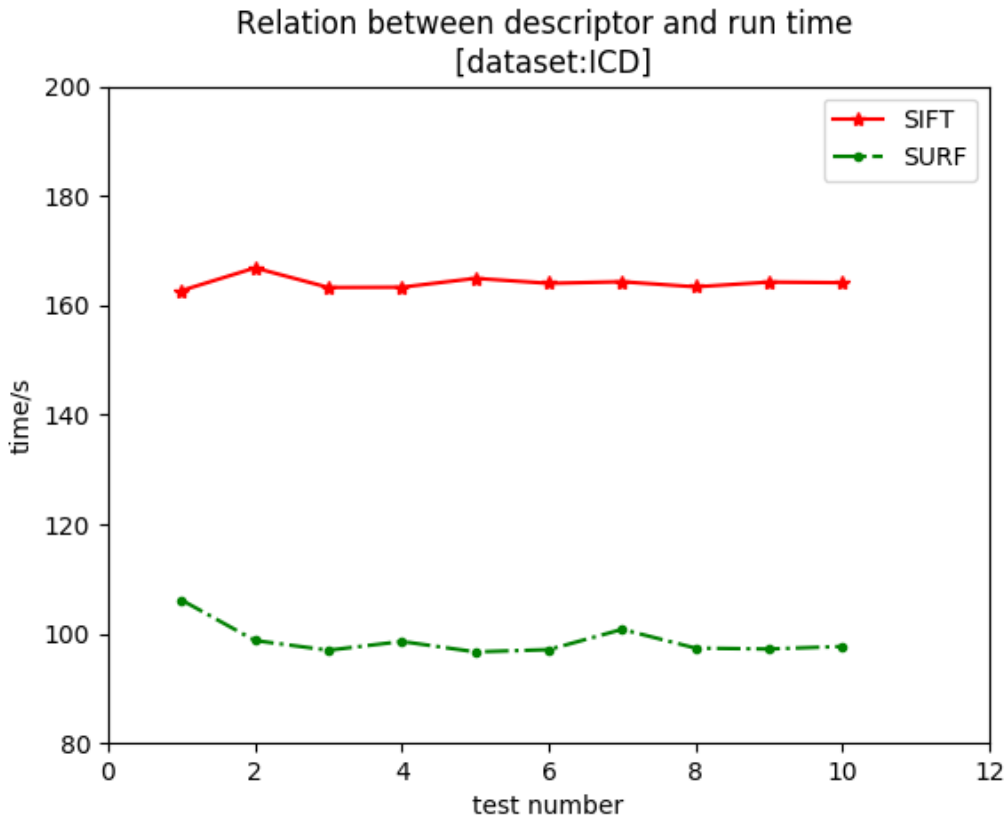

图 4.3 SIFT 与 SURF 算子在 10 次试验下分别运行的时间（ICD 数据库）

具体数据如下表所示：

表 4.3 数据集 TLD 计算结果

	SIFT	SURF	倍率
1	243.20373359	141.3975212858	1.7199999786
2	240.7801319414	142.1449950919	1.6939050987
3	241.3324973243	139.525804681	1.7296621071
4	240.0113347722	139.0882009411	1.7256052861
5	242.8807149581	140.357494666	1.7304435045
6	245.945679801	139.1990045074	1.7668637838
7	240.3560671966	137.9077757054	1.7428753815
8	240.3343928164	138.3521165097	1.7371211867
9	241.1239923655	139.6726427151	1.7263508993
10	240.1952418508	140.2049968408	1.7131717647
Average	241.6163786616	139.7850552944	1.7284850527

表 4.4 数据集 ICD 计算结果

	SIFT	SURF	倍率
1	162.6887353081	106.1477280985	1.5326633761
2	166.8339882231	98.7502322666	1.6894541349
3	163.2582303239	96.9801336657	1.6834193164
4	163.3079831955	98.5804656334	1.6565957783
5	164.9215688287	96.6791289358	1.7058652746
6	164.0139754058	97.0831906343	1.6894168222
7	164.3015413139	100.7933757537	1.630082732
8	163.3747556736	97.3642366569	1.6779750069
9	164.2382309725	97.2069982539	1.6895720876
10	164.1321963117	97.66955007	1.6804848204
Average	164.1071205557	98.7255039969	1.6622565995

结果显示在本文的工作环境下，SURF 算子对图像特征提取的效率比 SIFT 算子快约 1.7 倍。

（3）图像匹配率实验：为测试 SURF 特征点描述子提取的特征向量和 SIFT 特征点描述子提取的特征向量在进行模型训练后得到的模型分类器分类效果，本文对同济大学图书馆的数据集进行了测试，其中还涉及一个自变量是视觉词典的大小，即聚类的数目，实验结果如图 4.4：

