工学硕士学位论文

基于语义的视觉定位算法研究

RESEARCH ON VISUAL LOCALIZATION ALGORITHM BASED ON SEMANTICS

戴进

哈尔滨工业大学

2019 年 6 月

国内图书分类号：TN911.73　　学校代码：10213
国际图书分类号：654　　密级：公开

工学硕士学位论文

基于语义的视觉定位算法研究

硕士研究生：戴进
导　　师：谭学治教授
申请学位：工学硕士
学　　科：信息与通信工程
所在单位：电子与信息工程学院
答辩日期：2019 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index：TN929.5
U.D.C：621.3

Dissertation for the Master's Degree in Engineering

RESEARCH ON VISUAL LOCALIZATION ALGORITHM BASED ON SEMANTICS

Candidate： Dai Jin

Supervisor： Prof. Tan Xuezhi

Academic Degree Applied for： Master of Engineering

Speciality： Information and Communication Engineering

Affiliation： School of Electronics and Information Engineering

Date of Defence： June, 2019

Degree-Conferring-Institution： Harbin Institute of Technology

摘 要

随着现在网络的发展以及可穿戴设备的普及，使得人们对自身位置信息的需求日益提升，因此基于位置服务的发展愈发迅速。目前人类每天约有80%左右的时间在室内活动，因此室内定位凭借其独特的优势正逐渐获得研究人员的广泛关注。室内视觉定位技术更是凭借其内置传感器的独特优势，适用性远超其他诸多需要部署开销的定位系统。此外，以视觉信息进行定位的方式与人类自身通过眼睛确定位置过程近似，更值得进行深入研究。

本文将机器学习中的语义分割与定位算法相结合，首先研究了视觉定位技术与语义信息应用的国内外研究现状，并对机器学习的发展与视觉定位的结合进行了分析。其次，本文研究了机器学习分割出的语义成分在视觉定位系统中的应用。此外，本文针对传统算法系统中的不足针对性地做了以下研究：

（1）针对传统视觉定位系统离线阶段建立的数据库中数据量较大、图像检索耗时过长的问题，提出一种基于语义的离线数据库分类方法，该方法利用机器学习的方式对数据库中图像进行了语义提取并分类为语义子数据库，能够有效消除随着数据库容量增大，在线阶段检索时间延长的线性增长关系；

（2）针对在大型数据库中，检索效率与准确率不高的问题，提出了一种基于语义与内容的快速图像检索方法，该方法先在离线数据库中进行检索区域定位，再在已分类的语义数据库进行高精度检索，提高了在线阶段检索匹配的检索效率与准确率。并针对传统视觉定位系统中定位阶段对全局进行特征提取导致时间开销较大的问题，提出了一种基于语义约束的特征点定位算法，该算法有效地减少了定位匹配阶段提取特征点的区域与提取特征点的个数，降低了特征提取的复杂度及时间开销。此外，由于语义限制，该方法可以剔除大量的误匹配特征点，兼顾特征提取算法的速度与定位算法的精度，提高了系统的实时性。

综上，本文就论述的算法进行了仿真。仿真结果表明本文提出的基于语义与内容的快速图像检索方法能够在保证检索匹配图像精度的情况下，有效地减少了在线检索阶段寻找匹配图像的时间开销；基于语义约束的特征点定位算法可以减少特征提取所需时间，提高整体系统的实时性表现。

关键词：视觉定位；语义提取；语义数据库；语义约束

Abstract

With the development of the network and the popularity of wearable devices, people's demand for location information is increasing. Therefore, location-based services are gradually attracting researchers' extensive attention. As people spend about 80% of their time in the indoor environment, indoor positioning technology is a research hotspot in the current positioning technology. Indoor visual positioning technology, with its unique advantages of built-in sensors, is gradually replacing those indoor positioning systems required extra cost. In addition, the way of positioning by visual information is similar to the process by human eyes, which is more worthy of further study.

In this paper, semantic segmentation and localization algorithm in machine learning are combined. Firstly, the research of visual localization technology and semantic information are studied, and the combination of machine learning and visual localization is analyzed. Secondly, this paper studies the application of semantic components in visual localization system. In addition, this paper has completed the following research on the problems existing in the offline and online phases of the visual positioning system:

(1) Image retrieval stage of visual positioning system takes high time cost because of the large database, this paper proposes a fast image retrieval method based on the semantic and content, the method using machine learning divide image database into the semantic database, can effectively reduce the online retrieval time. Precision retrieval is carried out in the semantic sub-database, which ensures retrieval accuracy and saves time.

(2) Aiming at the problem of low accuracy in large database, the SCBIR method is used to perform accurate retrieval by using semantic database. Aiming at the problem of high time cost in global feature extraction of the online stage, a feature point location method based on semantic constraint is proposed, which effectively reduces the location area and the feature point extracted in matching phase, in order to reduce time cost of the features extraction. In addition, due to semantic constraints, this method can eliminate a large number of mismatched feature points, and take into account the speed of feature extraction method and the accuracy of positioning method, so as to improve the performance of whole visual indoor positioning system.

The simulation experiments above are presented. The results show that the proposed method can effectively reduce the time cost of retrieval in the online stage while ensuring the retrieval accuracy, and reduce the time cost for feature extraction. Finally, improve the performance of the whole positioning system.

Keywords: visual positioning, semantic extraction, semantic database, semantic constraints

目 录

第1章 绪 论

1.1 课题研究的目的和意义

本课题来源于国家自然科学基金《基于群智信息感知模式的WiFi室内定位系统中Radio Map构建方法》（项目编号：61571162）和黑龙江省自然科学基金《基于视觉的室内定位技术研究》（项目编号：F2016019）。随着现在5G网络的发展以及可穿戴设备的普及，人们对自身确切位置的需求也日益增加，因此对用户进行快速准确的定位算法正得到研究人员的广泛关注。经调查，人类在室内活动的时间占人类活动总时间的80%左右，因此对室内位置信息的需求要远大于对室外位置信息的需求，室内定位技术也成为了当前定位技术中的广泛研究的也因此获得了极大关注。其中室内视觉定位技术更是凭借其独特的优势在诸多室内定位技术中脱颖而出。诸多的室内定位技术需要进行额外的开销布置，例如WLAN定位中的WiFi接入点布置，蓝牙定位中的蓝牙设备安装等。但室内视觉定位技术只需要手机或相机内置的图像传感器即可完成定位，这种便利的定位技术正在逐渐替代那些需要在室内场景中安装额外定位开销的室内定位系统。此外，视觉定位技术与人类自身定位方式相类似，初到陌生场景中，都是先通过视觉信息来确定自身位置，其可以与人工智能相互融合的定位方式更加值得进行深入研究。

随着环境的变化，在任何场景均可靠并精准的定位方式可以让位置需求使用者拥有更好的体验。由于室外定位技术的研究略早于室内定位技术，因此大部分的室内定位技术是由室外定位技术演变改进而来的，而有些室外定位技术在演变到室内定位技术的过程中，却并不具有很强的实用性。例如在室外定位中，不得不提的是基于全球定位系统（Global Position System，GPS）以及基于GPS的地图位置服务和各种导航系统。GPS定位技术在室外具有高适用性，其高精度的定位使其成为了室外使用最广泛的定位技术。然而室外场景和室内场景存在相应的偏差，由于室内场景的特殊性，GPS信号在室内场景中由于信号遮挡等问题导致信号较弱，无法进行精确定位。此外，室外定位与室内定位的精确要求也略有不同，GPS信号只能提供大致位置信息，由于非军用，因此无法获得高精度的定位信息，因此无法满足室内定位的精度要求[1]。由于GPS定位系统室内应用的不理想，研究人员逐渐把目光放到其他定位系统上，例如蓝牙[2]、FID[3]、超声波[4]、超宽带[5]、WLAN[6]以及视觉[7]等定位系统。在目前的室内定位系统中，基于WLAN的定位方式是应用场景最多的定位技术之一。基于WLAN的室内定位系统需要在室内环境中提前部署无线接入点。然而，由于室内场景的复杂性，该技术需要部署大量无线接入点才能

完成精确定位，在某些室内外交界处，会由于墙壁遮挡、门窗遮挡等因素导致定位精度较低，不能满足用户在室内环境中对定位精度的需求。大量无线接入点的部署会产生高额的部署开销，对无线接入点数目的以及对人流密集度的过度依赖使得该方法具有很大的局限性。其他的室内定位系统同样需要进行定位设备的额外部署，因此在室内定位中同样具有局限性。

上述室内定位系统中在实际场景应用时均有不足，针对这些问题，研究人员把目光逐渐转移到了室内视觉定位系统上。室内视觉定位系统是一种融合了图像处理、模式识别、计算机视觉等多种学科理论形成的仅通过视觉信息进行定位的系统。该定位系统与其他定位系统相比，其显著优势是不需要对室内场景进行额外的开销部署，仅通过图像信息对用户周围环境进行感知，在复杂室内环境可以对用户进行精确位置确认。除此之外，视觉定位系统中需要的图像信息具有方便获取、不易篡改等特点，其自主性强、定位精度高也是独特优势。随着可穿戴设备及手机等设备的普及，室内视觉定位系统也正在逐渐替代那些需要在室内进行设计的其他室内定位系统，其仅通过图像信息进行定位的方式也与人类自身在位置环境中的定位方式想类似，可以与如今正在研究热点的人工智能领域向融合。

在实际应用中使用视觉定位系统进行定位时，根据系统所安装的成像设备的数量与种类的不同，可以将传统的视觉定位系统分为三种形式，单目系统、多目系统以及深度系统。单目视觉定位系统即只使用单个的图像采集设备进行离线数据库采集与在线阶段定位，其主要优势为成本低、设备简易以及在离线和在线阶段数据采集流程简单，无需进行额外参数标定。多目视觉定位系统即使用多个图像采集设备针对不用方向、同一方向的多个角度进行图像采集，其主要优势为实现了对待定位环境的多角度全方位的图像采集，并通过多目摄像头采集的图像可以直接对采集位置的三维信息进行确定。深度视觉定位系统在图像采集过程中应用了深度相机，采集的结果不仅有图像信息，还拥有红外或激光雷达探测得到的伴随图像信息生成的深度信息。利用深度相机采集的图像结果与其他两个视觉定位系统基本相同，但深度信息是另外两个视觉定位系统所不具备的，既节省了通过图像信息计算深度的复杂过程，又提高的深度信息结果的准确性。三个视觉定位系统各有优势，但从视觉定位系统的实用性角度分析，多目视觉定位系统需要对图像采集设备的多目摄像头位置进行确定，进行外参矩阵计算，并不可随意更改；深度视觉定位系统由于图像采集阶段应用了深度相机，在在线定位阶段对用户定位设备要求过高，不具有普适性。而且当前阶段大部分深度相机的达不到室内定位要求的测量范围、对噪声与光照十分敏感，也无法对透明物体进行测量，在实际应用中受到很大程度上的局限。因此，定位流程简单，对用户定位设备无特殊要求的单目视觉定位系统

与其他两种视觉定位系统相比，具有更高的实用性与便利性。

经过对定位系统设备资源成本与定位设备部署人工成本的综合考量，本文选用单目视觉定位系统进行定位，并结合机器学习的方式，在保证在线阶段检索及定位精度的前提下，有效地提升了定位的效率，节省了时间开销。此外，针对目前传统的室内视觉定位系统中普遍存在的在线检索时间随数据库容量增大而延长的问题，本文以语义数据库的形式来代替传统的定位数据库，该方法明显减小了在线检索阶段的时间开销，在整个定位系统实时性能上有了很大提升。

近年来，同时定位和地图构建（Simultaneous Localization and Mapping，SLAM）技术在计算机视觉与机器人领域都具有广泛的研究价值[8]。目前许多前沿研究将SLAM技术与地图构建中的语义信息相结合[9]，这使得最终构建的地图精度又有了大幅提升。此外，包含了语义信息的地图也可以应用到更多的场景，为许多其他技术提供了更精确的先验信息。综上所述，本文的研究意义如下：本文提出了一种基于语义与内容的快速图像检索方法，解决了在线阶段随着离线数据库容量增加检索时间延长的问题；并利用得出的语义位置信息对定位结果形成约束，减少定位阶段提取特征点的数量，将定位精度与定位效率同时考虑，达到平衡。将传统的图像地图构建与语义信息相结合将为SLAM提供一个新的地图构建形式，也可以为其他以视觉为基础的研究提供更精确的地图信息，其定位的高效性也提升了用户的使用体验。此外，利用AR形式对用户进行目的地路线规划，对目的地相关信息进行索引等功能也是该研究领域的延伸发展方向，旨在为用户提供便利性更强、精确度更高的多元化定位服务。

1.2 国内外研究现状

1.2.1 视觉定位技术在国内外的研究现状

室内定位技术以是否需要部署额外设备为前提可大致分类两类：基于设备部署与非基于设备部署[10]。其中，后者的研究还处于萌芽阶段，在硬件与软件的各个层面还有一些相应的限制，其目前的成熟度并不足以在实际应用场景中进行实验。基前者的定位研究由于设备的辅助性，并不存在如非基于设备部署的室内定位系统的诸多限制，因此是目前室内定位系统中的首要研究目标。

早基于设备部署的研究中，基于邻近信息、基于无线电指纹、基于几何特征以及基于航位推算几种方式最具有代表性。基于邻近信息技术对目标用户的定位需要一个已知位置作为先验条件，该定位技术通常包括两类：分别为基于射频识别与基于蓝牙定位[11]。基于无线电指纹技术以WiFi定位为主，其分为离线与在线两个

阶段。在离线阶段，需要进行指纹库的建立，将位置指纹与信号强度一起存储在建立起的指纹库中；在在线阶段，根据用户当前产生的信号强度值与位置指纹库进行最优匹配，即可对用户完成精确定位[12]。基于几何特征的室内定位技术顾名思义，其主要到了数学中的应用几何学原理，通过一些附加设备，诸如WiFi、蓝牙、RFID或UWB等获取定位所需信息，例如到达时间（Time Of Arrival，TOA），到达角度（Angle Of Arrival，AOA），到达时间差（Time Difference of Arriva，TDOA），接收信号强度（Received Signal Strength，RSS）等测量信息[13]。利用上述测量信息，即可以通过几何学原理计算目标与信号源间的距离，从而确定目标的位置。基于行人航位推算（Pedestrian Dead Reckoning，PDR）的室内定位技术主要利用惯性测量单元进行定位，例如加速计等设备进行偏差较小且实时性较高的定位操作[14]。

基于设备的位置感知技术已经日趋成熟，但是一个新兴的利用视觉信息进行定位的室内定位技术也得到了国内外研究人员的广泛关注，在SLAM技术中也是不可或缺的一环。视觉室内定位技术在传统的定位技术中做出了改进，以视觉信息作为定位的主要信息，而其大体步骤与传统定位方式较为类似，例如与WiFi定位系统相比，同样拥有离线阶段与在线阶段两个阶段。

离线阶段是一个数据采集并建立离线数据库的过程，离线阶段进行数据采集之后，需要将图像进行存储。由于图像中包含丰富的信息，有与定位相关重要信息，也有会对最终定位形成干扰的无用信息，因此存在于离线数据库中的图片需要进行后续处理，选择并提取出需要的信息加以利用，特征选择与提取的操作不仅能消除图片中存在的干扰，如果特征选择得当，还能加快在线阶段的检索速度。目前数据库构建方法主要分为两类：一是寻找复杂度更低分类更准确的特征对图像进行特征提取，降低特征的复杂程度，提高特征对图像描述的准确程度，在减少检索时间的同时提高精度；二是对离线数据库进行分类，以便在在线阶段进行搜索时，减小相应的搜索区域，达到加快检索速度的目的。如果对离线数据库不采用分类手段，使用CBIR等方法直接对数据库进行图像检索，则会产生较大时间开销。常用的特征提取方法有主成分分析法（Principal Component Analysis，PCA）、BoW（Bag of Word）、稀疏模型和空间金字塔模型等，但是上述几种特征提取方法，都存在一定的问题。PCA方法会使类内距离增大，导致依据特征分类的准确率降低[15]；BoW方法对关键词的聚类中心并不能明确定位，无法达到最优的检索效果[16]；稀疏模型所要求的纯净低噪声训练图片难以获取，对需要高准确率的检索环境来说，难度较大[17]；空间金字塔模型的复杂度过高不利于实时演算[18]等，在室内离线数据库的检索中都存在着弊端。常用的数据库分类方法有基于神经网络的分类方法

以及基于无监督学习的聚类方法等。文献[19]利用卷积神经网络与区域卷积神经网络对肺部疾病进行了分类，针对细节分类进行了性能提升。但该方法对于数据库中的一张图片包含多类别landmark的场景并不适用。文献[20]采用了聚类的方式将数据库中图像大致分为四类，然后再从每一类中进行精确检索。但是聚类的方法是一个无监督学习方式，每一类所代表的含义以及每一类中的对象是不确定的。此外，聚类方法最终聚为几类需要经过大量测试才能确定最佳效果，检索准确度较低。文献[47]将SVM分类方法与基于内容的图像检索方法相结合，对图像进行快速检索，虽然提高了检索效率，但由于SVM方法是二分类方法，导致检索准确率较低。

在线阶段是一个对待定位图像进行检索并获取最终定位结果的过程。在线阶段的匹配图像检索过程与离线阶段的离线数据库构建方法需一致，利用同样的特征提取方法定位到待定位图像的匹配图像。由于匹配图像所携带的信息包含离线数据库构建时其自身的位置，因此待定位图像的粗略位置已知，需要进行精确定位。常用的精确定位方法大致分为三类，分别为2D-2D方法，3D-2D方法以及3D-3D方法。最具代表性的2D-2D方法为对极几何方法，该方法仅利用图像中的2D信息进行定位。文献[21]利用多视角几何方法将多张图像进行多特征提取，并进行了包含颜色信息的3D点云构建；3D-2D方法相比于2D-2D方法多了三维信息，即真实世界坐标。文献[22]利用PNP方法用单相机进行了室内定位，该方法拥有很高的在线定位精度；3D-3D方法通常应用与三维定位方面，该方法利用了深度相机或激光雷达获取的三维深度信息，进行高精度定位或模型构建。文献[23]利用ICP方法进行了全局三维点云配准，该方法在保证全局最优的同时，加快了点云匹配的速度。

目前，国内对于视觉定位的研究还处于初步阶段，大多集中在机器人领域。文献[24]提出了一种基于辅助靶标的移动机器人定位方法，该算法具有算法复杂度小、实时性高等优点，通过与机器人移动路径区域视觉信息的特征提取与处理，并结合辅助靶标信息，可以达到较高的定位精度。文献[25]利用了卷积神经网络，将机器人运动的空间与时间相关性进行了约束，达到了较高精度的定位效果。大疆公司也一直在进行无人机视觉方面的深入研究，其将超声波和摄像头一起其他多种传感器进行融合，使无人机在复杂位置的室内场景中同样拥有很高的定位精度。

1.2.2 图像语义分割在国内外的研究现状

随着计算机功能的不断发展，深度学习的方法逐渐应用到了各个领域。深度学习概念最早在2006年由Hinton等人提出[26]，在机器学习中这是一种对数据进行表征学习的方法。卷积神经网络（Convolutional Neural Network，CNN）作为机器学习中最成功的一共模型，在近年来不断取得突破性进展[27,28]，以CNN为基础框架的机

器学习算法广泛的应用在了图像分类、机器翻译、语音识别等领域。与CNN并驾齐驱的几个深度学习网络模型还有循环神经网络（Recurrent Neural Network，RNN）以及生成对抗网络（Generative Adversarial Network，GAN）等。CNN作为最早的神经网络模型，其基本结构较为简单，由输入层、卷积层、池化层、全连接层以及输出层组成。图像由输入层输入，经过卷积层的多个卷积核进行卷积之后，再经过池化层进行特征降维，将底层较为粗糙的特征逐渐变为高层较为精细的特征，最后经过多层卷积层与池化层后，通过全连接层进行分类，再由输出层进行输出。CNN由于其卷积的特性，所提取的对平移、伸缩。倾斜等变形均不是十分敏感，因此十分适合处理图像数据。RNN是一种环状神经网络，网络自身拥有记忆系统，可通过记忆对任意时序的输入序列进行处理，例如对图像上下文之间的连续性信息可以进行更加合理的利用，长短期记忆神经网络（Long Short Term Memory，LSTM）是目前以RNN为网络基本架构最具有代表性的网络之一。GAN主要由生成器网络和判别器网络组构成，GAN网络的核心思想是不断对训练库中的训练样本进行学习，同时利用生成器网络根据训练样本模式不断产生人造样本，并输入到判别器网络中不断进行判断。在整个训练过程中，让生成器网络与判别器网络不断进行相互对抗，从而相互提高。

在图像语义分割领域，CNN网络也凭借其优良的学习性能与准确的应用效果得到了许多领域学者的青睐。语义分割概念由Ohta等人首次提出，其基本的定义为：为图像中的每一个像素分配一个预先定义好的表示其语义类别的标签[29]。语义分割技术是从传统的图像分割技术中演进而来的，与图像分割相比，图像语义分割技术在其基础上为图像前景中的目标分配了一定的语义信息。该语义信息可以从图像本身的纹理、图像中场景的关联以及其他高层语义特征来获得，相较于普通特征更具有实用价值。语义分割技术可以广泛的被应用在自动驾驶领域[30,31]，以及作为医学图像分割和行人目标检测等领域的技术基础。图像语义分割方法根据分割特点可以大致分为两类：分别为基于区域分类方法和基于像素分类方法。

基于区域分类的图像语义分割方法通常将传统图像处理算法与深度神经网络（Deep Neural Network，DNN）相结合。首先对原始图像进行划分，框选出一系列的不同目标的候选区域；其次将全部候选区域输入到深度神经网络中，让网络对区域内的每个像素进行语义分类；最后根据深度神经网络的分类结果，对原始图像进行语义标注，得到分割结果。基于区域分类的图像语义分割方法的分割效果与输入图像的质量息息相关，其核心在于如何生成准确的不同目标候选区域框。根据候选区域生成方法与图像划分的标准，可以将该方法分为两类：基于候选区域的方法

与基于分割掩模的方法。基于候选区域的算法的优点为将物体检测技术与语义分割算法相融合，使用物体检测技术得出较为准确的候选区域，并完成语义分割与目标检测两项任务。但是该方法对图像全局的语义信息没有充分考虑，对小尺度语义信息容易忽略导致分割出错。文献[32]利用生成敌对网络对目标遮挡进行训练，并利用Fast-RCNN网络进行了高精度的目标检测。文献[33]通过对RCNN添加了一个预测对象掩码的分支，提出了Mask R-CNN算法，相较于RCNN算法，Mask R-CNN速度更快且相较于其他网络模型表现更好。基于分割掩模的方法的优势为利用RCNN等物体检测技术进行分割掩模生成，并对掩模进行精炼与优化，可以对多尺度图像以及图像的隐藏含义进行深度挖掘。但是由于其对图像隐藏含义挖掘过于深入，因此对于背景较为复杂或被遮挡的物体分割准确率较低。文献[34]提出了一种新的自顶向下细化方法来增强前馈网络的目标分割，由此产生的自底向上、自顶向下体系结构，该结构能够有效地生成高保真对象掩码使得分割精度提升了10%-20%。文献[35]提出了一个Multipath结构网络，提高了对小目标对象分割的精确度，并解决了在物体检测过程中尺度变化、目标遮挡和集群等问题。

基于区域分类的图像语义分割方法虽然已经取得了不错的分割效果，但在图像分割精度与速度上还达不到理想的状态，因此基于像素分类的图像语义分割方法被提出。该方法从大量原始图像、标注图像及弱标注图像中提取出图像特征与语义信息并将其输入到深度神经网络中进行训练，以端到端的方式对每个像素进行分类，有效地提升了分割准确率。基于区域分类的图像语义分割方法可以以标注类型和学习方式的不同主要分为两类：全监督学习方法和弱监督学习方法。

全监督学习图像语义分割方法需要对图像进行像素级标注，从而才能利用丰富的视觉特征对图像进行像素级分类。文献[36]利用全卷积网络进行语义分割，并在其末端增加了一个全连接的条件随机场，并利用带孔卷积核扩大了卷积提取出特征图的感受野，增强了像素间的关联与空间一致性。文献[37]提出了一种扩张卷积算法，可以在不损失分辨率的条件下增加感受野，是对带孔卷积的进一步优化。文献[38]利用在[37]的基础上进行了改进，利用混合扩张卷积代替了扩张卷积，既增加了卷积核的感受野，又将局部相关性进行了有效保持。由于全监督学习方法需要进行像素级的图像标注，制作训练数据库的过程十分耗费人力物力，因此一些学者想对其进行简化，以弱监督学习的方式对图像进行训练，减少构建训练库的成本。弱监督学习方法只需要对图像进行弱标注，例如图像级标注、边框级标注或涂鸦级标注。图像级标注是指将整幅图像进行语义标注，对一副图像只以物体种类打上语义标签即可；边框级标注是指人工将语义主体部分用矩形框进行标记，并标注语义种类信息；涂鸦级标注则是指用类似涂鸦的点或线条对图像进行语义标注。文献[39]

对训练样本进行了边框级标注，并将其输入到全卷积网络中进行训练，利用循环迭代的方式提高准确率。文献[40]以随机涂鸦的像素点作为监督信息，提出了点监督方法，并将监督信息与CNN相结合，对网络的损失函数进行了优化，取得了较好的分割效果。文献[41]对训练数据采用了图像级标注方法，利用分类网络对图像中显著性区域目标按主次进行获取，并对显著性区域进行提高像素精度处理，解决了语义分割边缘模糊的情况，得到了很好的分类效果。

目前，国内也逐渐将语义分割技术投入到实际场景中进行应用。文献[42]对20m航拍高度影像进行了图像植被识别，其构建的FCN-VGG19植被识别正确率达到了85%以上。文献[43]提出了一种基于深度学习的实时图像语义分割框架RT-SegNet，利用编码阶段、解码阶段和降为阶段三个阶段的优化，对于多种公共数据集其语义分割准确率均有很大提升。文献[44]对遥感图像进行语义分割，采用了改进的U-Net网络实现了端到端的像素级语义分割，其分割准确率可以用于实际工程。

综上所述，本节主要分析了语义分割技术的国内外发展现状。结果表明，语义分割技术已经融入到了计算机视觉与模式识别的各个领域，也是目前深度神经网络优化发展的目标之一。该技术目前主要应用于自动驾驶领域，通过更精确的语义识别，给用户带来更优越的体验及更智能化的服务。

1.3 本文的主要研究内容和结构安排

本课题结合了图像处理与机器学习的相关方法，在离线阶段，主要进行了图像离线数据库的建立。针对目前视觉定位随着离线数据库增大而导致检索时间延长的问题，本课题提出了一种基于机器学习的语义分类算法对离线数据库进行细致分类，将包含大量图像的离线数据库分类为多类包含少量图像的离线子数据库，减少了待检索图像在数据库中搜索匹配图像时的搜索范围，从而缩短在线阶段的检索特征所需要的时间，实现在现阶段匹配图像的快速查找。在在线阶段，在图像特征匹配的过程中，通常算法为对整张图片进行全局的SURF、HOG或者其他特征匹配。为了提高算法的速度与精度，本文提出了一种基于语义与内容的快速检索（Semantic and Content-Based Image Retrieval，SCBIR）算法，跳过了对离线数据库全局进行检索的阶段，直接针对待检索图像所属的语义数据库进行精确检索，降低了传统算法搜索整个离线数据库所带来的高复杂度，加快了在线阶段针对于大型数据库的检索速度，同时也提升了在线定位阶段的实时性。此外，求解用户位置的过程中需要用到基本矩阵。在计算基本矩阵过程中，以特征点集作为输入，进行两个特征点集之间的粗匹配。可采用线性估计基本矩阵的算法，从而恢复对极几何。然而在点集的粗匹配过程中，会存在一定数量的误匹配点对，由此估算得出的基本

矩阵是不精确的。因此对匹配特征点进行剔除，并降低特征点提取复杂度是十分重要的，对提升系统在线定位效率有很大意义。

本文针对传统算法中存在的问题，提出了弥补其缺陷的改进算法。在离线阶段，通过基于语义的数据库分类方法，解决了随着离线数据库容量增大，在线检索时间延长的问题，以保证检索精度为基础，最大化地节省时间开销；在线阶段通过提出了基于语义与内容的快速图像检索方法找到待定位图像的匹配图像，并通过基于语义限制的特征点定位方法有效地提高了在线定位速度，给用户更好的体验。

本课题的主要研究安排如下：

（1）在复杂的室内场景下，针对离线时间建立的数据库中数据量较大，图像检索耗时过长的问题，本课题提出一种基于语义的离线数据库分类方法，其使用机器学习中的语义分割网络将离线数据库中图像进行分类，变为离线子数据库的算法。将图像分类后，在线阶段即可在相应小区域内进行检索，减小在线检索阶段的时间开销。该算法的优势为在数据库图像容量较大时明显减少图像检索时间开销；

（2）在基于语义的数据库分类算法基础上，利用提出的SCBIR算法找到待定位图像的匹配图像。在针对传统定位算法在线阶段中，根据对极几何关系，需要寻找两张图像中多对特征匹配点进行匹配而导致的特征提取复杂度较高，提取时间较长，且存在大量误匹配点的问题，提出了一种基于语义限制的特征点定位方法，根据在像检索中提取出的图像语义信息，在语义区域内进行特征提取与匹配，该算法不仅减少了图像中特征点的提取数量，而且由于语义限制，可以剔除大量的误匹配特征点，来提高特征提取算法的速度与后续定位算法的精度，将该算法应用于后续的基于对极几何求解地理位置的算法中，形成一种可以进行快速的室内视觉定位的方案，来提高整个系统的实时性与应用性。

针对本文上述提出的算法，本文分别对其进行了程序编写与仿真实验，其结果表明，基于语义的离线数据库算法能够准确地根据语义信息将离线数据库分类多类细致的语义数据库，并结合本文提出了SCBIR方法可以有效地达到在保证定位精度的前提下，减少在线检索阶段的时间开销的目的。此外，利用本文提出的基于语义约束的特征点定位方法，可以有效减少在线阶段提取特征点的数量并对误匹配的特征点对进行剔除，提高了在线阶段的定位效率。本文的章节安排如下：

第1章将对明确本文的目的意义；首先对课题来源进行说明。其次进一步对视觉定位技术以及语义信息提取在国内外的研究现状进行了详细分析，并通过对目前相关算法的对比，对其优缺点以及应用场景进行了说明。最后，将对本文的章节分布与研究内容进行阐述，明确本文的内容结构。

第2章对本文所需要的基本研究理论进行了说明。首先阐述了摄像机模型搭建，

对针孔成像模型与坐标系转换理论进行了说明，明确了视觉定位系统的基本框架。其次，阐述了本文主要用到的SURF特征提取方法，分别从特征提取与特征描述子的构造两个方面对其进行了详细说明，并对该算法的优良性能进行了分析。最后，对本文课题中基于语义的视觉定位算法结构流程进行了说明，并对其基本原理和层次结构进行了具体分析。

第3章对基于语义的离线数据库分类算法进行了研究。首先对语义分割网络进行了搭建，并对其中包含的子网络进行了功能分析，并进行网络优化函数的设置。其次对搭建好的语义分割网络进行了数据训练与参数调整，使得网络的语义识别性能达到最优。最后，利用训练完成的语义分割网络对事先采集好的离线数据库进行分类，形成多种语义子数据库。在网络训练阶段与离线数据库分类阶段，本文针对不同阶段均进行了仿真实验，对网络参数变化对网络分类性能的影响与最终离线数据库的分类结果进行了详细分析。

第4章对基于语义约束的检索定位算法进行了研究。首先在在线检索阶段，在语义数据库构建完成的前提下，对基于语义与内容的快速图像检索方法进行了性能分析，对其检索速度与检索精度均进行了对比仿真。其次针对在线定位阶段中进行全局SURF特征提取导致特征提取时间较长的问题，提出了一种基于语义限制的特征点定位方法，利用语义分割网络提供的语义区域对特征点对进行了匹配与剔除，提高了定位所需特征点对的质量。最后，针对本文算法在合适的场景中进行了仿真实验，并对实验结果进行了详细分析。

第2章 室内视觉定位相关理论分析

在进行视觉室内定位的过程中，需要用到一些的与摄像机模型、坐标系转换、特征点提取以及数学几何约束求取位姿方面相关的基础知识。本章将对本文提出的创新算法中所用到的一些视觉定位算法中的基础理论进行阐述，并对算法原理进行分析，为后文的应用提供坚实的理论基础。

2.1 摄像机模型搭建

2.1.1 针孔成像模型

视觉定位系统的成像模型对人眼成像模型进行了模拟，将采集到的三维图像信息投影到二维平面，形成了二维图像。该成像模型利用了透镜成像的非线性模型，其中是针孔成像模型最为常用。在针孔成像模型下，对相机参数的求解映射到数学模型上即为对线性方程组的求解。

在搭建摄像机模型时，其核心的步骤是四个坐标系的建立，即三维世界坐标系、三维相机坐标系、二维图像坐标系与二维像素坐标系，几个坐标系之间的关系如图2-1所示。坐标系的具体建立过程如下：

（1）三维世界坐标系建立：三维世界坐标系也被称为客观坐标系。通常可以根据具体研究环境选取适当选取合适原点O_w，以右手坐标系定义三维世界坐标系方向并形成坐标系O_w-xyz。三维世界坐标系的选取需要切合表示出研究对象在空间中的具体位置。

（2）三维相机坐标系建立：三维相机坐标系在定义原点位置时用到了相机的光心，相机的光轴作为w轴，u轴与v轴分别与图像坐标系X,Y轴平行，并与相机的镜头表面相重合，构成三维坐标系O_c-uwv。

（3）二维图像坐标系建立：图像坐标系通常以相机CCD（Charge-coupled Device）图像平面的中心作为坐标原点O_1，X,Y轴分别平行于图像平面的两条对应轴，并符坐标轴合旋转关系，构建出二维图像坐标系O_1-XY。

（4）二维像素坐标系建立：像素坐标系通常以左上角顶点作为原点O_2，两个坐标轴U,V轴分别与O_1-XY中的X,Y轴平行，构建出二维像素坐标系O_2-UV。

在相机成像的过程中由于干扰、器件老化等原因一般会产生畸变，但是可以通过相机标定记性驱除。在不考虑相机畸变的情况下，真实世界的某一个点，在三维

世界坐标系中用 P 表示。点 P 经过针孔 O_c，在成像平面进行投影，则投影到了 O_1-XY 坐标系中，而后经过变换映射到了 O_2-UV 上，成为了图像中的像素点。点 P 的整个成像过程一共经历的三次转换：O_w-xyz 与 O_c-uwv 之间的转换；O_c-uwv 与 O_1-XY 之间的转换；O_1-XY 与 O_2-UV 之间的转换。三维世界坐标系中所有的点都需经过上述坐标系转换过程才可以在图像中以像素的形式进行成像，呈现在二维像素坐标系中。

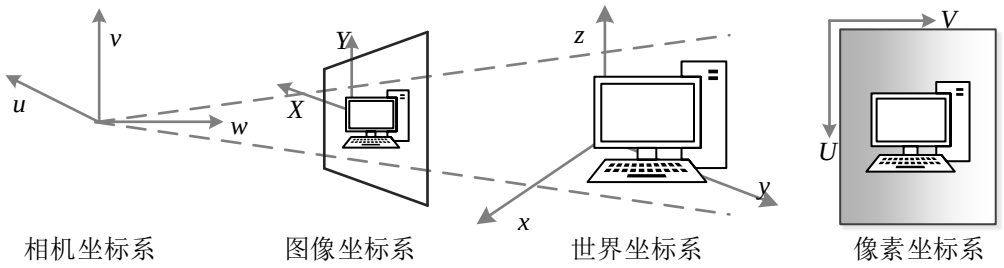


图 2-1 相机成像模型示意图

首先，先进行 O_1-XY 与 O_2-UV 之间转换公式的推导。由于 O_1-XY 和 O_2-UV 之间是通过放缩和平移变换得到的，其放缩的尺寸取决于CCD传感器的物理尺寸，则将CCD传感器的物理尺寸记为 dX 和 dY。具体的转换公式如式(2-1)所示，该公式为齐次坐标表达式。

$$\begin{bmatrix} U \\ V \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dX} & 0 & U_0 \\ 0 & \frac{1}{dY} & V_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{2-1}$$

其中，dX 与 dY 的单位为米/像素，U_0 和 V_0 为坐标系原点平移的像素距离。通过上述公式，即完成了在图像坐标系中的某一点到像素坐标系的转换。

其次，进行三维相机坐标系与二维图像坐标系之间转换公式的推导。两个坐标系之间的关系满足针孔成像模型，因此假设相机的焦距为 f，f 即为针孔与二维成像平面之间的距离。在不考虑镜头畸变影响的前提下，O_c-uwv 与 O_1-XY 之间的转换公式如式(2-2)所示：

$$z_c \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \\ w \\ 1 \end{bmatrix} \tag{2-2}$$

其中，f 的单位为米，z_c 是三维坐标系在降维成二维坐标系的过程中损失的深度信息。上述公式分别推导了 O_c-uwv 与 O_1-XY 之间的转换公式，O_1-XY 和 O_2-UV 之间的转换公式，将式(2-1)、(2-2)进行整合，即可得到三维相机坐标系与二维图像坐标系之间的转换公式，如式(2-3)所示：

$$z_c\begin{bmatrix}U\\V\\1\end{bmatrix}=\begin{bmatrix}f_x & 0 & U_0 & 0\\0 & f_y & V_0 & 0\\0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}u\\v\\w\\1\end{bmatrix} \tag{2-3}$$

其中，$f_x=f/dX$、$f_y=f/dY$，f_x 和 f_y 被称为有效焦距。

最后，进行 O_w-xyz 与 O_c-uwv 之间的转换公式推导。由于两个坐标系均是三维坐标系，因此经过一定的旋转和平移关系即可相互转换。利用坐标系间位姿变化关系，设旋转矩阵 $\mathbf{R}$，平移向量 $\mathbf{t}$，则两个坐标系之间的转换关系如式(2-4)所示：

$$\begin{bmatrix}u\\v\\w\\1\end{bmatrix}=\begin{bmatrix}\mathbf{R} & \mathbf{t}\\\mathbf{0}_3^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}x\\y\\z\\1\end{bmatrix} \tag{2-4}$$

其中，坐标系中坐标分别以其次的形式表示。图2-2更加直观的显示了 O_w-xyz 与 O_c-uwv 之间的旋转平移关系。

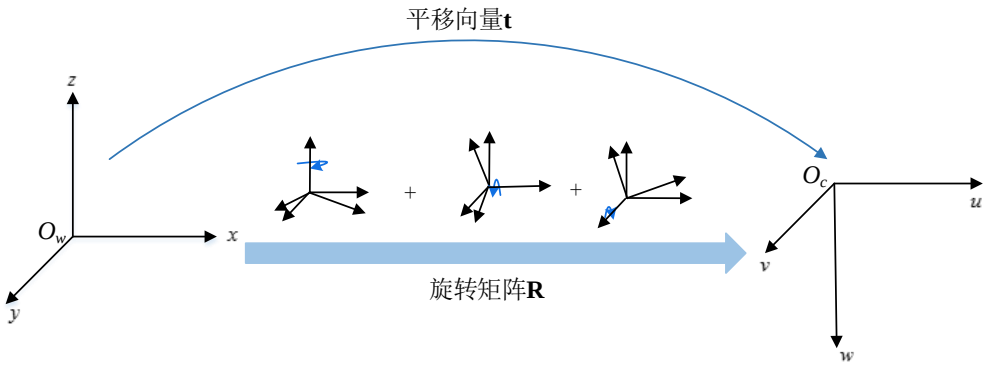


图 2-2 三维坐标系间的旋转和平移关系

如图2-2所示，坐标系之间的旋转关系主要是通过对坐标系的三个坐标轴分别依次按相应角度旋转，之后对两个坐标系坐标原点的位置进行平移，即可完成两个三维空间坐标系之间的相互转换。综上，四个坐标系之间的转换公式已经推导完成，将式(2-1)、(2-2)和(2-4)进行联立，即可找到 O_w-xyz 中三维空间点与 O_2-UV 中像素点的一一对应关系，如式(2-5)所示。

$$z_c\begin{bmatrix}U\\V\\1\end{bmatrix}=\begin{bmatrix}f_x & 0 & U_0 & 0\\0 & f_y & V_0 & 0\\0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}\mathbf{R} & \mathbf{t}\\\mathbf{0}^T & 0\end{bmatrix}\begin{bmatrix}x\\y\\z\\1\end{bmatrix}=\mathbf{KT}P=\mathbf{M}P \tag{2-5}$$