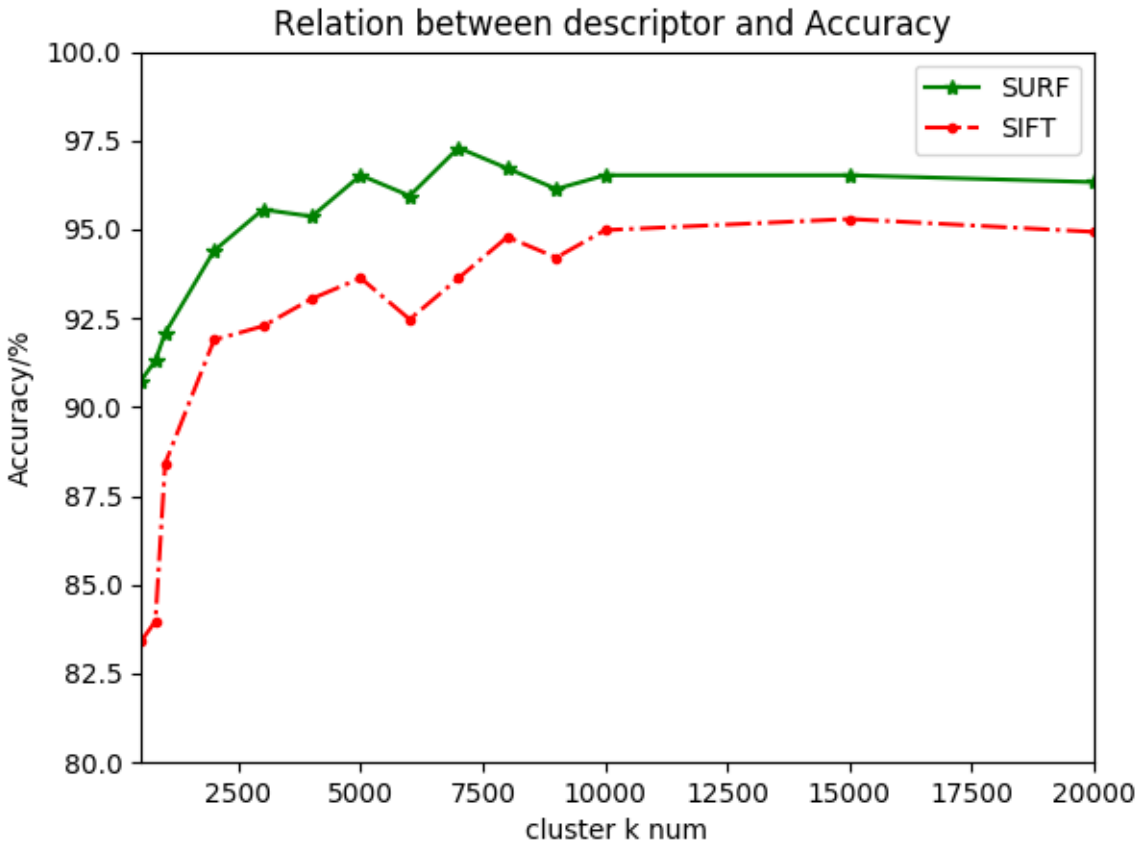

图 4.4 SURF 与 SIFT 算子在不同聚类数条件下的准确率

具体数据如下表：

表 4.5 不同聚类数下两种算子的准确率

聚类数 k	SURF	SIFT
500	0.907	0.834
800	0.913	0.840
1000	0.921	0.884
2000	0.944	0.919
3000	0.956	0.923
4000	0.953	0.931
5000	0.965	0.936
6000	0.960	0.925
7000	0.973	0.936
8000	0.967	0.948
9000	0.961	0.942

10000	0.965	0.950
15000	0.965	0.953
20000	0.963	0.949

可以明显看出，SURF 特征描述子训练的分类器准确率比 SIFT 特征描述算子的准确率高很多。

4.3.3 实验总结

通过实验结果发现，SURF 算子在

（1）图像特征点提取；

（2）计算时间效率；

（3）模型准确率；

的结果都优于 SIFT 特征描述子。而（3）的结果一定程度上来源于（1）的结果，即 SURF 对场景图像的描述更加丰富准确，细节更多，导致训练模型时可以更准确地描述一个场景，提高场景分类准确率。以内准确率的优先级在本文最高，另外两个条件 SURF 算子也均表现优于 SIFT 算子，所以本文选择的 SURF 特征描述子是正确的策略。