其中，$\mathbf{M}$为相机的透视投影矩阵，$\mathbf{K}$为相机的内部参数，矩阵中包含相机中心点的像素偏移信息、相机的焦距信息以及相机本身CCD传感器的物理尺寸等信息，相机的内部参数矩阵在相机出厂后一般不会随着外界条件的变化而改变。$\mathbf{T}$为相机外部参数，外部参数具体包含了O_w-xyz与O_c-uwv之间的旋转与平移关系，即旋转关系$\mathbf{R}$和平移关系$\mathbf{t}$。由于相机镜头可能存在畸变，因此在使用之前需要对相机进行标定操作。在相机标定的过程中，会对相机的畸变进行修正，同时也确定了相机的内部参数矩阵$\mathbf{K}$。如果是在三维世界坐标系确定的前提下，则在相机标定的过程中，外部参数矩阵$\mathbf{T}$可以被求出，在$\mathbf{K}$与$\mathbf{T}$均确定的情况下，$\mathbf{M}$即为已知。在相机的透视投影矩阵$\mathbf{M}$已知的情况下，三维空间中任意一点$P(x,y,z)$均能在二图像素坐标系中找到对应的像素坐标(U,V)。

2.1.2 对极几何约束

如果用两个相机或同一个相机分两次在不同位置对同一个景物进行拍摄，那么由于景物的重叠，两张图像一定含有某种隐含的对应关系。如何对这种隐含的对应关系进行求解则用到了对极几何。首先，对极几何的应用条件是两幅图像之间，它是描述两幅对应图像分别以相机光心连线为轴的平面束的交的几何。寻找隐含对应关系最直接的办法就是对像素点进行逐一匹配，然而这种匹配方法会导致搜索范围过大，匹配复杂度增加。如果对匹配点加以一定的约束条件，如对极几何约束，则搜索的范围会大大减小。为了更好地分析本文所要阐述的内容，以图2-3为例说明对极几何中的几个基本关键概念。

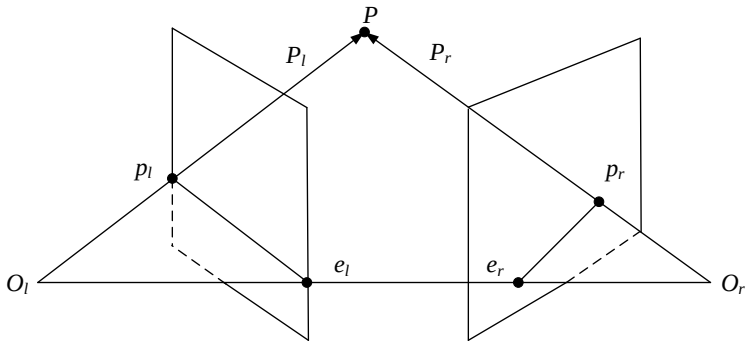


图 2-3 对极几何约束关系

1) 基线：拍摄两幅图像的相机光心连线，如图2-3中连线O_lO_r；

2) 对极平面束：以基线为轴的平面束；

3) 对极平面：包含基线的任意一个平面，如图2-3中平面PO_lO_r；

4) 对极点：基线与两幅图像的每个交点，如图2-3中点e_l与e_r；

5) 对极线：对极平面与图像平面的交线，如图2-3中直线e_lP_l与直线e_rP_r；

6) 极线约束：两根极线上的点存在的对应关系。

实际中的对极几何模型更具有普适性，是更一般的立体成像关系。在这种情况下求解两幅图像的对应关系就需要用到本质矩阵$\mathbf{E}$与基本矩阵$\mathbf{F}$。本质矩阵$\mathbf{E}$中对两个摄像机在物理空间中的旋转关系与平移变化进行了描述；基本矩阵$\mathbf{F}$则对本质矩阵$\mathbf{E}$中的所有信息进行了包含，还额外具有内部参数的隐含信息。本质矩阵$\mathbf{E}$是将左侧摄像机观测到的物理坐标与右侧摄像机观测到的点的相同位置进行了关联，而基本矩阵$\mathbf{F}$是对两个摄像机像平面上的点的像素坐标进行了关联。下面将对两个矩阵的获取与关系进行推导。

在视觉定位系统中，定义两个相机系统分别为用户相机系统与数据库采集相机系统。设P为空间中的一点，则在用户相机坐标系下，点P的三维坐标为$P_{\mathrm{l}}=(x_{\mathrm{l}},y_{\mathrm{l}},z_{\mathrm{l}})$，在数据库采集相机坐标系下，点$P$的三维坐标为$P_{\mathrm{r}}=(x_{\mathrm{r}},y_{\mathrm{r}},z_{\mathrm{r}})$。$p_{\mathrm{l}}=(U_{\mathrm{l}},V_{\mathrm{l}})$与$p_{\mathrm{r}}=(U_{\mathrm{r}},V_{\mathrm{r}})$分别为点$P$在对应相机成像平面上的像素坐标。根据上一节中推导的几个坐标系之间的转换公式，可得到如下关系：

$$\begin{cases} z_{\mathrm{l}}p_{\mathrm{l}}=\mathbf{K}_{\mathrm{l}}P_{\mathrm{l}} \\ z_{\mathrm{r}}p_{\mathrm{r}}=\mathbf{K}_{\mathrm{r}}P_{\mathrm{r}}=\mathbf{K}_{\mathrm{r}}(\mathbf{R}P_{\mathrm{l}}+\mathbf{t}) \end{cases} \tag{2-6}$$

其中，z_{l}与z_{r}是点P_{l}与P_{r}中已知的坐标位置，在公式中作为常数可以进行省略，因此可以将式(2-6)进行简化，可以得到：

$$\begin{cases} p_{\mathrm{l}}=\mathbf{K}_{\mathrm{l}}P_{\mathrm{l}} \\ p_{\mathrm{r}}=\mathbf{K}_{\mathrm{r}}P_{\mathrm{r}}=\mathbf{K}_{\mathrm{r}}\left(\mathbf{R}P_{\mathrm{l}}+\mathbf{t}\right) \end{cases} \tag{2-7}$$

由于$\mathbf{K}_{\mathrm{l}}$与$\mathbf{K}_{\mathrm{r}}$分别为不同相机内部参数矩阵，可以通过相机标定进行获取，因此可以作为已知量，现设变量s_{l}与s_{r}，令：

$$\begin{cases} s_{\mathrm{l}}=\mathbf{K}_{\mathrm{l}}^{-1}p_{\mathrm{l}} \\ s_{\mathrm{r}}=\mathbf{K}_{\mathrm{r}}^{-1}p_{\mathrm{r}} \end{cases} \tag{2-8}$$

将式(2-8)带入式(2-7)可得：

$$s_{\mathrm{r}}=\mathbf{R}s_{\mathrm{l}}+\mathbf{t} \tag{2-9}$$

对式(2-9)两侧进行平移向量$\mathbf{t}$的外积操作，可以得到：

$$\mathbf{t}\times s_{\mathrm{r}}=\mathbf{t}\times\mathbf{R}s_{\mathrm{l}} \tag{2-10}$$

对式(2-10)同时左乘$s_{\mathrm{r}}^{\mathrm{T}}$，可以得到：

$$s_{\mathrm{r}}^{\mathrm{T}}\mathbf{t}\times s_{\mathrm{r}}=s_{\mathrm{r}}^{\mathrm{T}}\mathbf{t}\times\mathbf{R}s_{\mathrm{l}} \tag{2-11}$$

根据外积定义可以得出，$\mathbf{t}\times s_{\mathrm{r}}$垂直于$\mathbf{t}$与$s_{\mathrm{r}}$，因此$s_{\mathrm{r}}^{\mathrm{T}}\mathbf{t}\times\mathbf{R}s_{\mathrm{l}}=0$。将该条件应用于式(2-8)，可以得出：

$$p_{\mathrm{r}}^{\mathrm{T}}\mathbf{K}_{\mathrm{r}}^{-\mathrm{T}}\mathbf{t}\times\mathbf{R}\mathbf{K}_{\mathrm{l}}^{-1}p_{\mathrm{l}}=0 \tag{2-12}$$

式(2-12)即为对极几何约束，将其进行简化可得：

$$p_{\mathrm{r}}^{\mathrm{T}}\mathbf{F}p_{\mathrm{l}}=s_{\mathrm{r}}^{\mathrm{T}}\mathbf{E}s_{\mathrm{l}}=0 \tag{2-13}$$

其中$\mathbf{E}$为本质矩阵，$\mathbf{F}$为基本矩阵，二者之间满足关系$\mathbf{F}=\mathbf{K}_{\mathrm{r}}^{-\mathrm{T}}\mathbf{E}\mathbf{K}_{\mathrm{l}}^{-1}$。在定位过程中，首先利用对极几何约束$\mathbf{F}$进行求解，在由二者间关系求解$\mathbf{E}$。在已知本质矩阵后，即可分解出旋转矩阵$\mathbf{R}$与平移向量$\mathbf{t}$，二者即为世界坐标系下，两个相机系统的位姿变换情况。

2.2 SURF 局部特征提取算法

2.2.1 SURF 特征点提取

在上一节中主要对对极几何约束进行了具体推导，根据推导过程可知，如果想获得拍摄两幅图像的两个相机之间的位姿关系，需要有世界坐标系中某几个点在两幅图像中对应投影点的坐标及其对应关系，因此如何只通过图像找出两幅图像中的像素坐标的对应关系则是重中之重。本文后文中阐述的基于语义约束的特征点选取方法也需要对图像中的特征点进行提取，因此需要对特征点提取算法进行分析与选取。由于室内视觉定位系统是对采集到的图像进行后续处理，该定位方法促使研究人员对图像处理的手段越来越多元化，对图像的处理手段也需要满足在各种环境条件下以及在各种干扰条件下都能够正常发挥作用。目前主流的图像处理手段为图像特征提取，根据特征提取的区域范围不同，特征提取的方法主要分为两类：全局特征提取和局部特征提取。全局特征提取方法以整幅图像作为单位，以单个向量映射图像标签；局部特征提取方法则对图像处理更加细致，将图像进行分割，对每一块区域进行特征表达。因此局部特征提取方法相较于全局特征提取方法更能反映出图像的一致性，而且如果要将两幅图像进行精细匹配，也需要进行基于

局部的特征提取方法。经过研究人员的不断分析与深入研究，目前常用且比较稳定、特征提取好的局部特征提取方法主要有尺度不变特征变换（Scale-invariant feature transform，SIFT），加速鲁棒特征（Speeded-Up Robust Features，SURF）以及定向旋转特征（Oriented FAST and Rotated BRIEF，ORB）。根据本文的实验环境场景，将选用SURF方法作为特征提取方案。

SURF局部特征算法作为SIFT算法的改进，大幅加快了特征提取的速度。此外，SURF针对噪声干扰、特征较少、有遮挡等情况也具有较强的鲁棒性，能够精准的对图像进行局部特征提取。该算法的具体流程如表2-1所示。

表 2-1 SURF 算法流程

输入：待特征提取图像 $\boldsymbol{I}$
输出：SURF 局部特征向量 $\mathbf{F}$
第一步：Hessian 矩阵生成；
第二步：高斯图像金字塔生成；
第三步：确定关键点；
第四步：确定关键点主方向；
第五步：特征描述子生成；
第六步：特征向量 $\mathbf{F}$ 生成。

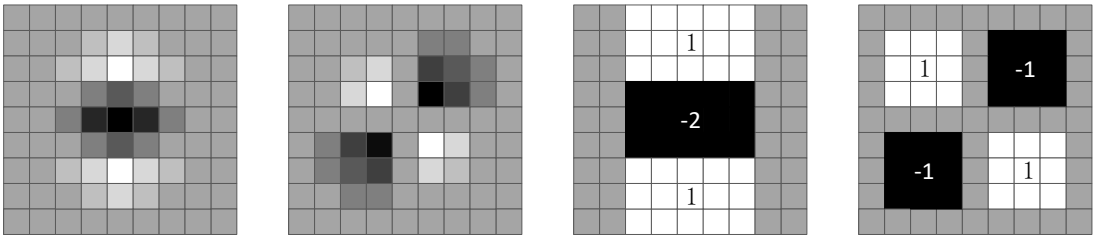
SURF特征点具有尺度不变性，图像的分辨率大小不影响其特征点的提取，为达到这一特性，需要对图像进行图像高斯金字塔构建，以满足不同尺度输入的情况。对于输入图像I，首先要进行图像平滑，主要采用了高斯滤波的方式。某一平滑后像素点$\mathbf{x}=(x,y)$，在$\mathbf{x}$处尺度为σ的Hessian矩阵$\mathbf{H}(\mathbf{x},\sigma)$的表达如式(2-14)所示：

$$\mathbf{H}(\mathbf{x},\sigma)=\begin{bmatrix} L_{xx}(\mathbf{x},\sigma) & L_{xy}(\mathbf{x},\sigma) \\ L_{xy}(\mathbf{x},\sigma) & L_{yy}(\mathbf{x},\sigma) \end{bmatrix} \tag{2-14}$$

其中，$L_{xx}(\mathbf{x},\sigma)$为图像$I$中$\mathbf{x}$处的高斯二阶导数$\partial^2/\partial x^2\, g(\sigma)$，同理$L_{xy}(\mathbf{x},\sigma)$和$L_{yy}(\mathbf{x},\sigma)$与其意义相似。Hessian矩阵行列式在$\mathbf{x}$处的近似响应值为：

$$\det(\mathbf{H}_{approx})=D_{xx}D_{yy}-(0.9D_{xy})^2 \tag{2-15}$$

原图像的尺度变换空间即为将求解的近似响应值代替原图像中的每一个像素值。在实际应用中，可以用一组积分模板进行代替上述复杂操作过程，其模板形式如图2-4所示。



(a) *yy* 方向　(b) *xy* 方向　(c) *yy* 方向　(d) *xy* 方向

图 2-4 积分模板示意图

其中，图2-4(a)、图2-4(b)为替代前不用方向的二阶导数；图2-4(c)、图2-4(d)为代替后不用方向的二阶导数。利用积分模板进行替代操作之后，即可进行图像金字塔的构建，以积分模板的形式同样也加快了SURF算法的计算速度。

在图像金字塔构建完成之后，SURF特征点提取的核心操作为非极大值抑制。在图像金字塔中，该算法将每一个像素点的近似响应值与其三维邻域的像素响应值进行比较，并对非极值进行抑制，对响应值是极大值或极小值的，则作为特征点进行提取，其算法示意图如图2-5所示。

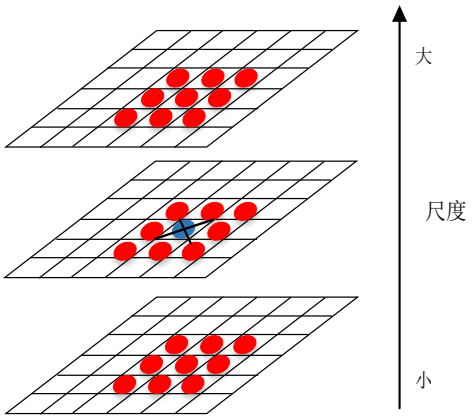


图 2-5 SURF 关键点非极大值抑制示意图

如图2-5所示，图中最中心点为选取像素，周围26个点即为选取像素点的三维邻域，特征点的选取即为对选取点及其邻域的特征点响应值进行分析，其尺度方向即为尺度变化方向，图中为选取像素点的三层尺度变化。关键点的主要信息为位置与方向，通过利用harr小波的方式对水平与垂直特征值进行计算，关键点的主方向为特征总和最大的方向。综上所述，对图像进行特征点提取首先需要对图像进行图像金字塔构建，再对特征点及其三维邻域进行非极大值抑制操作，最后通过harr小波特征值确定特征点的主方向。在特征点的位置与主方向都确定之后，即可进行特征向量的构建并完成特征点匹配。下一小节将对SURF特征点匹配进行详细分析。

2.2.2 SURF 特征点匹配

得到两幅图像的特征点之后，为了寻找两幅图像中对同一个三维空间点的映射值，需要对提取的SURF特征点进行匹配操作。SURF特征点的匹配操作主要是通过将两组特征点的64维特征描述向量进行欧式距离比对完成的。

在上一节中，特征点的位置与方向均被确定，由于SURF特征拥有旋转不变的特性，因此需要确定坐标轴。之后，需要在关键点的周围形成变长为20s（s代表关键点当前所在尺度大小）的正方形框，并将其划分为大小相等的4×4个子区域。对每个子区域中的像素点进行遍历操作，对其求取harr小波水平、垂直朝向特征值也特征值绝对值之和，最终会生成64维特征向量。该特征向量即为SURF特征描述子，其生成过程如图2-6所示。

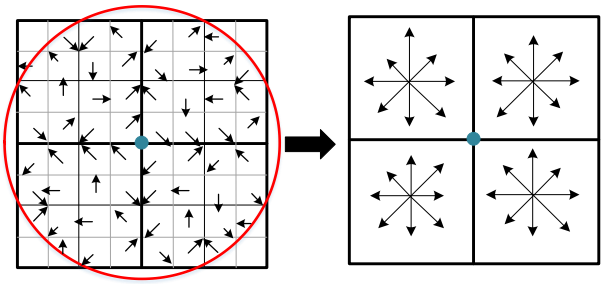


图 2-6 生成 SURF 局部特征描述子

如图2-6所示，中心点即为选取的关键点，将其周围分割成16个子区域，根据像素遍历得到特征向量。最后，通过对两幅图像中特征向量进行欧式距离计算，如果欧式距离小于一定阈值，即可判断两个特征点相互匹配，匹配结果如图2-7所示。

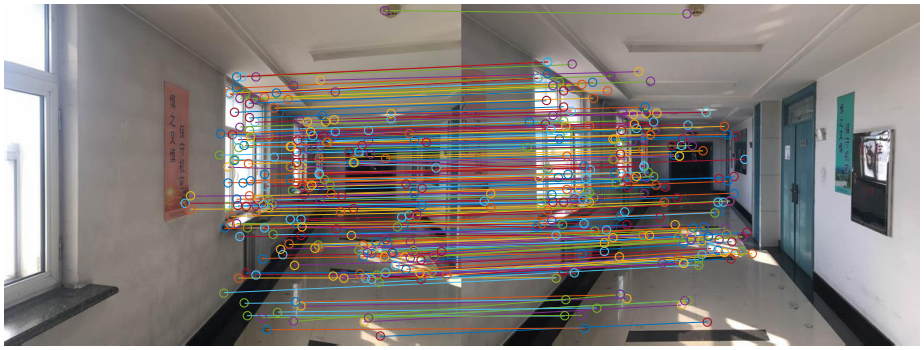


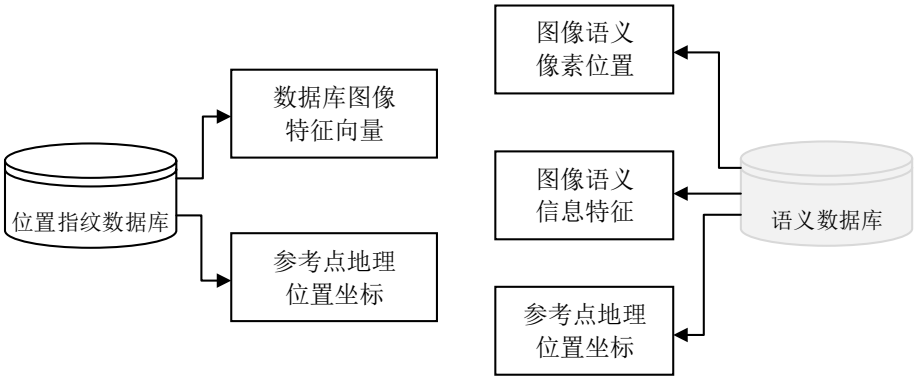
图 2-7 图像 SURF 局部特征描述子示意图

如图2-7所示，左右两张图像分别为在不同位置拍摄的同一场景图像，图中用连线表示匹配到的特征点对。从图中可以看出，虽然是两个不同位置拍摄的图像，但是由于图像中为同一场景，存在许多重叠内容，SURF特征点提取方法将重复内容中的特征点进行了匹配，匹配结果较为准确。

2.3 基于语义的室内视觉定位流程

2.3.1 基于语义的离线数据库建立

本文研究的基于语义的室内视觉定位系统与传统的室内视觉定位系统类似，同样分为两个阶段：离线与在线。但本文分别对两个阶段的算法进行了改进优化，离线阶段的优化旨在利用语义数据库代替传统的检索离线数据库，以减小数据库增大所带来的在线阶段检索的时间开销；在线阶段的优化旨在利用高精度高效检索方法，找到待检索图像的最佳匹配图像，并利用语义区域的优势在保证定位精度的情况下提高定位效率。传统的室内视觉定位系统中用来建立离线数据库的方法主要为位置指纹法，该方法同样从WiFi定位演进而来。位置指纹法通过在待定位环境中不断采集图像，并将此时拍摄图像的位置与图像同时绑定存入以建立离线数据库，数据库中包含的不仅仅为图像特征信息，还有此时的用户地理位置信息。在在线定位阶段，通过检索算法找到用户输入待定位图像在离线数据库中的匹配图像，根据匹配图像的地理位置信息大致推算用户的粗略位置，再用定位算法进行精确定位返还给用户，完成定位服务。但是该方法具有一定的局限性，随着目前室内场景的增加与室内面积的不断扩大，室内场景变得越来越复杂化，因此想到达到精确定位需要采集稠密的离线数据库。随着离线数据库容量的增大，由于检索算法需要对离线数据库中所有图像特征进行遍历才能找到匹配图像，因此在线检索的时间开销也随之增大，影响了用户的体验。为了解决上述问题，本文提出了一种基于语义的离线数据库分类方法，用语义数据库代替传统的离线数据库，旨在将离线数据库根据语义进行分类，避免了检索阶段对数据库全局进行遍历检索的缺陷，减少了时间开销。图2-8为传统室内视觉定位系统的离线数据库与本文所提出基于语义室内视觉定位系统的语义数据库中存储信息比较示意图。



(a)传统数据库　　(b)语义数据库

图 2-8 两种离线数据库存储信息比较示意图

图2-8(a)中为传统数据库中所包含的信息，图2-8(b)中为本文提出的语义数据库中所包含的信息，语义数据库包含的信息更加全面。本文用所提出的基于语义与内容的快速图像检索（Semantic and Content-Based Image Retrieval，SCBIR）方法对待定位图像进行了精确匹配，在保证检索精度的同时提高了检索效率。其次针对匹配图像，利用了基于语义数据的特征点定位方法，进行精确位置查找，有效地提高了定位效率。本文研究的基于语义的视觉定位系统的流程图如图2-9所示。

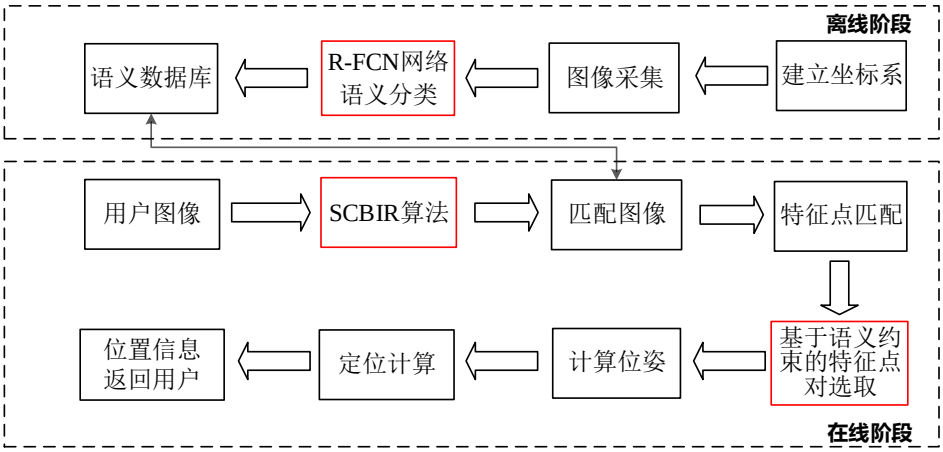


图 2-9 基于语义的室内视觉定位系统技术路线图

在离线阶段，首先要进行坐标系建立，由于采集的图像需要包含位置坐标信息，要求图像采集时坐标系已经建立并保持不变。在图像采集时，应保证采集图像完整覆盖待定位图像，采集图像间隔尽量保持一致并保证质量。待图像采集完成之后，此时已经得到了类似传统室内视觉定位系统中的离线数据库，将数据库中的每一张图像均输入到语义分割网络中，得到语义分割结果，根据语义分割结果将图像进行分类，对分类后每一类的图像进行结构与颜色特征提取，即完成了图像语义数据库的构建。在本文的语义数据库构建方法中，只需按传统的图像采集方式进行图像采集，语义分割提取等均为后续批量操作，减少了数据库建立过程中的复杂程度。在图像特征提取方面，由于语义提取包含语义特征，语义数据库分类较为精确，因此只针对语义数据库中的图像提取的颜色与结构特征，替代了传统方式的SURF特征提取，减小了特征提取阶段的复杂度与时间开销。在第3章，本文将对提出的基于语义的离线数据库分类方法进行详细阐述。

2.3.2 基于语义的在线检索定位

在线阶段，本文首先提出了一种基于语义与内容的快速图像检索算法，即SCBIR算法。对用户输入的待定位图像同样进行语义分割，将其定位到语义数据库

中的某一语义子数据库内，在语义子数据库内对其进行基于内容的图像检索，找到匹配图像。该方法的优势在于，传统定位系统中的检索算法针对整个数据库进行遍历检索具有盲目性，无法确定用户的起始位置，导致定位系统启动时间较长。而SCBIR算法根据图像中的语义组合信息确定了用户的起始位置，将检索范围大大减小，减少了在线阶段的检索时间，节省了定位系统的启动时间，提升了定位系统效率与用户体验。此外，在定位阶段本文提出了一种基于语义约束的特征点定位方法，利用语义分割网络输出的图像中语义的像素位置信息对图像中的特征点进行剔除与选择，以选取更优质的特征点进行定位。该方法在保证定位精度的同时有效地提升了定位效率。值得注意的是，在线阶段的基于语义与内容的快速检索算法和基于语义约束的特征点定位算法都是凭借语义分割网络的输出结果来提高定位系统的整体效率的。网络中输出的语义主体在图像中的像素位置矩形框不仅能根据语义对应剔除误匹配的特征点，还能以语义约束的形式减少特征点的提取区域，节省了在线阶段大量的时间开销，给用户更好的实时性体验。在线阶段的检索过程与定位过程将在第四章中进行详细的分析与实验验证。

综上所述，本文研究的基于语义的室内视觉定位系统分在两个阶段均进行了改进。在离线阶段，将离线数据库进行分类形成语义数据库，能够明确缩小在线阶段检索时进行检索的范围，对待定位用户的初始位置进行精确估计，有效地减小了随数据库容量增大而线性增长的检索时间开销。在在线阶段，首先通过基于语义与内容的图像检索算法对待定位图像的匹配图像进行了查找，再对两幅图像进行特征点的提取与匹配，利用语义分割网络输出的语义主体像素位置对匹配特征点对进行剔除与选取，将优质特征点对利用对极几何约束求解出两幅图像的位姿关系，从而进行精确定位，在保证了定位准确度的前提下提高了定位系统的效率。

2.4 本章小结

本章对本文算法需要的基础理论进行了详细阐述。首先，本文对摄像机成像模型进行了搭建，对针孔成像模型进行了相机叙述，并对基本的定位方式对极几何约束公式进行了推导。其次，对于对极几何约束所需要的匹配点对的由来进行了说明，并对适合本文环境的SURF特征点提取方式进行了原理叙述，分析了SURF特征提取算法的优良性能。最后，对本文提出的基于语义的视觉定位系统总体流程进行了分析，那个对离线与在线阶段的改进模块进行了分析，并将本文针对于传统定位系统的改进方面与优势进行了说明，为后续研究提供了充足的理论依据。

第3章 基于语义的离线数据库分类算法研究

在室内视觉定位系统的两个阶段中，离线阶段是一个图像数据采集并建库的过程，即利用图像采集设备将待定位区域的场景图像进行稠密存储，并附加地理位置信息。在传统建库过程中，为保证在线检索与定位阶段的准确性，要求离线数据库对待定位场景进行稠密覆盖。然而随着采集图像的增加，数据库容量会相应增大，导致在线检索阶段的检索时间相应变长，影响在线定位的实时性。为了解决这个问题，本文提出了一种基于语义的离线数据库分类算法，将传统离线数据库进行分类转化为语义数据库，有效地缩小了在线检索阶段需要进行特征比对的检索范围。

3.1 语义分割网络搭建

3.1.1 语义分割网络框架

在视觉定位中，离线图像数据库的稠密程度决定着在线定位阶段的定位精度。随着人们对定位精度要求的不断提高，其所依赖的图像数据库容量也越来越大，使得在图像检索过程中花费的时间越来越长。因此，本文提出了一种可标注多标签的基于语义的数据库分类算法，来提高分类的速度与精度，该方法主要应用到了机器学习中的语义分割网络。目前语义分割网络主要应用在目标检测领域中，其核心算法是基于区域的全卷积网络（Region-based Fully Convolutional Networks，R-FCN）算法。本文提出的基于语义的离线数据库分类算法的主要思想是利用目标检测原理，对离线数据库中的图片采用语义检测的方法，并针对每张图像的语义检测结果，对该图像打上不同的多种语义的标签。最终将相同语义排列的标签的图像分为一类，将离线数据库进行分类转化为语义数据库，以达到缩小检索范围、减少在线检索阶段检索时间开销的效果。下面将对R-FCN进行详细分析。

R-FCN主要由全卷积网络（Fully Convolutional Network，FCN）、候选区域生成网络（Region Proposal Network，RPN）和ROI（Region Of Interest）子网络三个部分组成。其中FCN主要用于提取特征，RPN网络根据提取的特征与事先在图像中标注的语义区域生成ROI，ROI子网以FCN提取的特征与RPN输出的ROI作为输入，并对ROI内的成分进行语义分类。R-FCN的目标检测分为两个步骤。首先，先进行目标位置的确定；其次，对确定了位置的目标进行具体类别的分类。R-FCN算法首先利用了全卷积网络生成特征映射图，并利用候选区域生成网络RPN对生成的特征映射图进行全图的前景目标搜索和筛选，以确定目标的边框；在这个基础上，利用分类的网络对选定的目标进行分类识别，R-FCN构架流程图如图3-1所示。

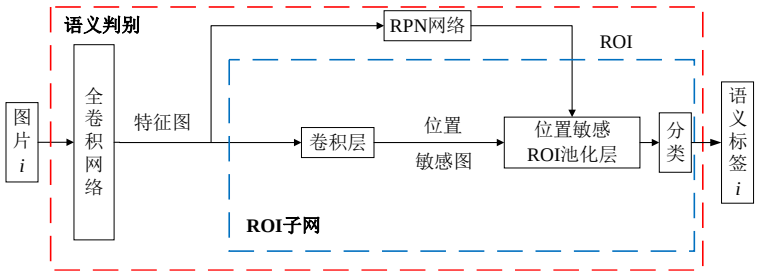


图 3-1 R-FCN 构架流程框图

R-FCN 的输入为图像信息，首先图像经过全卷积网络，利用不同卷积核对其深层特征进行提取，得到图像的特征图信息。图像的特征图中包含着图像中各种边缘、纹理、颜色以及隐含特征信息。其次，将卷积得到的图像隐含特征作为输入，经过RPN网络。RPN网络利用图像特征以及先验的标注信息，会生成候选区域矩形框，即语义目标所在位置信息。最后将特征图与候选区域信息一起作为输入，输入到ROI子网中，ROI子网将会对RPN输出的多个候选区域框进行语义信息判定，并对特征图进行卷积操作，得到位置敏感分数图，对每一个候选区域生成框进行分类，给出语义归属。

3.1.2 语义分割子网分析

本文在R-FCN前段的全卷积网络使用了ResNet网络，ResNet全称为 Deep Residual Network，即深度残差网络，为解决网络中的“退化”问题而被提出。“退化”问题即为当网络模型层次加深时，可能会发生梯度弥散甚至梯度爆炸，导致错误率提高的一种现象。其运用残差的方法较好的解决了该问题，保证了该网络随着网络层数的加深，其性能不会随之下降。该网络输出一个三维向量，是由多张feature map组成，其中每一张 feature map 都代表了原图像上某个层次的特征。在最终输出的 feature map 中并没有舍弃图像中的空间信息，因此该网络对精确的分类任务提供了很大帮助。其方法原理图如图3-2所示。

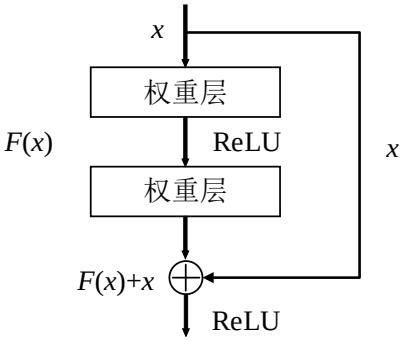


图 3-2 ResNet 网络原理图

如图3-2所示，如果在深层神经网络中，后续每一层之间没有累积误差与数据损失，则根据恒等估计可将其转化为一个非深层网络，但直接拟合恒等关系 $H(x)=x$ 比较困难。如果把网络设计为 $H(x)=F(x)+x$，则可以对优化方式进行转换，可以转换为学习一个残差函数 $F(x)=H(x)-x$。只需让 $F(x)=0$，即可完成恒等映射 $H(x)=x$，减少了残差拟合的复杂度。 ResNet 同样是全卷积网络，它可以提取处图像内部包含的隐藏信息，如纹理特征、颜色特征、边缘特征等。目前，ResNet网络可以分为50层、101层、152层等等，经过实验验证，这几种网络针对本实验数据正确率差别不大，因此选用了层数最少的ResNet-50网络。

R-FCN网络使用了ResNet-50网络对图片进行卷积、池化等操作，该操作最后生成了一个 $W\times H\times 1024$ 的feature map。W 与 H 不是训练图片的宽高尺寸大小，但是与原图的宽高尺寸有一定的比例关系。ResNet-50网络的输出层是一个三维向量，RPN在ResNet-50的输出层上完成了候选区域的搜索。其搜索形式是利用512个大小可以调整的卷积核对输出层进行卷积操作，其卷积核大小一般为 $3\times 3\times 1024$，最终会输出一个 $W\times H\times 512$ 的三维向量。其卷积后的尺寸未发生改变，仍为 W 与 H 是因为在卷积的过程中对原图像的边界以0进行了填充。以输出层向量为 $224\times 224\times 1024$ 为例，其卷积后结果为 $224\times 224\times 512$ 的向量；随后将每一个单独的512维向量作为RPN中两个独立卷积层的输入，从而将特征映射图中的信息转换为候选区域的位置信息和前后景的概率信息。图3-3所示为RPN示意图。

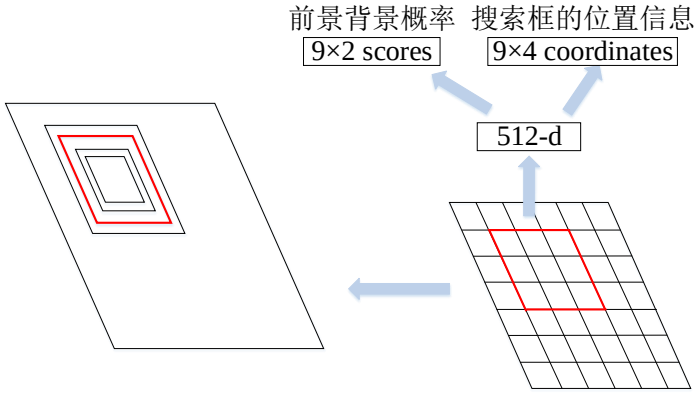


图 3-3 RPN 网络示意图

在图3-3中，scores表示该基准矩形框内为前景或者背景的得分；coordinates表示生成的基准矩形框距离标准的矩形框的位置修正参数。9表示在RPN对原有FCN的输出层进行卷积生成的新feature map中，一个点对应着原输出图像上的9个基准矩形框，其维度是网络事先规定的，维度大小分别为128、256、512；矩形框的比例为1:1、1:2、2:1。最终，一共得到9个基准矩形框。而每个基准矩形框都会得

出4个修正参数t_x、t_y、t_w、t_h。利用这4个修正参数可以对基准矩形框进行修正即可得出候选区域，修正公式如下所示：

$$\begin{cases} x = w_a t_x + x_a \\ y = h_a t_y + y_a \\ w = w_a \exp(t_w) \\ h = h_a \exp(t_h) \end{cases} \tag{3-1}$$

式中，x、y、w、h表示候选区域的中心横坐标、中心纵坐标以及矩形框的宽度与高度，x_a、y_a、w_a、h_a表示基准矩形框的相应四个参数。

对于原始的输入图片，RPN网络会得到约两万个搜索框，对下一步的分类来说候选区域太多会造成负担过大。因此需要对搜索框进行筛选和删除。如果搜索框超出图片的边界，那么判定该搜索框是无效的，对其进行删除操作。此外，如果对于同一目标存在重叠覆盖的搜索框，且搜索框的大小与置信概率均不同。该情况可以采用非极大值抑制（Nonmaximum suppression，NMS）方法进行处理，以达到将重叠的搜索框删除的效果，并显著提高搜索框的搜索效率。NMS方法是将通过RPN网络得出的搜索框进行置信程度排序，其置信程度是根据前景和背景的得分情况确定的。NMS方法将置信程度最大的候选框选取出来，并将与该候选框有重合部分的候选框全部选取，进行IoU的计算。IoU（Intersection over Union）是测量矩形框位置准确率的标注，其经常用Jaccard系数来进行评估：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3-2}$$

其中A，B分别代表预测像素范围与真实像素范围。在训练过程中，根据实验数据规定Jaccard系数的范围在[0.7,1]为正样本，表示的为语义的主体部分；Jaccard系数的范围在[0.1,0.3]为负样本，表示的为背景类。由于实验场景与拍摄步长选取的约束性，本文随机选择比例为1:1的正负样本作为ROI子网中位置敏感ROI池化层输入。

当选择的置信程度最大的候选区域与与其重叠的候选区域的IoU大于设定的阈值时，此时就将该置信度较小的候选区域进行删除；当选择的置信程度最大的候选区域与与其重叠的候选区域的IoU小于设定的阈值时，该候选框就进行保留。当这一组中所有候选框操作完成后，选择除该组之外置信程度最大的候选框重复该操作，最终剩余的候选矩形框的数量就会大量较少，并且其候选矩形框的准确程度都很高。RPN网络的作用就是为之后的ROI分类网络生成较为少数的高质量候选矩形框以方便其进行准确分类。