4.4 机器学习分类算法的性能评估实验

4.4.1 实验方法

为了比较 SVM 支持向量机、逻辑回归和贝叶斯分类器的分类效果，本文通过对两种数据集用 SURF 算子提取特征向量，然后分别进行三种分类器的训练。

4.4.2 实验过程和结果

通过对两种数据集分别进行试验，得到了准确率关于视觉词典大小的关系如图 4.5 和图 4.6：

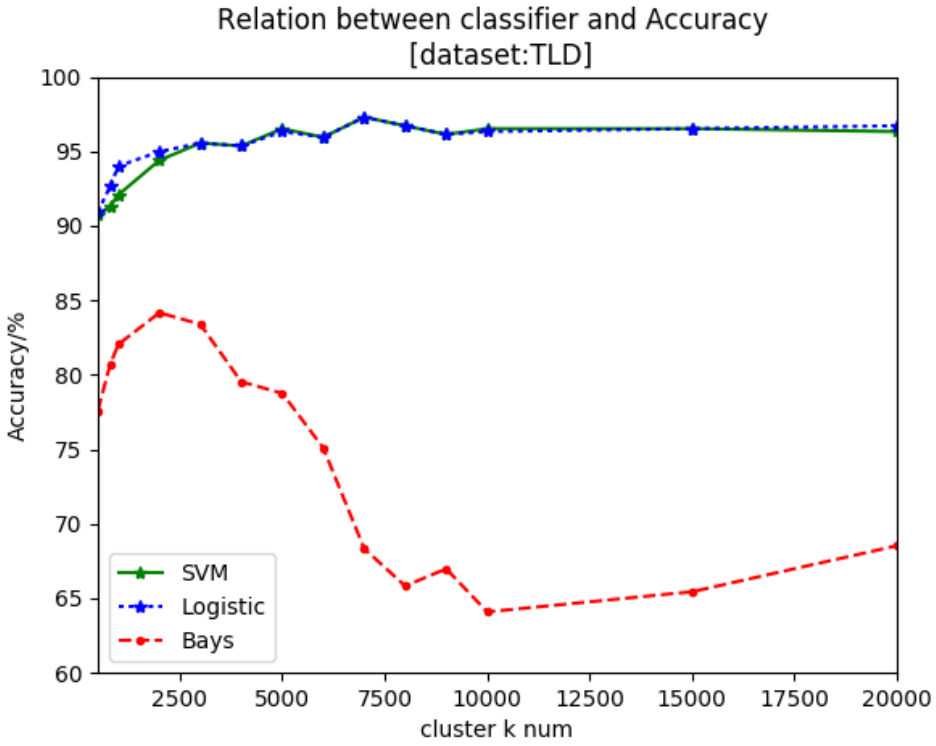

图 4.5 不同分类方法的准确率（TLD 数据库）

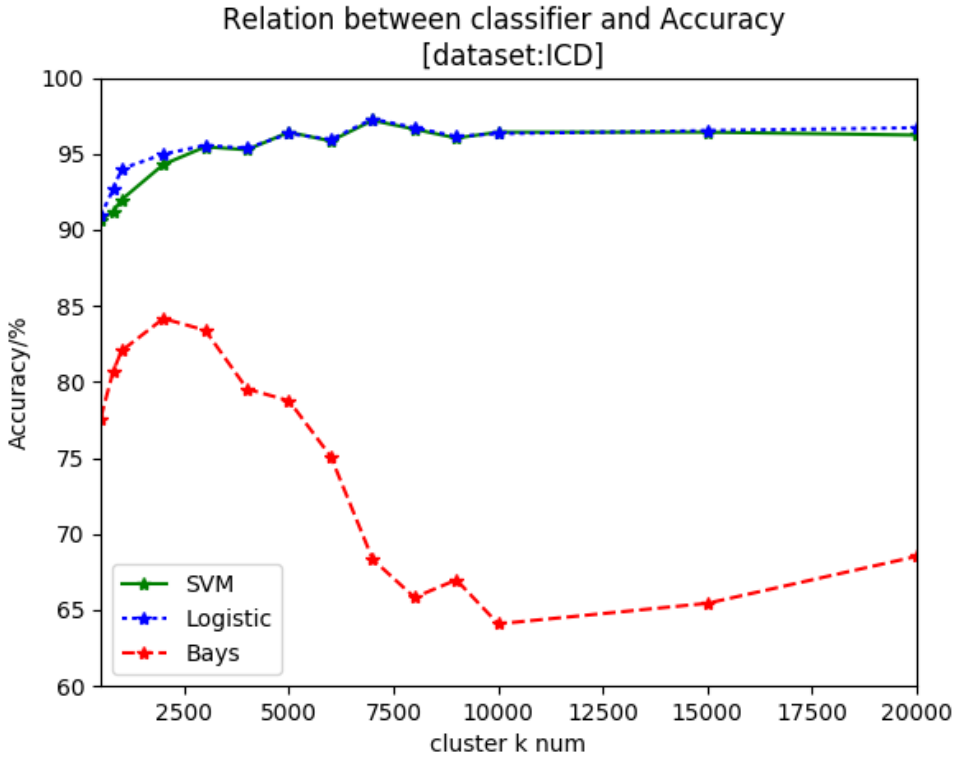

图 4.6 不同分类方法的准确率（ICD 数据库）

装 订 线

具体数据如下表：

表 4.6 两种数据集下不同分类器的准确率

	Dataset TLD			Dataset ICD		
	SVM	Logistic	Bayes	SVM	Logistic	Bayes
800	0.90733591	0.90926641	0.77606178	0.9436225	0.9295664	0.796062
1000	0.91312741	0.92664093	0.80694981	0.9661733	0.9469403	0.826950
2000	0.92084942	0.94015444	0.82046332	0.9723455	0.9604544	0.840463
3000	0.94401544	0.94980695	0.84169884	0.9841342	0.9701695	0.861699
4000	0.95559846	0.95559846	0.83397683	0.9723552	0.9859846	0.853977
5000	0.95366795	0.95366795	0.7953668	0.9923553	0.9996795	0.815367
6000	0.96525097	0.96332046	0.78764479	0.9920352	0.9836246	0.807645
7000	0.95945946	0.95945946	0.75096525	0.9912631	0.9897946	0.770965
8000	0.97297297	0.97297297	0.68339768	0.9844235	0.9932297	0.703398
9000	0.96718147	0.96718147	0.65830116	0.9952713	0.9878147	0.678301
10000	0.96138996	0.96138996	0.66988417	0.9952239	0.9868996	0.689884
15000	0.96525097	0.96332046	0.64092664	0.9917423	0.9962046	0.660927
20000	0.96525097	0.96525097	0.65444015	0.9956242	0.9955097	0.674440

可以看出，SVM 支持向量机的分类效果与逻辑回归的分类效果相近，而贝叶斯分类器分类效果较差，最高只达到了 86%，而且并不稳定。另外，因为 ICD 数据库的场景图像采集的比作者自己采集的 TLD 数据库更加清晰科学，所以分类器的效果有显著地提升。