ROI子网即为分类网络，该网络同样也是对FCN网络中ResNet-50网络输出的feature map上行卷积操作。ROI子网利用了 $k\times k\times(c+1)$ 个1×1×1024卷积核，将ResNet-50输出的三维 $W\times H\times 1024$ 的feature map卷积生成了新的 $W\times H\times k^2(c+1)$ 的feature map。其中，k代表要将RPN生成的候选区域矩形框的行和列分成几等分，一共分成 k^2 块。c 表示要训练的图像集中一共有多少类的目标。由于除了想要分类的目标外，还有背景层，因此，训练的类别一共有 $c+1$ 类。生成的新的feature map被称为位置敏感分数图，该分数图为一个三维向量。如果对新生成的位置敏感分数图进行细化分类，$W\times H\times k^2(c+1)$ 分别表示为，每一张位置敏感分数图的大小为 $W\times H$，每一个子块中有 $c+1$ 张位置敏感分数图，整个的位置敏感分数图中一共有 k^2 个字块。本课题中令 $k=3$，因此，ROI区域被平分为9块，位置敏感分数图同样也拥有9个子块，如图3-4所示。

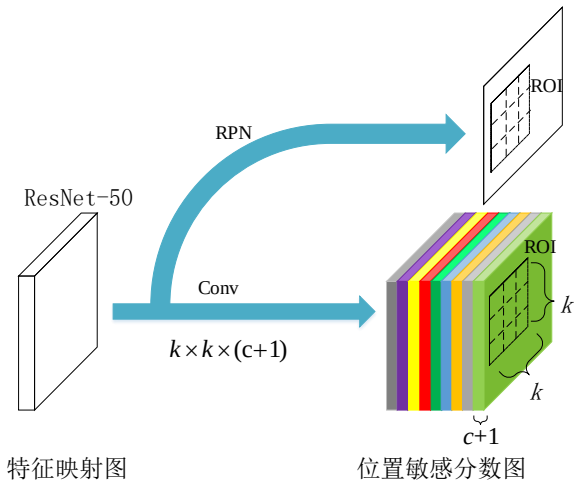


图 3-4 位置敏感分数图

ROI 子网利用卷积操作在输入图像上为每一个类别生成 $k\times k$ 个位置敏感分数图，描述了空间网格位置，其对应的位置分别为ROI区域的9个位置。每一个位置敏感分数图上的值，代表了该空间位置上为该种类元素的得分情况。每个位置敏感图有 $c+1$ 个通道输出，对于RPN输出的 $R\times S$ 尺寸的ROI，将目标框划分为 $k\times k$ 个大小为 $R\times S/k^2$ 的子区域，该子区域中包含有多个位置敏感分数。由于过多数据会对后续分类操作形成干扰，因此需要用池化操作对数据进行下采样。对任意一个子区域 $bin(i,j),0\le i,j\le k-1$，定义池化操作：

$$r_c(i,j\mid\Theta)=\sum_{(x,y)\in bin(i,j)}\frac{1}{n}z_{i,j,c}(x+x_0,y+y_0\mid\Theta) \tag{3-3}$$

其中，$r_c(i,j|\Theta)$是子区域$bin(i,j)$对c个类别的池化响应，$z_{i,j,c}$是子区域$bin(i,j)$所对应的位置敏感分数图，(x_0,y_0)代表ROI左上顶点像素坐标，n是子区域$bin(i,j)$中的像素数目，Θ代表了网络的所有学习所得到的参数。位置敏感池化的操作框图如图3-5所示。

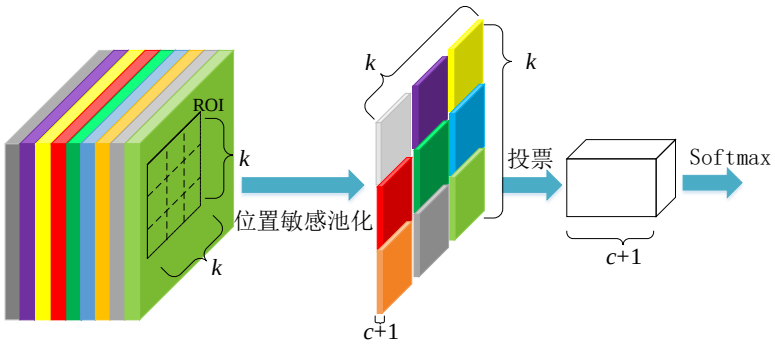


图 3-5 位置敏感池化框图

经过位置敏感池化之后，原来$R\times S/k^2$大小的每个子区域变成了一个值，对于每一类来说，ROI被划分为的9块子区域变成了9个位置敏感分数，分别代表了该位置对应该类别9个空间方位的得分。最终，计算$k\times k$个子区域池化响应输出$r_c(i,j|\Theta)$的均值，将ROI池化层输出的c+1维特征按维度求和得到一个c+1维的向量：

$$r_c(\Theta)=\sum_{i,j}r_c(i,j|\Theta) \tag{3-4}$$

下一步，将这个向量代入到多项逻辑斯蒂回归（Softmax）公式，就可以利用Softmax回归类方法获得该搜索框中的目标属于每个类别的概率，并按照目标所属的最大概率将其归类：

$$s_c(\Theta)=e^{r_c(\Theta)}/\sum_{c'}^{c}e^{r_{c'}(\Theta)} \tag{3-5}$$

伴随每一个语义种类同时输出的还有一个四维向量，记作$\{x,y,w,h\}$，其分别表示当前语义ROI区域的中心横坐标、中心纵坐标以及矩形框的宽度和高度。在训练过程中，为了确定网络训练时的准确程度和最佳迭代次数，需要设置相关的损失函数。语义检测预测了网络语义区域与实际语义区域之间的损失最小化，在网络中通常会采用随机一度下降法去更新网络参数，使损失函数最小化。损失函数L由分类损失函数L_{cis}和位置损失函数L_{reg}组成：

$$L(s,t_{x,y,w,h}) = L_{cis}(s_{c^*}) + \lambda \text{sign}(c^*) L_{reg}(t,t^*)$$
$$\text{sign}(c^*) = \begin{cases} 1 & c^* > 0 \\ 0 & else \end{cases} \tag{3-6}$$

其中，c^* 代表 ground truth，$c^* > 0$ 表示分类正确。λ 表示超参数，表示分类损失和位置损失的重要性差异，用以调整两种损失的重要程度。t 表示预测的语义区域位置，t^* 代表了 ground truth 的真实位置。

分类损失函数用交叉熵损失表示为：

$$L_{cis}(s_{c^*}) = -\log(s_{c^*}) \tag{3-7}$$

位置损失函数用了平滑 L_1 损失函数[45]来控制梯度大小：

$$L_{reg}(t,t^*) = \sum_{i \in \{x,y,w,h\}} S_{L_1}(t_i, t_i^*) \tag{3-8}$$

平滑 L_1 损失函数公式为：

$$S_{L_1}(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & else \end{cases} \tag{3-9}$$

为了研究语义信息被遮挡或残缺等情况，本文采用了OHEM（Online Hard Example Mining）络[46]进行训练。在训练期间，该网络释放的正负样本的约束，将负样本的阈值调整为0，并取消了正负样本的比例。

3.2离线数据库分类算法研究

3.2.1 数据库分类基本流程

在离线阶段，首先进行图像采集工作，使用视觉存储设备对待定位环境的视觉特征以图像的形式进行存储，建立待定位区域的图像数据库。其次，需要对数据库中图像进行特征提取或者图像分类工作，形成新的特征数据库或子数据库，方便在在线阶段进行高效快速的检索。

针对特征提取算法，假设在待定位区域采集了 m 张图像，每一张图像用 I 表示，则图像数据库可以表示为 $P = [I_1, I_2, \cdots, I_m]$。如果之后对图像数据库 P 进行了特征提取，则形成的特征数据库可以表示为 $F = [f_1, f_2, \ldots, f_m]$，其中 f_i 为从图像 I_i 中采取某种特征提取方法生成的特征向量，$f_i = \mathcal{F}(I_i), \forall i = 1, \cdots, m$。将二者进行匹配组合，即可形成在线阶段所需的离线检索数据库 D，可以表示为：

$$D=[d_1,d_2,\cdots,d_m]^{\mathrm{T}} \qquad \forall i=1,\cdots,m \tag{3-10}$$

其中,$d_i=[I_i,f_i]$ 表示为每一张照片的像素数据与特征数据的组合。因此随着离线数据库的图像存储数量增多，其特征与像素的维数也随之增多，大大降低了在线阶段的检索速度。

针对图像数据库分类算法，假设按照某种方法将其分为了 k 类，原始图像数据库可以表示为 $P=[P_1,P_2,\cdots,P_k]$，其中 $P_i,\forall i=1,\cdots,k$ 为原始数据库分类形成的子数据库，并满足条件 $\sum_{i\in k}P_i=P$。检索时先将待检索图片的检索范围缩小到子数据库范围，之后再对子数据库按合适的精确检索方法进行详细检索。因此该方法检索效率的提升在于能否对原数据库进行准确细致的分类，子数据库占原数据库的百分比越小，则检索效率越高。本文所提出的离线数据库分类算法即将离线数据库进行细致准确分类，将待检索图像直接定位到缩小检索范围的语义数据库中，提高整个视觉定位系统在线阶段的检索效率，给用户更好的实时性。基于语义的离线数据库分类算法的基本流程如图3-6所示。

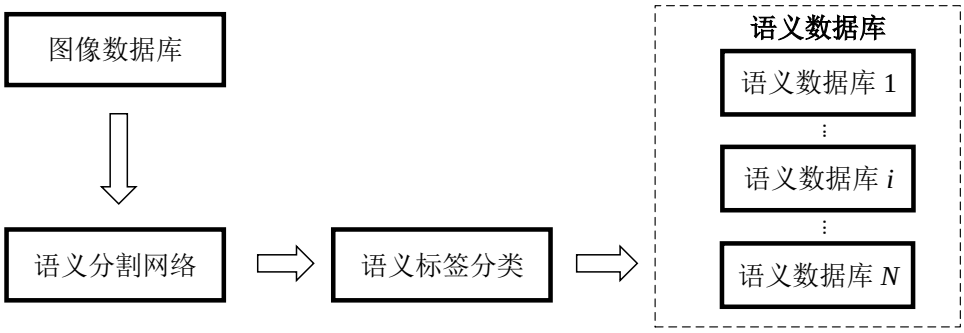


图 3-6 基于语义的离线数据库分类算法流程

由图3-6可以看出，基于语义的离线数据库分类算法的核心为将图像数据库分类为语义数据库的过程，该过程可以大大减少传统检索算法的检索范围，减少第一步检索所用的时间。首先将离线阶段构建的离线数据库中的图像作为语义分割网络的输入；其次，语义分割网络对于输入的每一张图像进行语义识别，识别的结果过图像中的每种语义类别和该类语义在图像中的像素位置；最后，利用图像所属的多种语义类别，根据语义的排列组合形式将图像进行分类形成语义数据库。在语义数据库中，可以采用一些拥有高复杂度但有较高精度的检索方法检索出最匹配输入待检索图像的结果。第二步的精确检索过程以精度为前提，由于其搜索范围较小，因此不会有过高的时间开销。将图像数据库转化为语义数据库的过程如图3-7所示。

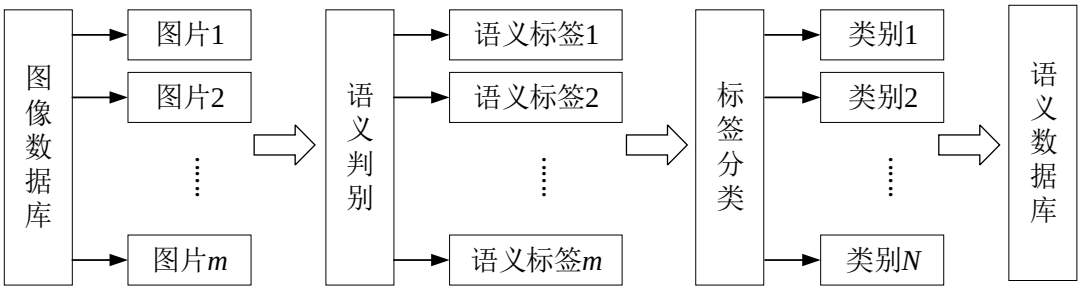


图 3-7 图像数据库分类算法流程图

在将原始图像数据库分类成语义数据库的过程中，考虑到分类类别的个数对分类算法复杂程度的影响以及对子数据库中图像检索速度影响，本文主要对将图像数据库划分为语义数据库的阶段进行了算法改进，对数据库中的每张图像进行了语义特征的提取，并构成了一种基于语义的图像数据库。

3.2.2 语义数据库构建

假设对待定位环境中规定的语义共计 c 类（不包含背景类），定义总语义库为 $S=[S_1,S_2,\cdots,S_c]$，其中 $S_i,\forall i\in c$ 表示对应的语义。对于每一张图像，可能图像中的内容包含多种语义，因此对每张图像上的类别分类用到了不同语义的组合。根据排列组合可以得出，如果将待定位区域中分为 c 类语义，那么得到的语义组合种类数目为 $N=2^c$。最终，可以得到的语义数据库 $D_S=[D_S^1,D_S^2,\cdots,D_S^N]$，其中 $D_S^i,\forall i\in N$ 为语义子数据库，是对图像离线数据库进行语义判别后将有相同语义的图像进行集中形成的数据库。虽然语义组合的种类数目与定义语义数目成指数关系增长，但是就实际问题而言，大量语义基本分布在较广的范围中，在某一区域集中出现的概率极低，因此一些语义子数据库 D_S^i 中，不存在相应的分类图像，这使得分类复杂度大大降低。

图像经过语义分割网络后，图像中的语义信息均以标签的形式与图像绑定。对于输入图像 I_{input}，如果图像中含有语义成分，其经过语义分割网络后绑定的语义标签可能为 $S_{input}=[S_1,S_2,\cdots,S_k]$，其中 $1\le k\le c$。最终每个图像绑定的语义标签会出现各种语义排列组合的情况，因此本文定义语义判别向量 $\Omega=[\omega_1,\omega_2,\cdots,\omega_c]^{\mathrm{T}}$，其中

$$\underset{i\in c}{\omega_i}=\begin{cases}1 & S_i\subset\varnothing\\ 0 & else\end{cases} \tag{3-11}$$

对于每一张输入图像 I_{input}，均有对应的语义标签 S_{input} 与语义判别向量 Ω_{input}。由于语义标签中语义成分过多，其排列顺序也会相应延长检索时间，因此本文针对

于每一张图像，将其中包含的语义信息转化成相应的数字标签。定义转化向量 $\Lambda=[2^0,2^1,\cdots,2^c]$，即可将多种语义标签转化成的唯一数字标签：

$$l=\Lambda\cdot\Omega \tag{3-12}$$

因此，语义分割网络输入端的每一张图像 I_{input} 在输出时均附带一个数字标签 l_{input}，该数字标签代表了图像中的各种语义组合成分，也是数据库分类中的分类标记。将离线数据库中的每一张图片均输入到网络中，将输出数字标签 l_{input} 相同的图像归为一类，形成语义子数据库 D_S^i。在 D_S^i 构建完成之后，对其中的每一张图像进行颜色与结构特征提取，并以向量的形式保存在数据库中，待检索图像查询时使用。

3.3 基于语义的离线数据库分类算法性能分析

3.3.1 语义分割网络训练数据库构建

在完成网络搭建之后，需要采集足够的图像来构建机器学习的训练库，本文训练集的环境取自哈工大科学园2A栋12楼走廊，这是一个典型的办公环境，采集区域在该场景平面图中用深色区域表示，如图3-8所示。

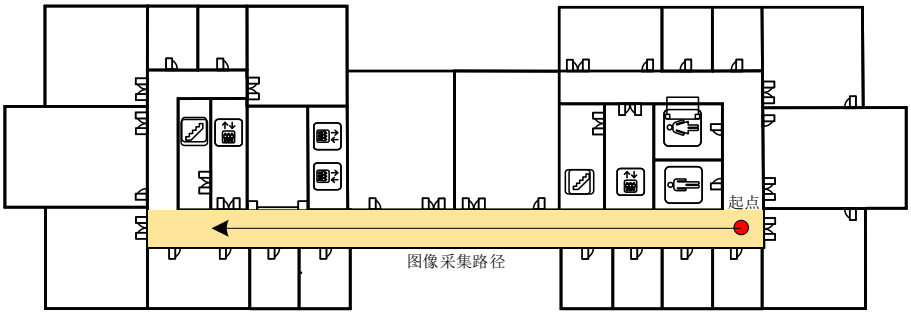


图 3-8 实验采集区域平面示意图

本实验以0.5m作为离线数据库数据采集间隔，并从正反双向分别采集数据。拍照时不要求镜头光轴与走廊中轴线平行。遇到语义场景时，进行多角度的拍摄，以便对该语义进行准确分类；遇到多语义场景时，缩短拍照间隔，使训练集中语义场景更丰富；遇到走廊中出现次数较少的语义目标时，例如通风口、消防栓、垃圾桶、安全出口标识等，应对其进行多次拍摄，以增加出现频率较少的语义在训练集中出现的次数，提高分类准确率。此外，对于门、窗等有多种状态（开、关）的语义，需要对其的不用状态分别进行采集与标注，以便在任何状态中都会对该种语义进行准确识别。最终将采集到的800张图片作为训练数据库输入到语义分割网络进行训练，训练数据集中样本如图3-9所示。

图 3-9 训练数据集的样本

训练数据集采集完成之后，需要其中每一张图像进行人工语义标注，以保证训练时 ground truth 的准确性。本文将走廊内的语义（除背景类）分为了9类，分别为：门，窗，暖气片，消防栓，垃圾桶，通风口，海报，展览板和安全出口标识。其标注图像如图3-10所示。

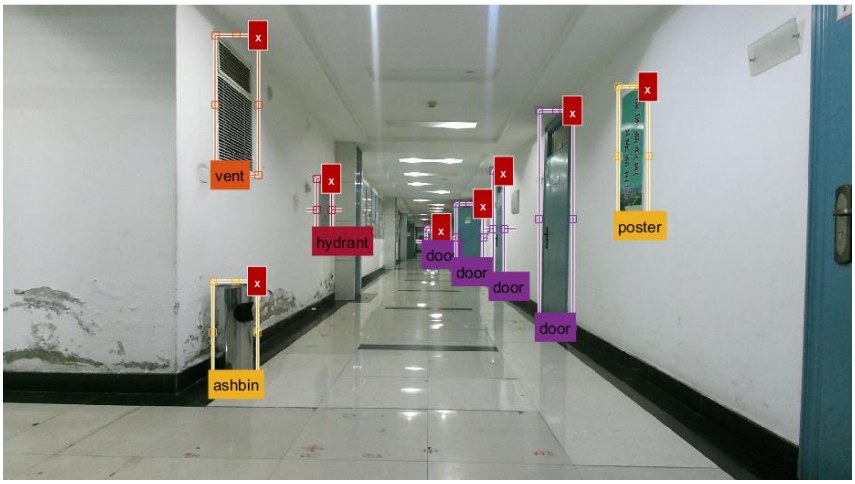


图 3-10 实验环境中语义种类标注图

由于训练的严谨性，需要将训练库中的图像均进行如图3-10一样的标注，以确定训练出该环境内各个角度的多种语义信息。矩形框标记位置需要包含完整语义，矩形框边界要将语义信息与背景信息分割开，确保网络训练的准确性。

3.3.2 语义分割网络性能分析

在网络训练数据库构建完成之后，需要对网络进行优化设置，其主要是对网络中的可调参数进行修改，使网络针对与本实验所采用的室内环境数据集有较好的适应性以及更高的精确度。所需要注意以下两个参数的设定：学习率和迭代次数。学习率的大小影响着网络损失值的下降速度，设置过大会使网络陷入局部最优，设置过小会使网络收敛缓慢。因此选取合适的初始学习率并随着训练过程实时调整

是十分重要的。常用的初始学习率设置模式为先设置较大学习率，然后每乘以0.3进行递减，直至找到合适初始学习率。在设置初始学习率之后，在网络的训练过程中，一般会使用指数衰减模式，保证后续学习率不断减小，符合网络训练过程。图3-11为不同学习率下，神经网络的损失与迭代次数的对比曲线，其中横轴表示神经网络的迭代次数，纵轴表示神经网络的损失值。