4.4.3 实验总结

经过实验，可以得出 SVM 支持向量机的分类效果与逻辑回归的分类效果基本相同，在数据更规范的前提下 SVM 的表现略优于逻辑回归。另外在以上所有试验比对中，涉及到的视觉词典大小可以得出，在聚类数 k=8000，即视觉词典中包含 8000 个视觉词汇的时候，分类准确率和分类稳定性已经达到要求。

4.5 演示

本文提出的基于视觉的高效定位系统主要由三部分组成：数据预处理部分、服务器部分和客户端部分。数据预处理阶段将事先采集好的数据库训练成场景分类模型，主要通过上文所述的图像特征提取，构建视觉词典，训练场景分类器；服务器端完成问询图像输入场景分类器，同样经过图像特征提取和构建视觉词汇包的部分，然后分类器给出位置信息；客户端主要工作为通过浏览器给服务器上传问询图像并接受服务器的运算结果。主要结构如图 4.7。

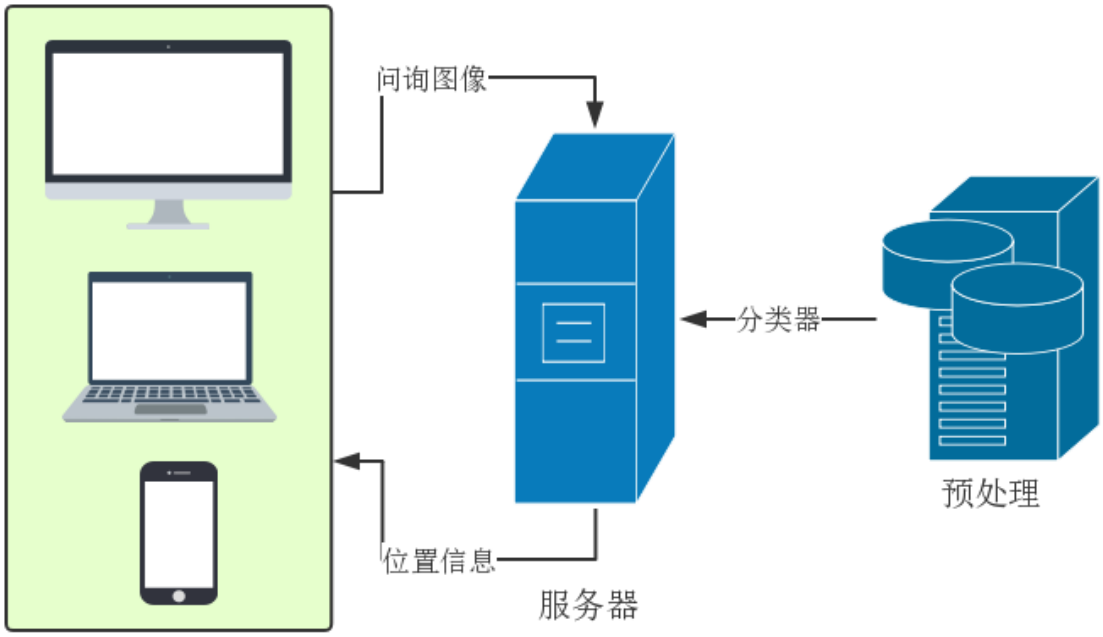

图 4.7 系统框架

4.5.1 数据预处理

该部分用了 Python+openCV 来处理图像数据。Python 具有丰富和强大的库。openCV[32]有可以为 Python 调用的接口。Python 的库很丰富，Numpy/Matplotlib/pandas 为数据处理提供了简洁强力的解决方案，而 Scipy/scikit-learn 给开发者提供了丰富强大的机器学习模块，操作简单，使数据挖掘和数据分析高效快捷，而且无访问限制。所以我们只要提供相应的参数，就可以训练出我们需要的模型，而不用自己完全实现

该系统的应用是为同济大学图书馆内的用户服务，所以只用了 TLD 数据库来实现该系统。数据预处理部分将 TLD 数据库的 12 个场景的图像进行特征提取，然后把特征向量聚类构建视觉词典，然后再将所有图像的特征向量量化到视觉词典上，得到每张图片的视觉词袋，输入 SVM 多类分类器进行训练，得到场景分类模型。

4.5.2 服务器端处理

服务器端采用了跨平台的设计，运用了 openCV 计算机视觉库和 Django 网络框架。

OpenCV 是一个开源的跨平台计算机视觉库，可以运行在 Linux、Windows、Android 和 Mac OS 操作系统上。它轻量级而且高效——由一系列 C 函数和少量 C++ 类构成，同时提供了 Python、Ruby、MATLAB 等语言的接口，实现了图像处理和计算机视觉方面的很多通用算法[32]。

Django[33]是一个高级的 Python Web 框架，它鼓励快速开发和清洁，务实的设计。 由经验丰

富的开发人员构建，它负责 Web 开发的许多麻烦，因此开发者可以专注于编写应用程序，而无需重新创建底层协议。 它是免费的和开源的。Django 框架的的特点有：1、构建快速； 2、安全可靠；3、自由的延展性。

服务器的主要工作有对问询图像预处理和位置信息生成。

对问询图像的预处理包括之前所说的提取 SURF 特征向量、聚类并量化到之前生成的视觉词典上，生成该图像的词袋模型。可以说词袋模型是本文的核心之一。

生成位置信息就是讲问询图像的词袋向量输入之前训练好的 SVM 场景分类器，然后就可以得到该图像所属场景类别。

4.5.3 客户端设计

客户端是浏览器的实现，采用了 HTML5 技术，和 Django web 框架结合。

HTML5 的设计目的是为了在移动设备上支持多媒体，比如 video 和我们需要的图像[34]。

客户端的主要工作是上传问询图像到服务器以及将服务器反馈的位置信息展现在浏览器上。图像可以并行上传多张，因为考虑到提高识别的准确度，用户可以上传多张某个场景的图像，选取得票数最多的那个场景类别作为用户的位置。

客户端运行过程如图所示：

（1）选择图片

点击绿色按钮添加图片文件。可以一次添加多张图片。点击黄色按钮可以取消选择的文件，如图 4.8。

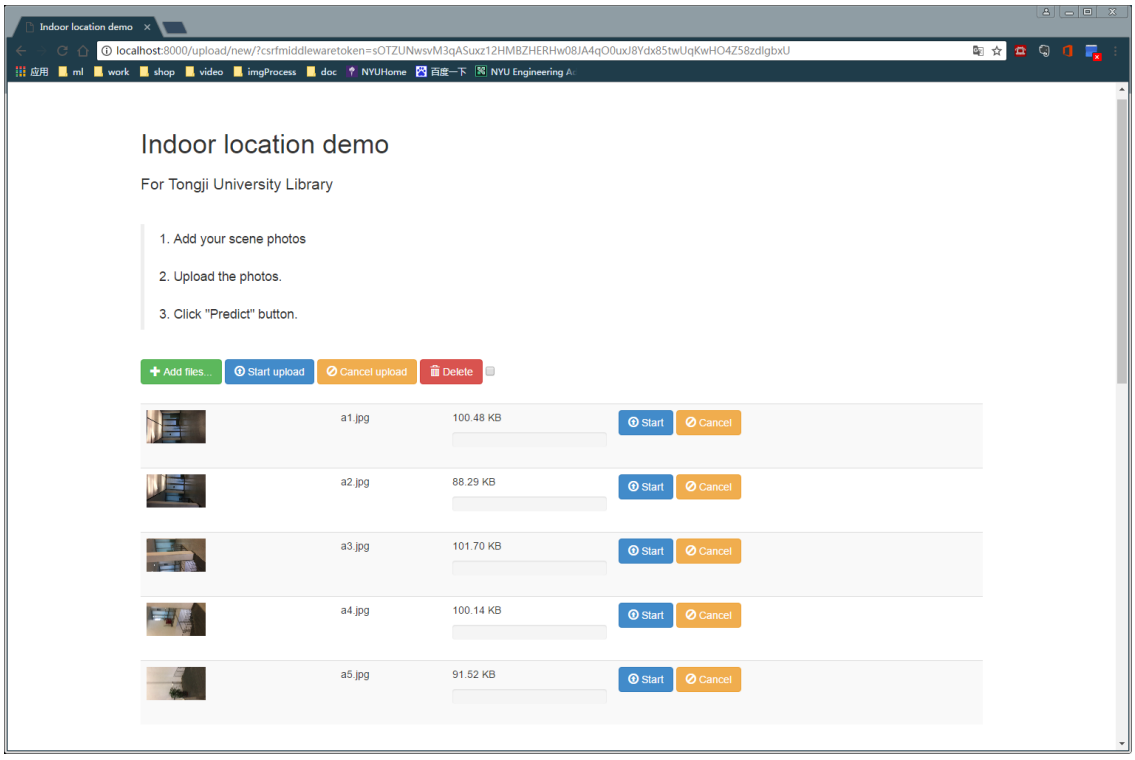

图 4.8 选择图片

（2）上传图片

点击蓝色按钮上传图片至服务器。点击红色按钮可以删除已经上传到服务器的图片，如图 4.9。

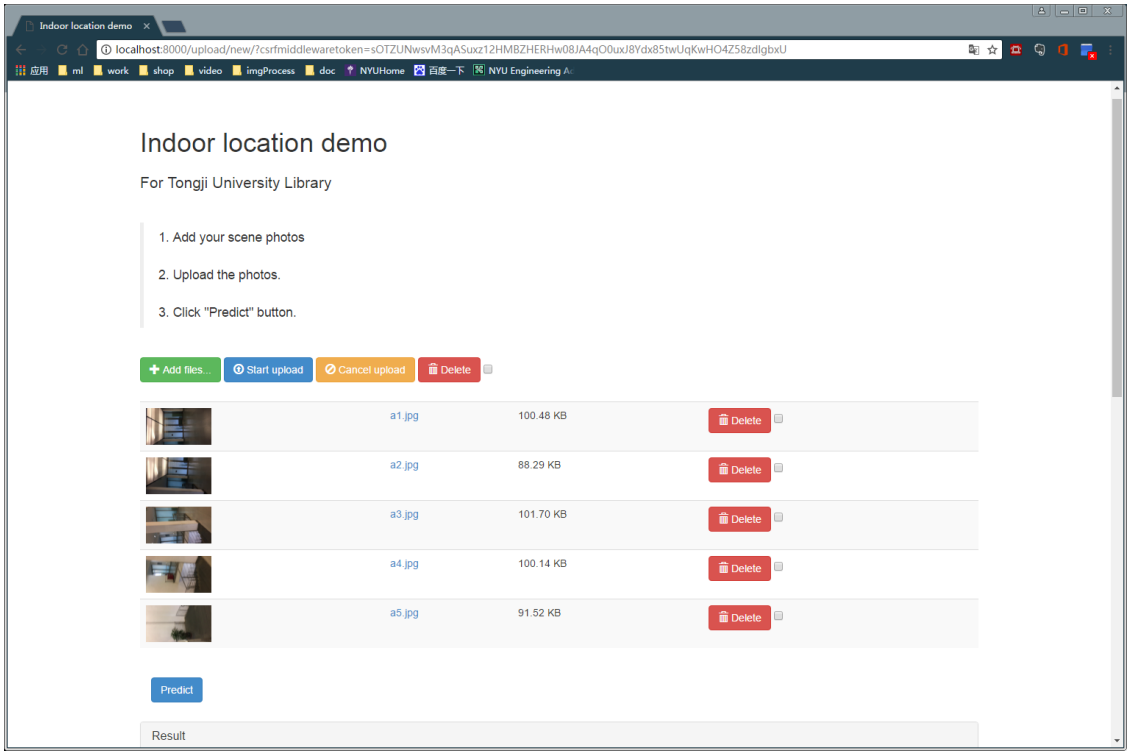

图 4.9 上传图片

（3）预测结果