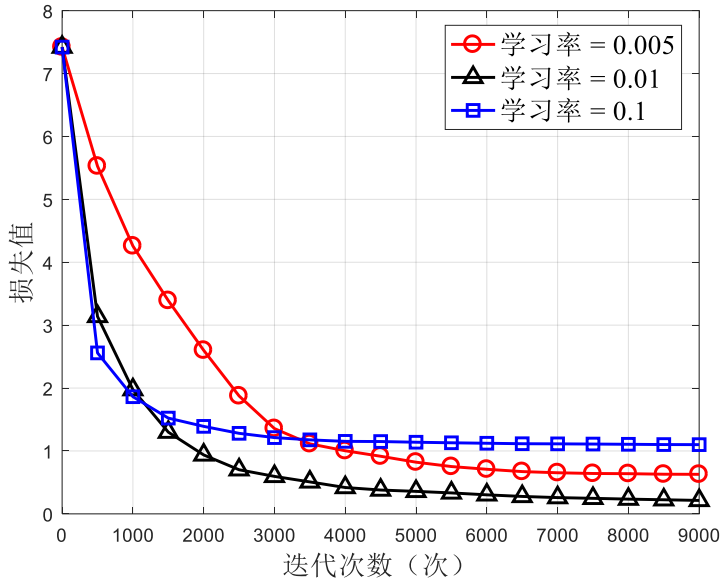
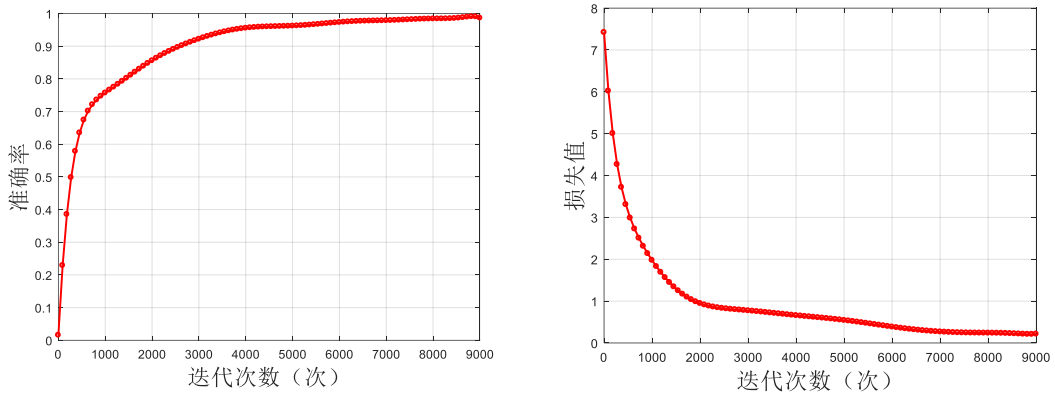


图 3-11 不同学习率下迭代次数与神经网络损失值关系曲线

由图3-11可以看出，当初始学习率设置为0.005时，神经网络的损失值下降的较为缓慢，收敛过慢；当初始学习率设置为0.1时，损失值在迭代开始时下降很快，但随着迭代次数的增加，损失值下降速率逐渐放缓，无法达到最低的损失值。当初始学习率设置为0.01时，损失值的下降速率与最终的收敛值均符合要求，因此本实验暂将初始学习率设置为0.01。

机器学习的迭代次数对最终的网络准确率有很大的影响。如果迭代次数过少，网络容易出现欠拟合现象，即最终网络的输出结果不够准确，精确度不高；如果迭代次数过多，网络容易出现过拟合现象，即最终网络的输出结果虽然十分准确，但只针对于训练库中的测试数据。如果换成其他数据，可能准确率会大幅下降。如图3-12所示为机器学习迭代次数与准确率和损失值之间的关系曲线，该曲线可以给出何时对网络进行停止训练的信息。



(a) 迭代次数与准确率间变化曲线　　(b) 迭代次数与损失值间变化曲线

图 3-12 神经网络训练过程中参数随迭代次数变化曲线

由图3-12中可以看出，神经网络迭代次数上限为9000次，当迭代次数超过7000次时，其准确率基本不再上升，损失值也保持在一个水平范围内。本文采用“no-improvement-in-n”策略，即当在一定迭代次数内，准确率不再上升，维持在一个稳定值或呈现下降趋势，即可采取 Early Stopping，提前停止神经网络的训练，起到防止网络过拟合的效果。在确定好网络的基本参数之后，即可将之前采集的训练数据集输入到网络中进行训练。网络训练好之后，可以在训练集采集场景中随机选取测试图像去验证网络的准确性，本文训练的神经网络语义分割效果如图3-13所示。

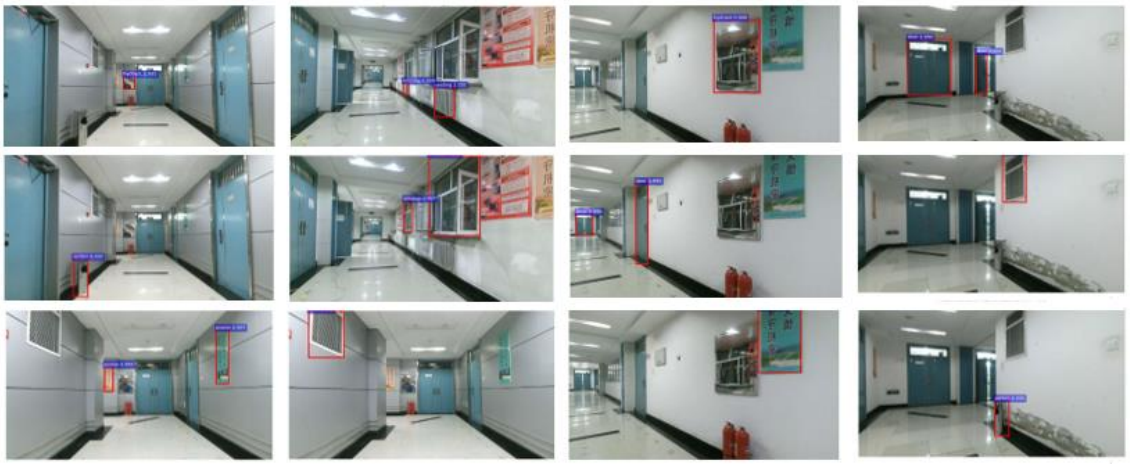


图 3-13 神经网络语义分割效果

由图3-13可以看出，针对每一个语义成分，神经网络输出的为其种类与位置信息，将每一个语义成分在图中用红色矩形框进行标注，并在其上方用蓝色矩形框进行其种类与所属该种类的概率进行说明。

3.3.3 离线数据库分类算法性能分析

离线数据库分类的准确性有很大程度依赖于离线数据库中图像的语义判别准确性。为了确保该语义分割网络针对本实验环境拥有较高的语义分割准确率，同时为了验证网络对每一类语义识别的准确性，本文对实验环境进行了测试图像采集。测试图像采集方式与训练图像采集类似，但有所不同的是，测试图像采集设备并不单一，测试图像采集角度模仿用户在位置环境中的图像拍摄角度。且测试图像采集时间与质量需要有细微差别，旨在检测出不同光照、轻微遮挡、微量噪声以及运动模糊等情况对于语义分割网络识别准确率的影响。本文针对大量测试图像，利用其图像中的多种语义进行了统计分析，其结果如表3-1所示。

表 3-1 不同语义识别准确率

语义种类	识别准确（个）	识别错误或未识别出（个）	识别准确率
门	531	8	98.52%
窗	148	1	99.33%
暖气	135	3	97.83%
海报	180	6	96.77%
展览板	314	7	97.82%
垃圾桶	19	3	86.36%
消防栓	65	1	98.48%
安全出口标识	61	2	96.83%
通风口	25	1	96.15%

由表3-1可以看出，每张测试图像上分别存在不同数量的不同种类语义。针对实验环境，该语义分割网络对每种语义的识别准确率符合离线数据库分类的精度要求，语义识别准确率最可达到99.33%，其语义类别为窗。语义识别准确率最低为86.36%，其语义类别为垃圾桶。垃圾桶类语义识别准确率较低的原因为，其在走廊中出现的次数较少，且由于语义高度问题，在以用户为参考角度的测试图像拍摄中，垃圾桶在图像中出现的数量也较少。因此垃圾桶出现的少量识别错误在垃圾桶类别少量的测试基数下，降低了该类别的识别准确率。每张图像被赋予正确的语义标签即可进行后续的离线数据库分类算法。图3-14为基于语义的离线数据库分类算法对离线数据库的分类混淆矩阵结果图。本文算法将数据库自动根据语义信息分为了35类，图中每一行表示每张图像的真实类别，每一列表示图像经过神经网络后的预测类别，对角线上的值越接近1代表该类别分类越准确。

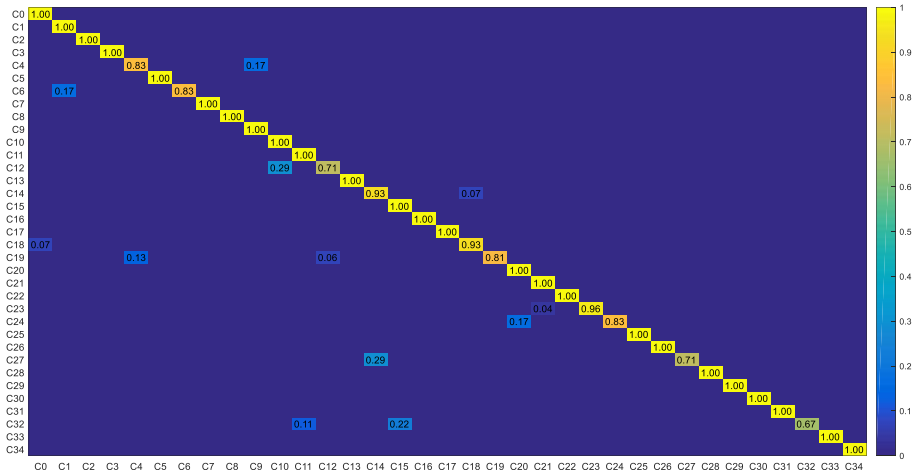


图 3-14 离线数据库分类混淆矩阵

由图3-14可以看出，针对与大部分数据库中的图像，基于语义的离线数据库分类算法对其分类均十分准确，针对某些少数图像，该算法对其分类准确率最低为67%。但对图像数据库的误分类对接下来的检索工作不会带来很大影响。当对处于误分类场景的某一张待检索图片进行分类时，可能由于某种影响因素，该张待检索图像同样被进行了误分类，其相同的影响因素会导致该张图片被分类到其场景的误分类类别中，在该分类中同样拥有该场景的类似图像，在检索阶段同样会匹配到其场景的正确匹配图像，不会影响最终的检索结果。尽管最终检索结果可能正确，但图像中的语义种类信息与位置信息并未进行正确标注，因此会对后续的基于语义约束的特征点选取方法造成影响。本文提出的分类算法可以在保证以高分类准确度为前提的情况下，尽量减小误分类对整体系统的影响。

3.4 本章小结

本章首先对语义分割网络的基本框架进行了详细分析，对语义分割原理进行了公式推导，并对整个网络的优化函数进行了定义。其次，详细分析了基于语义的离线数据库分类算法，分析了利用语义分割网络将图像信息转化成语义标签的过程，并推导了从语义标签转化为数字标签以及进行分类的过程。最后，本章对所提出的基于语义的离线数据库分类算法进行了实验仿真。给出了针对于不同语义信息的分割准确率。对于离线数据库，本文提出的基于语义的离线数据库算法将其共分为了35类，并对分类混淆矩阵进行了仿真。由仿真结果可以看出，基于语义的离线数据库分类方法能够有效地对离线数据库进行精确分类，降低在线检索阶段的检索范围，增加整个系统的实时性。

第4章 基于语义约束的在线检索定位算法研究

在上一章中，本文对基于语义的离线数据库分类算法进行了详细分析，重点对语义分割网络的准确率与离线数据库的分类准确率进行了实验验证，有效地对离线数据库进行准确细分，减小了在线检索阶段的检索范围。对于整个定位系统来说，上一章进行了离线阶段的工作，本章将进行在线阶段的工作。在检索阶段针对随着离线数据库的增加在线检索时间变长的问题，利用上一章分类后的离线数据库，根据本文所提出的基于语义与内容的快速图像检索算法对图像进行快速准确检索，旨在减小在线检索阶段的时间开销，在保证检索精度的前提下加快检索速度。此外，针对定位中提取特征点存在较多误匹配且时间开销较大的问题，利用上一章得出的语义区域像素位置与类别，根据本文提出的基于语义约束的特征点定位方法，可以对误匹配特征点进行剔除，在保证定位精度的情况下提高整个系统的实时性。研究结果表明，在线阶段通过本文所提出的算法，可以明显提升整个系统的实时性。

4.1 SCBIR 算法研究

4.1.1 SCBIR 算法基本流程框架

传统视觉定位系统中存在在线阶段因离线数据库容量过大而导致检索时间较长的问题，为了解决该问题，本文针对性地提出了基于语义与内容的快速图像检索（Semantic and Content-Based Image Retrieval，SCBIR）算法，旨在消除在线检索时间随离线数据库增加而增大的线性依赖关系。通过使用语义分割网络对离线数据库进行语义标签提取与分类，缩小在线检索阶段寻找匹配图像的检索范围，并用颜色与结构特征的方式找到最佳匹配图像，其检索时间开销随离线数据库容量的增大并无明显线性关系。SCBIR算法流程图如图4-1所示。

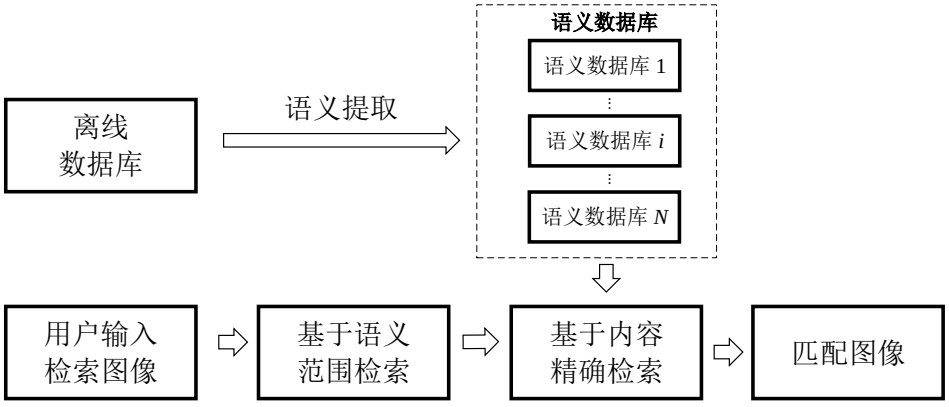


图 4-1 SCBIR 算法流程图

SCBIR算法将基于语义的检索方法与基于内容的检索方法进行了融合。传统的基于语义的检索方法是对图像进行文本描述，图像只有单个语义标签。而本文利用多语义标签信息定位待检索图像的初始位置区域，以解决传统检索方式启动较慢的缺陷；再用基于内容的检索方式对小范围内的语义进行精确检索，达到提升在线阶段检索速度的目的。在SCBIR算法中，基于语义的检索范围确定方式与第三章的离线数据库分类方式类似，对于用户输入图像 I_{user}，首先让其通过语义分割网络对其进行语义特征提取，得到图像中的语义种类组合 S_{user} 与语义判别向量 Ω_{user}。之后将 S_{user} 中的各种语义根据语义判别向量 Ω_{user} 确定其语义的稀疏程度，再通过转化向量转化为唯一的数字标签 l_{user}，该数字标签为对应语义数据库中与用户输入待检索图像拥有相同语义排列的语义子数据库的检索标签。通过上述基于语义的范围检索方式，将待检索图像的检索范围从整个离线数据库定位到了某一种语义排列的语义子数据库中，在语义子数据库中即可对待检索图像进行基于内容的检索方式实现精确检索，找到匹配图像，下面将对基于内容的检索方法进行详细叙述。

4.1.2 基于颜色的特征提取方法

以图像内容为基础的检索方法主要利用了图像的颜色、纹理以及形状结构等特征，弥补了单纯基于单语义标签检索的缺点，和本文提出的基于多语义标签范围检索方式形成了互补。基于内容的检索方法流程图如图4-2所示。

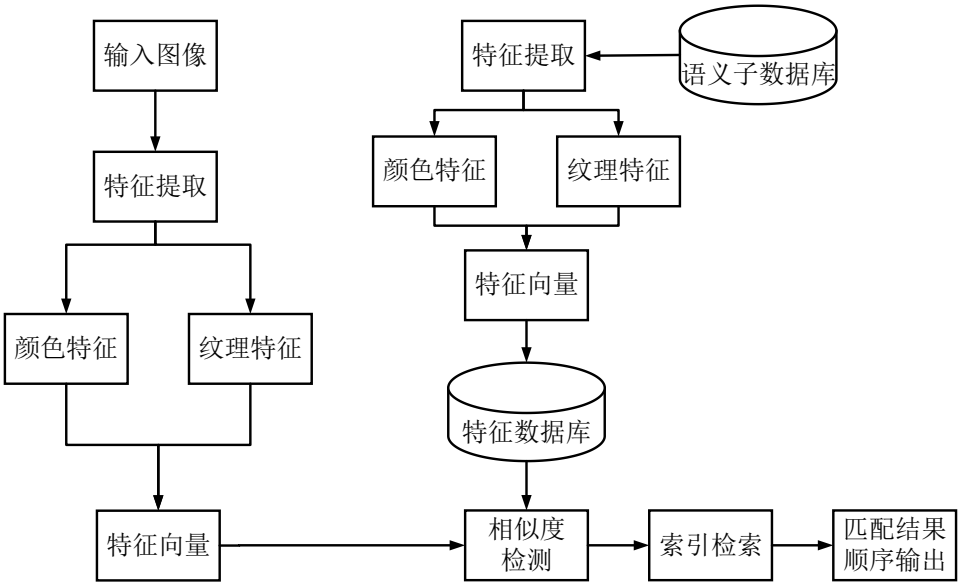


图 4-2 基于内容检索算法流程图

颜色特征是基于内容的检索方法中最重要的特征之一，该特征利用了图像底层的像素点。在三维空间中，有许多表示颜色的坐标系，例如RGB、HSV 以及

CMY 等。最常用的为RGB模型。假设彩色图像 I 的尺寸为 $n_1 \times n_2$，其颜色表示模型为RGB模型。在RGB模型中，每个像素的颜色被表示成一个三维向量 (R,G,B)，向量中每个值的取值范围为0到255，因此对于该三维向量，一共有 256^3 种颜色表现形式。为了减少颜色相关图，需要对颜色进行量化，将每种颜色量化为 m 阶 $h_1,...,h_m$。对于图像 I 中的某像素 $p=(x,y)\in I$，$p\in I_h$ 表示为 p 点的像素为 h。那么图像 I 颜色相关图可以表示为：

$$\gamma_{h_i,h_j}^{(k)} \triangleq \Pr_{\substack{p_1\in I_{h_i}\\ p_2\in I}}\left[p_2\in I_{h_j}\,\middle\|\, p_1-p_2\right|=k\Big] \tag{4-1}$$

其中 $\left|p_1-p_2\right|=\max\left\{\left|x_1-x_2\right|,\left|y_1-y_2\right|\right\}$，$i,j\in\{1,2,...m\}$，$k\in\{1,2,...n\}$。因此，给出任意一个颜色为 h_j 的像素，颜色相关图则会出在距离给定像素 k 处找到颜色为 h_j 的像素的概率。对于图像 I 的颜色相关图来说，$\gamma_{h_i,h_j}^{(k)}$ 表示了 h_i 与 h_j 之间的空间相关性。为了减小空间相关图的尺寸，一个新的概念颜色自相关图产生了，它是颜色相关图的子集。图像 I 的颜色自相关图只表示相同颜色之间的空间相关性，其可表示为：

$$\alpha_h^{(k)}(I)\triangleq\gamma_{h,h}^{(k)} \tag{4-2}$$

对于给定的 h_i，h_j 和 k，为了衡量两个颜色相关图或颜色自相关图之间的相似性，提出了相关距离 $d_1(r,s)=\left|r-s\right|/(1+r-s)$，分母上的1是为了防止被0整除。颜色相关图与颜色自相关图描述子的相关距离被定义为：

$$\left|I_1-I_2\right|_{\gamma,d_1}=\sum_{\substack{i,j\in\{1,...,m\}\\ k\in\{1,...,n\}}} d_1\left[\gamma_{h_i,h_j}^{(k)}(I_1),\gamma_{h_i,h_j}^{(k)}(I_2)\right] \tag{4-3}$$

$$\left|I_1-I_2\right|_{\alpha,d_1}=\sum_{\substack{i\in\{1,...,m\}\\ k\in\{1,...,n\}}} d_1\left[\alpha_{h_i}^{(k)}(I_1),\alpha_{h_i}^{(k)}(I_2)\right] \tag{4-4}$$

通过式(4-3)与(4-4)，即可通过颜色相关图或颜色自相关图对两张图像的颜色相关性进行比较，从而根据相关距离找到最佳匹配图像。

4.1.3 基于结构的特征提取方法

图像的纹理特征与颜色特征类似，也是检索图像的重要特征之一。纹理特征也称为结构特征，其主要与一个特定的、具有空间重复性的表面结构有关，这些表面结构是由重复一个或多个特定的元素在不同的相对空间位置所形成的。纹理特征

很难用语言去描述，但是在图像中被进行了抽象细化定义，例如：细度、平滑度、粒度、线条、方向性以及粗糙度等。这些特征定义了纹理成分的空间布局，使得选取所需要的纹理特征更加容易。通常情况下，纹理特征的提取可以与小波分析的方法相结合。通过小波变换的方式，信号可以分解为一系列的基本方程$\psi_{mn}(x)$，这些方程可以由小波生成方程获得，其形式如下所示：

$$\psi_{mn}(x)=2^{-m/2}\psi(2^{-m}x-n) \tag{4-5}$$

其中，m,n均为整数，因此对于信号$f(x)$可以表示为：

$$f(x)=\sum_{m,n}b_{mn}\psi_{mn}(x) \tag{4-6}$$

由于图像表示的为二维信息，因此需要用到二维小波变换，其将图像分解为四个子带，分别称为LL、LH、HL和HH，根据频率特性，每个子带可以用于每个分解层次，均值和标准差的能量分布即表示图像纹理特征。这些特征较好的捕捉了纹理的高级感知属性，对于图像浏览非常有效，但是对于图像的纹理识别与检索较为不敏感。因此本文应用了马尔科夫随机场纹理模型，随机场模型将图像看作是由随机标量或向量组成的二维数组，即每个像素位置的信号都是一个随机变量。每一种纹理都由信号的联合概率分布来表征，这种分布解释了空间相互依赖，或信号之间的相互作用。相互作用的像素对通常是邻域，随机场纹理模型的特征是邻域之间的几何结构和相互作用的定量强度。像素之间的随机场分布如图4-3所示。

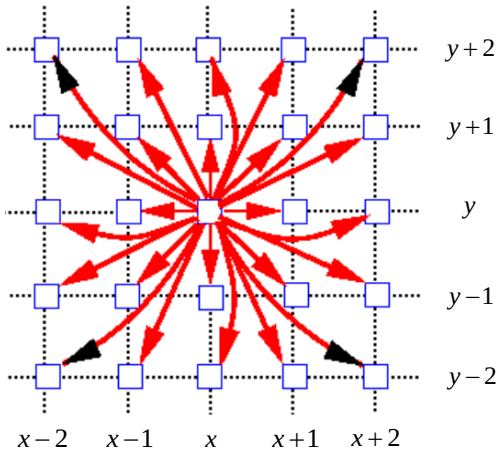


图 4-3 像素间随机场分布模型

图4-3中，每一个方块表示一个像素，若选取正中心(x,y)的像素位置，则箭头表示了选取像素位置及其邻域的随机场分布。假设像素之间的相互作用是平移不变的，那么根据每个像素的特征邻域就可以得出一个交互结构N。该交互结构即

为马尔科夫随机场模型，其中每个像素 $p=(x,y)$ 中信号的条件概率只取决于邻域 $\{(x+m,y+n):(m,n)\in N\}$。在联立自回归高斯-马尔科夫模型的特殊情况下，纹理特征由一组自回归参数表示：

$$g(x,y)=\sum_{(m,n)\in N}a(m,n)g(x+m,y+n)+sw(x,y) \tag{4-7}$$

其中，w 独立的白噪声，$a(m,n)$ 和 s 是确定联立自回归模型的参数。更一般的吉布斯随机场模型具有多个成对的像素相互作用，允许将期望的邻域与一组最有活力的邻域对关联起来。然后，在所选的像素对中，交互结构本身和信号同现关系的相对频率分布可以作为纹理特征。最终将颜色特征与纹理特征相结合，即可作为图像的内容特征。最后，根据待检索图像与检索范围内颜色特征向量与结构特征向量的欧式距离，根据最小欧式距离准则即可找到待检索图像的匹配图像。SCBIR算法的具体流程如表4-1所示。

表 4-1 SCBIR 算法具体流程

输入：待检索图像 I_{user}
输出：待检索图像 I_{user} 的检索匹配结果 I_{match}
第一步：将离线数据库中的 m 张图像输入到语义分割网络，得到每张图像的分类标签 $l_i, i\in m$；
第二步：将分类标签相同的图像分为一类，形成子数据库 D_S^i，其标签代表相应类别的检索标记，并对其中图像进行颜色与结构特征提取，以向量形式保存；
第三步：输入待检索图像 I_{user}；
第四步：确定待检索图像 I_{user} 对应的分类标签 l_{user}；
第五步：根据分类标签 l_{user} 寻找对应的子数据库 D_S^i；
第六步：提取 I_{user} 的颜色结构特征，在子数据库 D_S^i 中找到与 I_{user} 特征向量欧式距离最小的匹配特征向量代表的图像，即为 I_{user} 的检索匹配结果 I_{match}。

4.2 基于语义的特征点选取及视觉定位

4.2.1 基于语义约束的特征点选取方法

上一节中，本文对图像检索算法进行了详细分析。对于用户输入的待定位图像，利用SCBIR算法在离线语义数据库中找到与其最匹配的数据库图像。由于在数据库建立时，每一张采集的图像都伴随其真实世界的位置坐标存入数据库，因此在检索过后，对于用户输入的待定位图像，其粗略的真实世界位置已经得到，需要对其进行精确定位。精确定位即当前粗定位所获得的真实世界坐标为离线数据库中采集

的匹配图像坐标，由于数据库中采集图像时相机所在的位置与用户定位拍摄图像时相机所在的位置不能保证完全一致，可能会存在位姿变化等差异，因此需要对两个相机坐标系的旋转平移关系进行求解，从而得到用户定位拍摄照片时，其相机所在的真实世界位置，实现精确定位。

由于图像为二维信息，所得信息中不存在真实世界的三维空间点坐标。因此，对于两张二维图像与一张图像所在真实世界坐标作为先验条件的定位方法只适用2D-2D定位。由第二章推导的对极几何约束关系可知，只要能获取两张图像中足够的对应像素匹配点，就可以通过对极几何约束关系得到拍摄两幅图像时三维相机坐标系之间的旋转与平移关系。而像素匹配点可以由第二章阐述的SURF特征提取匹配算法得到，因此根据对极几何约束关系，可以得到：

$$p_{match}^{\mathrm{T}}\mathbf{F}p_{user}=0 \tag{4-8}$$

其中，p_{match}表示语义数据库中匹配图像的匹配特征点对像素坐标，p_{user}表示用户输入待定位图像的匹配特征点对像素坐标。两幅图像根据SURF特征点的提取与匹配，可以找出多对匹配特征点对，对于第i对匹配特征点，定义

$$\begin{cases} p_{match_i}=(x_{m_i},y_{m_i},1) \\ p_{user_i}=(x_{u_i},y_{u_i},1) \end{cases} \tag{4-9}$$

将坐标表示为齐次的形式方便后续矩阵运算。由对极几何约束可知，根据两幅图像的匹配特征点对可以得到基本矩阵$\mathbf{F}$。该矩阵是一个3×3的矩阵，根据其性质可知矩阵的秩为2，自由度为7，即存在7个自由取值变量。因此如果想求出基本矩阵$\mathbf{F}$的值，至少需要8对匹配特征点，才可以进行基本矩阵的恢复。传统的特征点对选取方法为，将图像所有的特征点对进行提取，得出每个特征点的描述向量，并根据最小欧式距离将两幅图像中的特征向量进行匹配，得到所有的特征点对。最后，将匹配的特征点对根据欧式距离从小到大进行排序，排序越靠前，证明该对匹配点的相似度越高，选取前面几对即可进行基本矩阵$\mathbf{F}$的求解。然而，对于所处空间状态任意的两幅图像，有很大概率会出现误匹配的现象。此外，SURF特征提取的方式对于分辨率越高的图像，由于其图像金字塔的特征，特征提取时间也就越长，特征向量的匹配也就更加复杂。因此，本文根据第三章语义分割网络生成的语义约束，对特征点提取范围进行规划，对误匹配点进行剔除。由于SCBIR检索方法是根据语义进行检索范围缩小的，因此输入图像I_{user}与匹配图像I_{match}附带的语义标签$S_{user}=[S_1,S_2,\cdots,S_k],1\le k\le c$与$S_{match}=[S_1,S_2,\cdots,S_k],1\le k\le c$是完全一致的。定义$S_{user_j}$

表示S_{user}中的第j个语义，S_{match_j}表示S_{match}中的第j个语义，其中$1 \le j \le k$。对于其中任意一个语义，语义分割网络均会生成语义在图像中的像素位置，为一个四维向量，记作$\{x, y, w, h\}$，其分别表示该种的中心横坐标、中心纵坐标、宽度以及高度，该向量表示了一个中心坐标为$\{x, y\}$的矩形区域。因此，对于重新进行大小重置得到的大小尺寸一致的用户输入图像I_{user}与匹配图像I_{match}，定义大小为$n_1 \times n_2$，则语义提取区域为：

$$Z_{ext} = \bigcup_{\substack{x_j \le n_1, y_j \le n_2 \\ j \in k}} \{x_j, y_j, w_j, h_j\} \tag{4-10}$$

其中，$Z_{ext} \le n_1 \times n_2$，由式(4-10)即可得出两幅图像的分别的语义提取区域。由原来的整幅图像缩小到了图像中每个语义区域的集合，减少了提取特征点的个数，节省了特征提取时间开销，减小了特征提取的复杂度；又从基数上减少了特征点个数，减小了特征点无匹配的概率。

在两幅图像分别从$Z_{ext_{user}}$与$Z_{ext_{match}}$中提取了SURF特征点，并根据最小欧式距离原则将匹配特征点从小到大进行排序之后，定义语义匹配特征点对集合M_1，语义不匹配特征点对集合M_2。对于匹配好的特征点对集合中的第i对匹配点：

$$\begin{cases} (p_{user_i}, p_{match_i}) \in M_1, & \left\{\{x_{u_i}, y_{u_i}\} \in \{x_{u_j}, y_{u_j}, w_{u_j}, h_{u_j}\}\right\} \cap \\ & \left\{\{x_{m_i}, y_{m_i}\} \in \{x_{m_j}, y_{m_j}, w_{m_j}, h_{m_j}\}\right\} \\ (p_{user_i}, p_{match_i}) \in M_2, & \text{其他} \end{cases} \tag{4-11}$$

由式(4-11)可以看出，当该对匹配特征点在两幅图像同种语义类别中存在时，将其放入到语义匹配特征点对集合；如果该对匹配特征点分别在两幅图像中的不同语义类别中，则将其放入到语义不匹配特征点对集合。

4.2.2 基于对极约束的视觉定位方法

在语义不匹配特征点对集合中的点即为误匹配点，需将其剔除。对于在语义匹配特征点对集合中的匹配特征点对，将其按照最小欧式距离准则进行排序，从前到后依次选取符合条件的匹配特征点对，进行基本矩阵$\mathbf{F}$的求解。定义基本矩阵$\mathbf{F}$为：

$$\mathbf{F}=\begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \tag{4-12}$$

将式(4-8)与式(4-12)进行合并，可得：

$$p_{match}^{\mathrm{T}}\mathbf{F}p_{user}=(x_{m_i},y_{m_i},1)\begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix}\begin{pmatrix} x_{u_i} \\ y_{u_i} \\ 1 \end{pmatrix}=0 \tag{4-13}$$

将式(4-13)展开，得到齐次线性方程：

$$\left(x_{m_i}x_{u_i},x_{m_i}y_{u_i},x_{m_i},x_{u_i}y_{m_i},y_{m_i}y_{u_i},y_{m_i},x_{u_i},y_{u_i},1\right)\begin{pmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{pmatrix}=\mathbf{w}_i^{\mathrm{T}}\mathbf{f}=0 \tag{4-14}$$

其中，$\mathbf{w}_i$ 表示齐次方程系数向量，$\mathbf{f}$ 表示基本矩阵转化向量。将得到的多对特征匹配点带入式(4-14)，可以将其表示成如下形式：

$$\mathbf{Wf}=0 \tag{4-15}$$

其中，由于最少需要8对匹配特征点才能求得基本矩阵，因此 $\mathbf{W}=\left(\mathbf{w}_1,\mathbf{w}_2,.....,\mathbf{w}_8\right)^{\mathrm{T}}$。由于基本矩阵 $\mathbf{F}$ 的秩为2，因此式(4-15)存在非零解。矩阵 $\mathbf{W}$ 在理想的情况下秩为8，可以按照求解线性方程组的方式对其求解。然而在实际情况中，由于图像中存在的噪声干扰， $\mathbf{W}$ 是满秩矩阵，式(4-15)只有零解。因此需要将求解式(4-15)的问题进行转化，变为另一种基本矩阵求解方式：

$$\min\|\mathbf{Wf}\|,\|\mathbf{f}\|=1 \tag{4-16}$$

为了对基本矩阵进行准确估算，因此需要保证 $\|\mathbf{Wf}\|$ 最小且接近于零，=该条件受限于 $\|\mathbf{f}\|=1$。将矩阵 $\mathbf{W}$ 进行奇异值分解可得：

$$\mathbf{W}=\mathbf{UDV}^{\mathrm{T}} \tag{4-17}$$

首先，对$\mathbf{WW}^{\mathrm{T}}$求取特征值并将其进行降序排列，将特征值对应的特征向量进行排列即可得到矩阵$\mathbf{U}$；其次，对$\mathbf{WW}^{\mathrm{T}}$或$\mathbf{W}^{\mathrm{T}}\mathbf{W}$的特征值平方根进行求解，并将结果降序排列构成的对角阵即为矩阵$\mathbf{D}$；最后，对$\mathbf{W}^{\mathrm{T}}\mathbf{W}$求取特征值并将其进行降序排列，将特征值对应的特征向量进行排列即可得到矩阵$\mathbf{V}$。为确保$\|\mathbf{Wf}\|$值最小，因此需要选取矩阵$\mathbf{V}$中最后一列向量构造基本矩阵转化向量$\mathbf{f}$，从而得出基本矩阵。

在获取了$\mathbf{F}$之后，需要进一步对$\mathbf{E}$进行求解。$\mathbf{E}$是由$\mathbf{F}$变换得到的，其转换关系如下：

$$\mathbf{E}=\mathbf{K}_{\mathrm{d}}\mathbf{F}\mathbf{K}_{\mathrm{u}} \tag{4-18}$$

其中，$\mathbf{K}_{\mathrm{d}}$为构建离线数据库时所用相机的内部参数矩阵，$\mathbf{K}_{\mathrm{u}}$为用户进行图像拍摄时，图像采集设备的内部参数矩阵。在获取了两个内参矩阵之后，即可对本质矩阵$\mathbf{E}$进行分解，得到两个图像采集设备之间的位姿关系：

$$\mathbf{E}=\mathbf{t}\times\mathbf{R} \tag{4-19}$$

其中，$\mathbf{R}$和$\mathbf{t}$分别表示两个图像采集装置三维相机坐标系之间的的旋转矩阵和平移向量。由于三维世界坐标系已经提前确定并保持不变，因此对应匹配点在三维相机坐标系中的旋转关系即为在真实三维世界坐标系中的旋转关系，但平移关系会有所改变。假设相机坐标系下用户拍摄照片中某点为$P_u=(x_u,y_u,z_u)$，数据库匹配图像中该点对应点在三维相机坐标系下的坐标为$P_d=(x_d,y_d,z_d)$，由于都在三维相机坐标系下，二者满足如下关系：

$$P_u=\mathbf{R}P_d+\mathbf{t}=\mathbf{R}(P_d+\mathbf{R}^{-1}\mathbf{t}) \tag{4-20}$$

由式(4-20)可以看出，在三维相机坐标系下，数据库相机坐标系与用户相机坐标系的相机光心平移向量为$\mathbf{R}^{-1}\mathbf{t}$。假设三维相机坐标系下的两点$P_u$与$P_d$在三维世界坐标系中的对应点为$P_{uw}$与$P_{dw}$，数据库相机坐标系与三维世界坐标系之间的旋转矩阵为$\mathbf{R}'$，平移向量为$\mathbf{t}'$，则：

$$\begin{cases}P_{d_1}=\mathbf{R}'P_{dw_1}+\mathbf{t}'\\P_{d_2}=\mathbf{R}'P_{dw_2}+\mathbf{t}'\end{cases} \tag{4-21}$$

将式(4-21)中上下两式做差可得：

$$P_{d_2}-P_{d_1}=\mathbf{R}'(P_{dw_2}-P_{dw_1}) \tag{4-22}$$

由式(4-22)可以看出，从三维世界坐标系转换到数据库相机坐标系的平移关系

只与旋转矩阵$\mathbf{R}'$有关，与平移向量$\mathbf{t}'$无关。因此，在三维世界坐标系下，用户相机位置与数据库图像采集相机位姿的平移关系为：

$$\mathbf{t}_{wd2u} = (\mathbf{R}')^{-1}\mathbf{R}^{-1}\mathbf{t} \tag{4-23}$$

在已知三维世界坐标系下数据库图像的真实位置与两幅图像拍摄位置位姿关系的情况下，即可对用户所在的精确位置进行求解。首先通过SCBIR算法找到用户输入待定位图像在语义数据库中的匹配图像，得到用户的初始粗略位置。之后根据二者之间的特征点匹配关系，利用基于语义约束的特征点定位方法计算出在世界坐标系下二者之间的位姿关系，最终完成定位操作。

4.3基于语义的检索定位方法性能分析

4.3.1 SCBIR 方法性能分析

本课题中 SCBIR 算法旨在加快在线阶段的检索速度，且消除掉离线数据库容量与在线检索时间之间存在的线性增长关系。其主要研究的是数据库容量与在线检索时间之间的关系，但是研究该方面的前提条件是检索准确率达到一定的标准。因此本课题将传统基于内容的检索方法（Content-Based Image Retrieval，CBIR）[47]与本文提出的 SCBIR 方法就检索准确率进行了比较，采用了平均检索准确率（Mean Average Precision，MAP）这一衡量指标，其中 MAP 值越高（越接近 1），代表系统检索出的相关信息排列越靠前，即检索方法性能更好。如图 4-4 所示为两种方法 MAP 的对比曲线。