点击下部蓝色按钮，上有‘predict’字符，即可以得到预测结果如图 4.10：

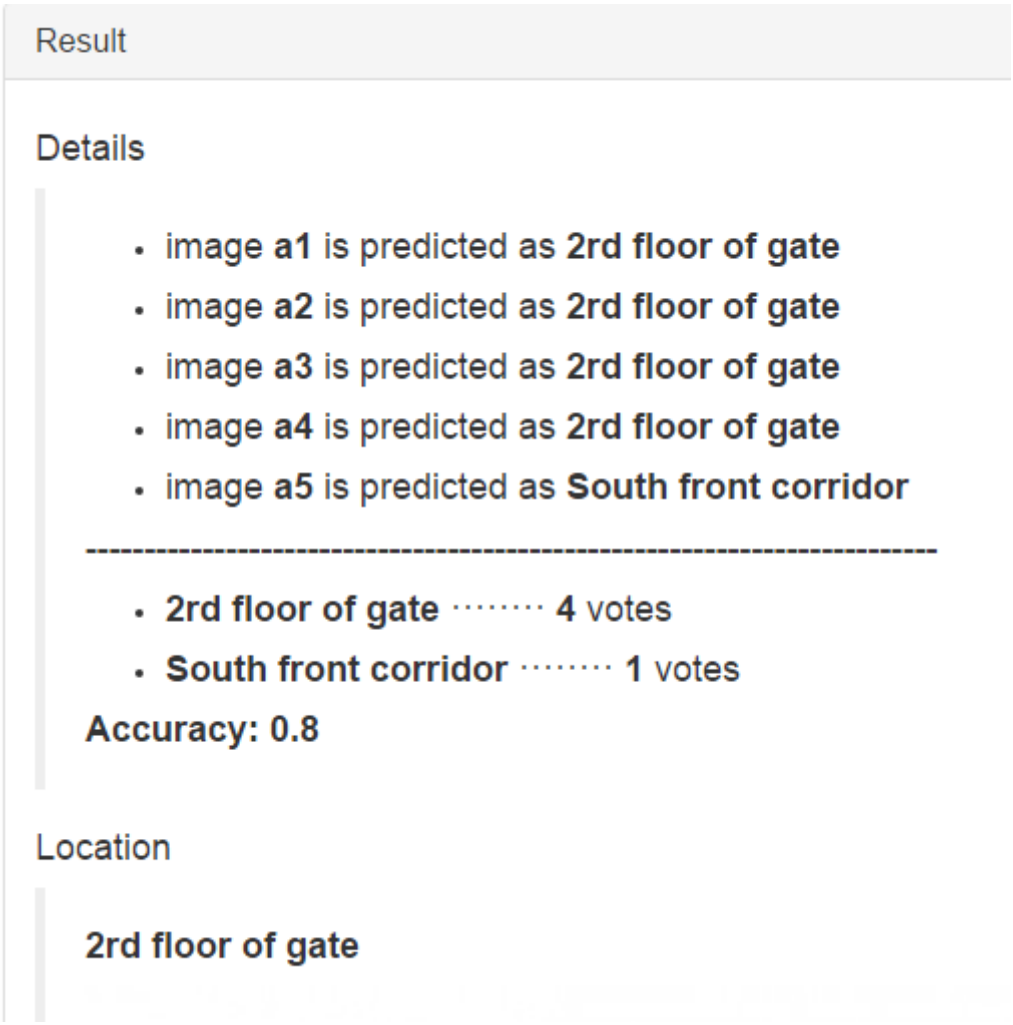

图 4.10 预测信息

该图展示了一些预测细节，比如图片 a1 被预测为大门二楼的场景，而图像 a5 被预测为南部走廊场景。总共有四张图像被认为是属于大门二楼场景的，一张图像被认为是属于南部走廊的。而准确率是 0.8，表示上传的 5 张图像里有一张被分析错了。因为我的图像都有标签，所以可以看出图像 a5 也应该属于大门二楼，但是分类器将他分类到了南部走廊，这可能因为这张测试图像和南部走廊更像。

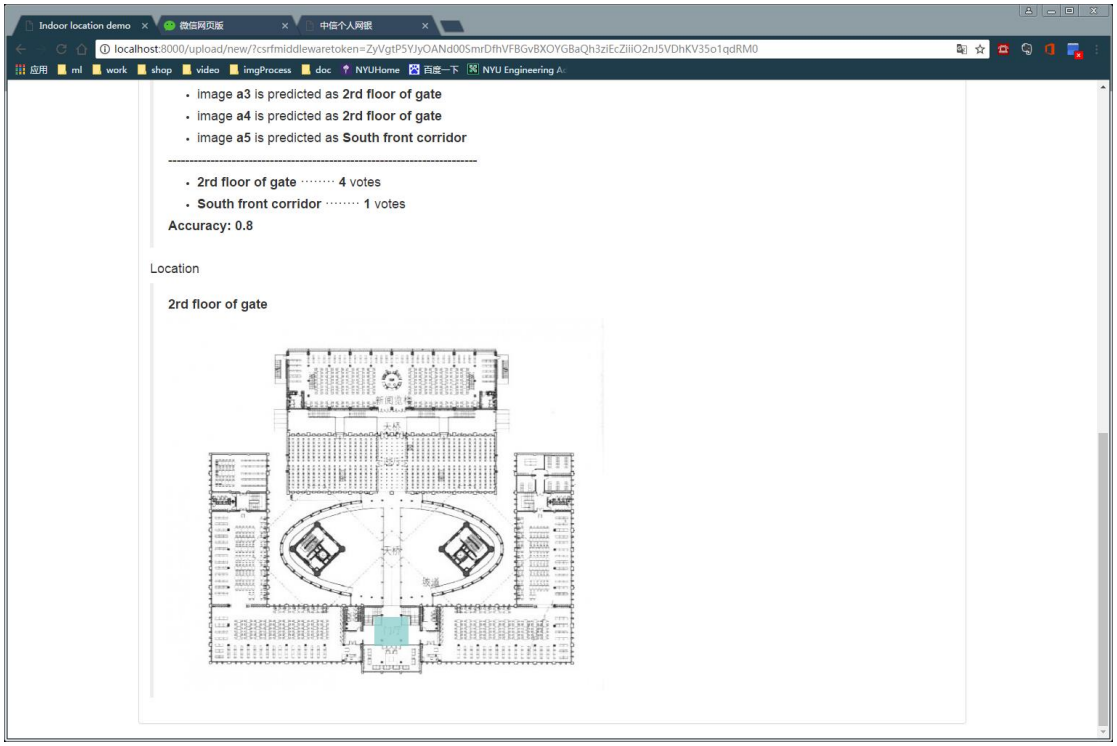

图 4.11 图书馆平面图

最后图 4.11 表示网页上展示了一副图书馆的平面图，并用颜色标记了预测的场景位置，如上图淡蓝色区块。不同的预测结果会在平面图上标记不同的位置。

4.5.4 实验结果及分析

服务器对 10 张问询图像处理和返回定位结果的平均用时是 2.4 秒，由于本文使用的训练图库是具有重复性和相似性的图书馆室内场景，比如相似的楼梯和走廊，所以有时也会出现匹配错误的情况，但是如果用户输入的图片在三张或以上并且图像描述的信息较为全面，那么实际使用的准确率可以是完全准确的。

因为选择使用的数据集是同济大学图书馆一楼即 TLD 数据库，数据集中包括了 1317 张图像，场景包括了图书馆的各个房间和走廊。在测试中，在不同的光照下，结果显示在夜晚的条件下准确率会下降一些。因为提取图像特征时没有考虑到图像的颜色信息。总体可以到 95%以上。

4.5.5 实验总结

本文的系统利用图像特征描述算子 SURF 对图像进行特征提取，再利用机器学习中的 SVM 支持向量机建立分类模型。现在该系统的只能满足图书馆的需求，在下一步希望扩展整个系统，添加更多室内场景，覆盖整个校园。

5 总结与展望

5.1 本文工作总结

本文介绍了利用在室内极易获取的图像作为信号源结合机器学习实现室内定位系统。用户可以将自己所处的地点拍照上传到服务器，服务器端通过构建词袋模型，再输入场景分类器，从而得到更准确的定位信息并返回结果。为了提高提高数据处理的效率和和分类器准确率，本文选取了 64D 的 SURF 特征描述子进行特征提取，在提高了运算效率的同时还提高了训练出的分类器的准确率；利用词袋模型将图像特征进行聚类，形成视觉词典，选择了更适合的密度聚类算法 Gaussian Mixture Model，用视觉词汇来表示图像，在大规模图像特征数据的条件下为场景分类器的训练提供了更高效的输入向量；在场景分类器的选择上，选取了更加稳定准确率更高的 SVM 支持向量机，并计算了更加适合图像特征的核函数以及构建了多类 SVM 分类器。

本文选取了 Django 网络框架，使得服务器的构建更加快捷高效。用 HTML5 技术简单地编写了浏览器页面，可以通过该接口上传问询图像到传服务器请求定位，服务器根据问询图像进行判断后返回问询图像的位置。 整个用户操作流程简单快捷，响应时间也较快，不会让用户厌烦的等待。

5.2 展望与建议

现在虽然室内定位技术层出不穷，但是并没有成熟的普及的商用系统。我觉得随着计算机视觉研究的不断发展，以及深度学习作为机器学习里的重要领域在其他学科的应用也更加成熟，两者的结合必定会为室内定位的实现带来质量和效率的突破。加上现在是大数据的时代，谷歌地图、高德地图等等都拥有海量的数据可供深度学习，可以支持准确率。图片的采集和预处理是一项比较繁重的工作，但是谷歌街拍等项目让这个任务不是不可能。

另外，室内定位技术完全可以通过糅合多种位置信号来实现精准定位，比如地磁信号。在现代建筑中，钢筋水泥的建筑框架给室内磁场造成了一种系统上的扰动，加强了室内环境里不同位置磁场的信号特征的差异。除此之外，还有红外、WiFi、GPS 等其他信号源可以作为辅助信号。多种信号的共同作用和矫正可以在实际使用中大大提高准确率。

参考文献

[1] Durrant-Whyte H, Balley T. Simultaneous localization and mapping: part I. IEEE Robor Autom Mag, 2006, 13:99-110

[2] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, 2006. 2169-2178

[3] Xiao J X, Hays J, Ehinger K A, et al. Sun database: large-scale scene recognition from abbey to zoo. In: Proceesings of IEEE Conference on Computer Vision and Pattern Recognition, San Franciso, 2010. 3485-3492

[4] Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91-110

[5] Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vision, 2001, 42: 91-110

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346–359, 2008

[7] Hugh Durrant-Whyte. What is a robot?[EB/OL]. http://www.abc.net.au/tv/bigideas/stories/2012/06/25/3530523.htm

[8] Dalal N, Triggs B. Histograms of oriented gradients fir human detection. In: Proceesings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 886-893

[9] Felzenszwalb P, Girshick R B, McAllester D, et al. Object detecetion with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 1627-1645

[10] Pandey M, Lazebnik S. Scene recognition and weekly supervised object localization with deformable part-based models. In: Proceedings of IEEE International Conference on Computer Vision, Barcelona, 2011. 1307-1314

[11] Wu J X, Rehg J M. CENTRIST: a visual descriptor for scene categorization. IEEE Trans Pattern Anal Mach Intell, 2011, 33: 1489-1501

[12] Li L J, Su H, Lim Y, et al. Objects as attributes for scene classification. In: Proceedings of European Conference on Copmuter Vision, Heraklion, 2010. 57-69

[13] Sadeghi F, Tappen M F. Latent pyramidial regions for recognizing scenes. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 228-241

[14] Juneja M, Vedaldi A, Jawahar C V, et al. Blocks that shout: distinctive parts for scene classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 923-930

[15] J. Sivic, A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos . IEEE International Conference on Computer Vision, Volume 2, page 1470--1477, 2003.

装 订 线

[16] Herbert Bay, et al. Speeded Up Robust Features. ETH Zurich, Katholieke Universiteit Leuven

[17] Binmore, Ken; Davies, Joan. Calculus Concepts and Methods. Cambridge University Press. 2007: 190. ISBN 9780521775410. OCLC 717598615

[18] Koenderink, J.: The structure of images. Biological Cybernetics 50 (1984) 363 –370

[19] R.A. Haddad and A.N. Akansu, "A Class of Fast Gaussian Binomial Filters for Speech and Image Processing," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 39, pp 723-727, March 1991

[20] Haar, Alfréd (1910). "Zur Theorie der orthogonalen Funktionensysteme", Mathematische Annalen, 69 (3): 331–371

[21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 110 (2008), pp. 346–359

[22] Fei-Fei Li; Perona, P. (2005). "A Bayesian Hierarchical Model for Learning Natural Scene Categories". 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2: 524

[23] Cortes, C.; Vapnik, V. Support-vector networks. Machine Learning. 1995, 20 (3): 273–297

[24] Hosmer, D. W. and S. Lemeshow: Applied logistic regression. New York; Chichester, Wiley, 2000

[25] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach 2nd. Prentice Hall. 2003 [1995]

[26] Huang, Z. (1998). Extensions to the K-means Algorithm for Clustering Large Datasets with ategorical Values. Data Mining and Knowledge Discovery, 2, p. 283-304

[27] FreeMind. 漫谈 Clustering (3): Gaussian Mixture Model [EB/OL](2009-02-02)[2015.5.12]. http://blog.pluskid.org/?p=39.

[28] 刘礼. 基于视觉的室内高效定位研究[D]. 成都：电子科技大学,2015

[29] K. Grauman and T. Darrel. The Pyramid Match Kernel:Discriminative Classification with Sets of Image Features. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Beijing, China, October 2005.

[30] 张贺. SVM 实现多分类的三种方法 [EB/OL](2016-03-11)[2017/6/4]. http://www.cnblogs.com/CheeseZH/p/5265959.html

[31] Andrew Gilbert, er al. ImageCLEF 2012-2015 WEBUPV Image Annotation Datasets[EB/OL]. http://www.imageclef.org

[32] Opencv team. Overview of opencv[EB/OL]. http://www.opencv.org

[33] Django team. Meet Django[EB/OL]. http://www.djangoproject.com

[34] W3school team. Html5 教程[EB/OL]. http://www.w3school.com.cn/html5/index.asp

谢 辞

四年的大学生涯转瞬即逝，我仿佛感觉我昨天才迈入同济的校园，然而明天就要离开可爱的校园，时间过得真快啊。

首先十分感谢赵钦佩老师在选题之初对我的宝贵建议，让我能一步一步从懵懂到实现整个项目。同时也同样十分感谢饶卫雄老师，和赵钦佩老师一起，在每周的组会上都对我进行了严格的监督和悉心的指导，并为我所遇到的难题提供了许多解决思路和帮助，非常感谢你们的认真负责。

同样十分感谢软件学院的各位老师，特别是数字媒体专业的老师们，你们在图像等方面对我的培养让我在毕业设计里如虎添翼，少走了很多弯路。也很感谢你们对我的兴趣培养，让我对编程、游戏设计等方面产生了极大的兴趣。

最后，十分感谢一路陪伴我的同学和朋友，和你们度过的每一天都是我大学生活珍贵的记忆。有了你们我的生活才丰富多彩！

谢谢大家！

装

订

线