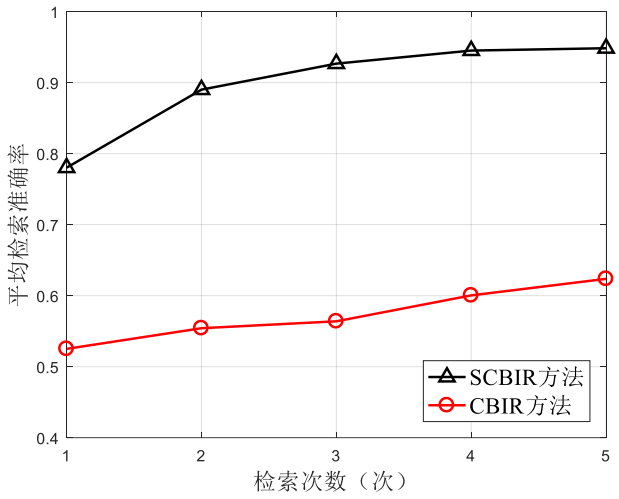


图 4-4 两种检索算法平均检索准确率对比曲线

由图 4-4 可以看出，本文提出的 SCBIR 方法的平均检索准确率随着检索次数的增加均在 90%上下浮动，而传统的 CBIR 算法的平均检索准确率在 60%上下浮动。CBIR 方法平均检索准确率较低的主要原因为，本文选取的待定位环境是一种循环结构，每隔一段距离其视觉特征可能较为类似，因此 CBIR 方法在检索时将不同循环结构的场景混淆，导致其平均检索准确率下降。而本文提出的 SCBIR 方法则利用语义组合的形式，很好地避免了循环结构场景对检索的影响。可以看出本文提出的 SCBIR 算法在检索准确率上优于 CBIR 算法。在同一张待检索图片的前提下，本文还针对不同数据库容量下，两种算法的检索时间开销进行了对比。在对时间开销进行说明之前，需要对进行本实验的实验环境进行说明。本实验 Linux 操作系统下进行，电脑 CPU 配置为 Intel I7-7700，显卡配置为 NVIDIA GTX1060，8G 内存的笔记本平台下进行。如图 4-5 所示为两种算法的时间开销对比曲线。其中横轴为检索数据库中的图像张数，纵轴为检索单张图像的时间开销。

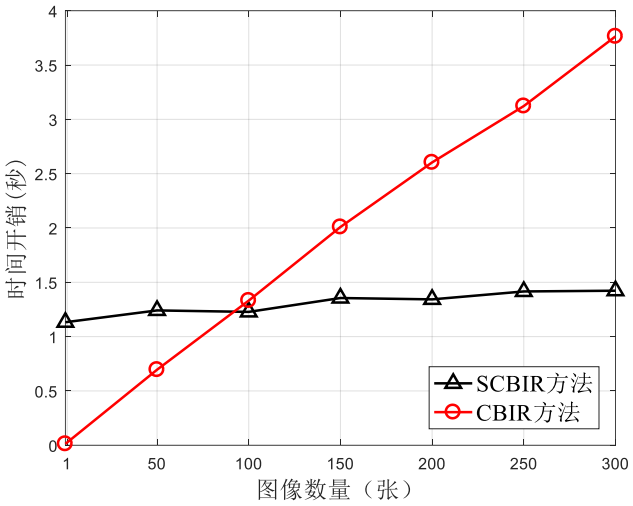


图 4-5 两种检索算法时间开销对比曲线

由图 4-5 可以看出，当检索数据库容量较小时，CBIR 算法占有一定的优势，其省略了待检索图像经过语义分割网络进行语义判别的过程，因此检索速度较快。但在检索数据库容量超过 100 张图片之后，SCBIR 算法的优势显现了出来，其检索的时间开销基本维持在 1-1.5s 之间，随检索数据库容量增加，并没有太大的时间开销变化。形成该种曲线趋势的主要原因是数据库容量的增加主要是因为场景的面积增大，而对于拥有同类语义组合的场景，采集的图像数量有限，因此检索的范围也基本一致，使得检索时间维持在一个水平范围内而不是线性增长。影响 SCBIR 方法检索时间的因素有两个：一是对同一语义组合的场景，缩小图像采集间隔，使

语义子数据库中图像数量增加，检索范围也相应增加；二是在带定位环境中，有多个场景均存在相同的语义组合，导致语义子数据库容量增大。而在目前的场景中，由于采集间隔基本固定，且场景中不会存在完全一致的循环结构，因此 SCBIR 方法的检索时间并不会有大幅度的变化。CBIR 方法则随着检索数据库的容量增加，其对应的检索时间开销也成正比增长。SCBIR 算法消除了随着离线数据库容量增大，在线检索时间相应变长的线性增长关系。因此本文提出的方法在针对大型检索数据库时，对减小时间开销以及提高检索准确度具有重要的作用。

4.3.2 基于语义约束定位方法性能分析

在得到用户输入的待定位图像匹配图像与用户的粗定位位置之后，需要对用户输入图像与匹配图像进行特征点提取与匹配。室内场景下的 SURF 特征点主要分布在门、窗、海报以及展览板等存在语义信息的位置附近，四周的白墙上含有极少的特征点。但是白墙场景在室内十分常见，且在室内图像采集中占据着图像像素的大部分。传统的特征提取匹配算法会对整幅图像进行特征点提取，对整幅图像进行 SURF 特征点提取时会有大量的时间开销以及较高的复杂度。此外，过多的特征点会出现大量的误匹配问题，如图 4-6 所示。

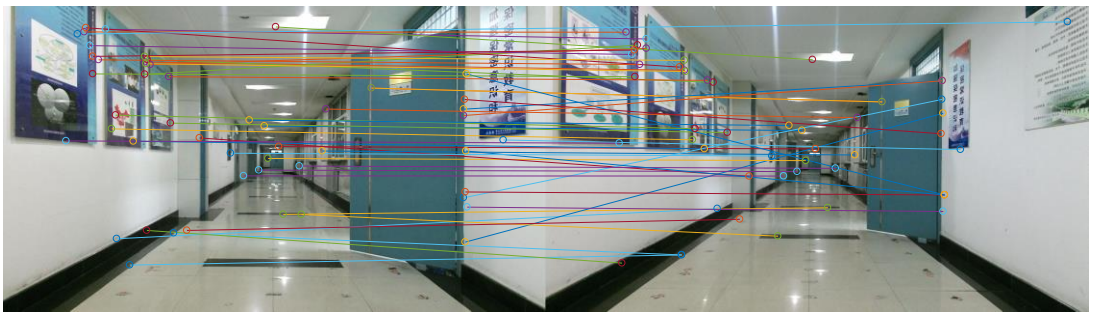


图 4-6 室内场景下误匹配 SURF 特征点

如果在定位环节中选取了误匹配的特征点对，对后续求解旋转平移矩阵的过程会带来巨大影响，导致定位精度偏低，定位误差过大。而且提取整幅图像特征点的操作含有很高的计算复杂度，对于在线定位阶段来说会降低定位的实时性，同时影响了用户体验。因此本课题利用语义分割网络中输出的语义种类与伴随其输出的位置向量$\{x, y, w, h\}$进行了 SURF 特征提取区域的规划，对用户输入的待定位图像进行了语义约束操作，将图像中不需要的背景信息驱除，只留下含有语义种类的前景信息。以图 4-6 中的展览板为例，由语义分割网络输出可得两张图像中展览板的位置信息，在相应位置区域内进行 SURF 特征提取，其结果如图 4-7 所示。

图 4-7 语义约束条件下室内场景下 SURF 特征点匹配

由图 4-7 可以看出，其特征点数量对比图 4-6 有明显减少，特征点提取匹配时间开销与非语义限制方法相比有明显降低，误匹配点通过语义区域的限制也基本消除。在对其他语义，例如门、海报等进行同样特征提取匹配之后，即可对特征点对按欧氏距离进行排序，选取优质的特征点对进行后续定位工作。本文根据图像中语义种类的数量将语义数据库中图像分为 4 类，语义种类在 0-2 类的为类别 1，语义种类在 3-4 类的为类别 2，语义种类在 5-6 类的为类别 3，语义种类在 7-9 类的为类别 4，分别对每一类图像其进行整幅图像的SURF特征点提取与基于语义约束区域的特征点提取，之后在对特征点取平均值，其特征点数量对比如图 4-8 所示。

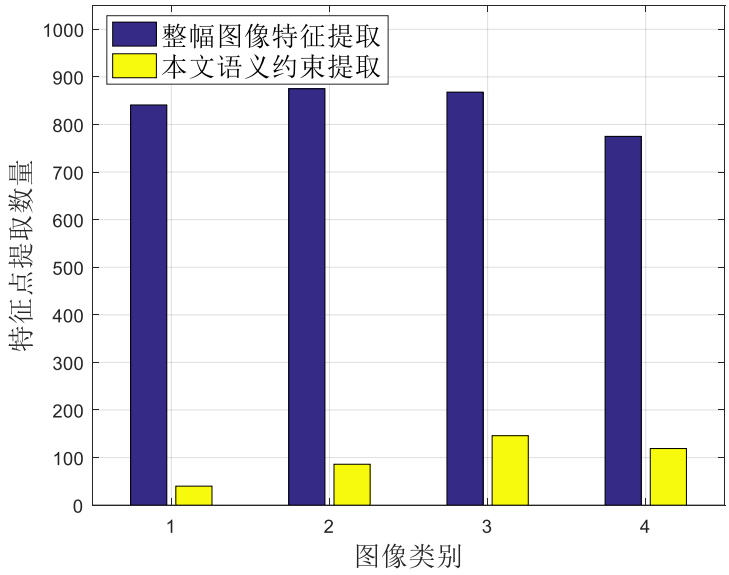


图 4-8 两种不同特征提取范围的特征点数目比较

由图 4-8 可以看出，对于实验场景中的不同类别图像进行整幅图像范围的特征点提取，其特征点数目大致在900左右；只针对语义约束区域进行特征点提取，其特征点数目大致在100左右，相较于原有方法特征点数目减少了 80%，其特征提取时间开销随特征点提取数目的减少也相应减少，增加了定位系统的实时性。此外，

通过在语义约束区域中进行特征点选取，由于语义种类的匹配性，会大量减少误匹配的特征点对，使最终提取出的特征点均为优质特征点。在定位阶段即可在优质匹配特征点集中选取定位所需特征点进行精确定位。最后，针对相同的输入图像，本文将对所提出的算法与对比算法进行相应的性能仿真分析。首先，对本文提出的 SCBIR 检索方法与基于语义约束的特征点定位方法进行定位性能分析；其次，对利用 SVM+CBIR 检索算法与基于对极几何的定位算法进行定位性能分析，分析其在检索性能上的差异对最终定位结果的影响；最后，对利用 SURF 特征提取匹配检索算法与基于对极几何的定位算法进行定位性能分析，分析传统对整幅图像进行 SURF 特征提取与匹配所带来的误匹配点对定位结果带来的影响。其最终结果如图 4-9 所示。其中，横轴为输入测试点的定位误差，单位为厘米，纵轴为定位误差的累积分布函数曲线（Cumulative Distribution Function，CDF）。

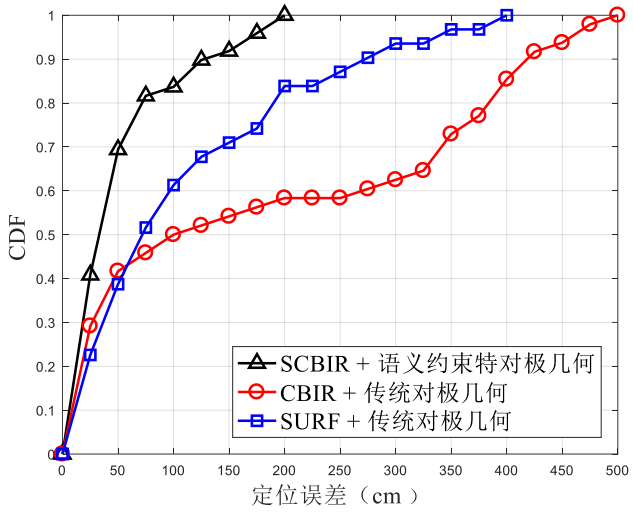


图 4-9 三种不同方法下定位误差对比曲线

本文为了确保实验环境的公平性，对三种方法使用的离线数据库与训练库进行了限制，让实验环境在最大限度上保持一致。由图 4-9 可以看出，采用 SCBIR 算法检索到的匹配图像与测试点位置比较接近，因此通过对极几何方法计算距离之后，与实际位置相差不大，定位误差较小；而采用SVM+CBIR 算法检索到的匹配图像由于和测试点图像相差较大，且由于定位环境具有循环结构，因此容易将某一场景的图像定位到下一个循环结构的场景中，导致初始粗定位结果误差较大。通过传统SURF方法进行特征提取匹配在精细程度上不如其他两种算法，但在最大定位误差上要小于SVM+CBIR方法，由于实验环境的循环结构，使得SVM+CBIR方法检索到的初始匹配图像误差较大。本文提出算法性能优势的根本原因在于，本文

利用了大量的离线阶段工作量换取了在线阶段的精确检索，其离线数据库的分类精确程度远高于其他两种对比算法，在粗定位结果更准确的情况下进行精确定位，使得整个定位系统整体定位性能更好。由仿真结果可知实际检索的准确度决定着用户的初始位置，一旦初始位置定位准确，后续的定位算法会更加精确；如果用户的初始位置确定错误，则后续的定位算法会在用户初始位置的基础上进行定位计算，最终结果会远远偏离用户准确位置。通过CDF曲线可知，本文提出的算法累积概率在1σ时，定位误差在50cm以内；而对比算法累积概率在1σ时，定位误差分别为100cm与260cm。在室内定位中，由于用户本身位置与手机之间也存在一定距离，因此一般以100cm作为定位误差衡量标准。本文提出的基于语义约束的特征点定位方法定位误差在100cm以内的概率为85%左右，最大定位误差为190cm；而对比算法定位误差在100cm以内的概率分别为60%与50%左右，最大定位误差分别为390cm与480cm。根据以上分析结果可以看出，本文提出的基于语义的视觉定位方法性能要优于使用CBIR与SURF等检索方法再进行对极几何定位的对比算法。

4.4 本章小结

本章主要对视觉定位系统在线阶段的检索算法与定位算法进行了详细阐述与算法仿真，并对仿真结果进行了对比分析。本文提出的在线检索阶段的SCBIR算法检索准确度要高于传统的CBIR算法，且消除了离线数据库与在线检索时间开销之间的线性增长关系。此外本文在定位阶段提出了基于语义约束的特征点提取方法与视觉定位算法，减小了在线定位阶段的特征点提取数量且提高了整个定位系统的定位效率。本文针对不同的检索算法得到的匹配图像运用不同的定位方法进行了定位结果仿真，仿真结果表明本文提出定位算法的特征点提取数量明显减少，定位精确满足要求，并提高了整个系统的实时性，给用户更好的体验。

结　论

定位是人类生活中不可或缺的一环，目前人类每天约有80%左右的时间在室内活动，因此室内定位凭借其独特的优势正逐渐获得研究人员的广泛关注。而室内视觉定位技术更是凭借其内置传感器的简单与便利，正在逐渐替代那些需要安装额外开销且设计特殊的室内定位系统。此外，以视觉信息进行定位的方式与人类自身通过眼睛确定位置过程近似，更值得进行深入研究。

本文将机器学习中的语义分割与定位算法相结合，首先研究了视觉定位技术与语义信息应用的国内外研究现状，并对机器学习的发展与视觉定位的结合进行了分析。其次，本文研究了机器学习分割出的语义成分在视觉定位系统中的应用。除此之外，本文针对传统视觉定位系统在离线阶段以及在线阶段存在的一些问题完成了以下研究成果：

首先，针对传统视觉定位系统离线阶段建立的数据库中数据量较大、图像检索耗时过长的问题，提出了一种基于语义的离线数据库分类方法，该方法利用机器学习的方式对数据库中图像进行了语义提取并分类为语义子数据库，能够有效地减少数据库容量增大，在线阶段检索时间延长的问题。该方法利用语义分割网络对图像中语义信息进行提取，并按照不同语义组合对图像进行多标签分类，其分类结果相较于无监督聚类方法或有监督分类算法都有很高的分类准确率。此外，语义分割网络的语义种类与位置的输出也会应用在后续定位算法中，减小系统复杂度。

其次，在进行离线数据库分类转换成语义数据库后，利用基于语义与内容的快速图像检索方法在语义子数据库中进行精确检索，在保证了检索准确率的同时节省了大量时间开销。此外，针对传统视觉定位系统中定位阶段对全局进行特征提取导致时间开销较大的问题，提出了一种基于语义约束的特征点定位算法，该算法有效地减少了定位匹配阶段提取特征点的区域与提取特征点的个数，减少了特征点的提取基数并减少了特征提取的时间开销。此外，由于语义限制，该方法可以剔除大量的误匹配特征点，兼顾特征提取算法的速度与定位算法的精度。

最后，本文为验证算法准确率与性能在合适实验环境中进行了仿真分析。仿真结果表明，本文提出的基于语义与内容的快速图像检索方法相较于传统的基于内容的检索算法提升了30%左右的检索准确率，并消除了数据库容量与检索时间开销之间的线性增长关系。本文基于语义约束的特征点定位算法在特征点提取数目方面相较于整幅图像特征点数目约减少了80%，定位结果相较于对比算法也有很大提升，提高了整个室内视觉定位系统的实时性。

本文主要以视觉定位系统在线阶段检索定位的实时性及用户体验为出发点，解决了在线阶段检索时间随离线数据库增大而线性增长的问题，以及在线定位阶段特征点提取数目过多，时间开销较大的问题。然而对于整个视觉定位系统来说，离线数据采集阶段同样存在一些问题亟待解决。

（1）在离线阶段构建离线数据库的过程中，需要人为对待定位环境图像进行采集，随着待定位场景的增大，采集图像所耗费的人力也逐渐增加。此外，本文采用了神经网络训练的方式，还需要进行训练数据库的采集与标注，更加大了离线阶段数据库采集的人工成本。如何快速地进行离线数据库采集与训练数据库标注是一个值得深入研究的问题。

（2）在离线阶段确定定位数据库位置的过程中，传统方式为在采集图像时同时记录图像采集的世界坐标位置，然而这种位置确定方式构建的定位地图并不稠密。定位系统最终的到的定位结果与定位数据库的稠密程度息息相关，因此如何构建稠密定位数据库也是一个需要深入探索的问题。

在未来的工作中，将针对以上提出的两点问题进行深入研究，对离线数据库进行稠密构建并对训练数据库进行快速标注，以在较少人力开销的前提下，实现实时性较高的视觉定位系统。

参考文献

[1] V. Indelman, Pini Gurfil, E. Rivlin, H. Rotstein. Real-Time Vision-Aided Localization and Navigation Based on Three-View Geometry[J]. IEEE Transaction on Aerospace and Electronic Systems. 2012, 48(3): 2239-2259.

[2] G. De Blasio, A. Quesada-Arencibia, C. R. García, J. C. Rodríguez-Rodríguez and R. Moreno-Díaz, "A Protocol-Channel-Based Indoor Positioning Performance Study for Bluetooth Low Energy," in IEEE Access, vol. 6, pp. 33440-33450, 2018.

[3] A. Aguilar-Garcia, S. Fortes, R. Barco and E. Colin, "Enhancing Localization Accuracy With Multi-Antenna UHF RFID Fingerprinting," 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Banff, AB, 2015, pp. 1-9.

[4] L. C. Png, L. Chen, S. Liu and W. K. Peh, "An Arduino-Based Indoor Positioning System (IPS) Using Visible Light Communication And Ultrasound," 2014 IEEE International Conference on Consumer Electronics - Taiwan, Taipei, 2014, pp. 217-218.

[5] A. Vecchio and G. Cola, "Fall Detection Using Ultra-Wideband Positioning," 2016 IEEE SENSORS, Orlando, FL, 2016, pp. 1-3.

[6] S. He and S. -. G. Chan, "Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 466-490, Firstquarter 2016.

[7] M. Yuda, Z. Xiangjun, S. Weiming and L. Shaofeng, "Target Accurate Positioning Based on The Point Cloud Created by Stereo Vision," 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP), Nanjing, 2016, pp. 1-5.

[8] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.

[9] N. Brasch, A. Bozic, J. Lallemand and F. Tombari, "Semantic Monocular SLAM for Highly Dynamic Environments," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, 2018, pp. 393-400.

[10] G. Zou, L. Ma, Z. Zhang and Y. Mo, "An Indoor Positioning Algorithm Using Joint Information Entropy Based on WLAN Fingerprint," Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Hefei, 2014, pp. 1-6.

[11] Chung-Hao Huang, Lun-Hui Lee, Chian C. Ho. Real-Time RFID Indoor Positioning

System Based on Kalman-Filter Drift Removal and Heron-Bilateration Location Estimation[J]. IEEE Transactions On Instrumentation And Measurement, 2015, 64(3): 728-739.

[12] Yang C, Shao H R. WiFi-based Indoor Positioning[J]. Communications Magazine IEEE, 2015, 53(3): 150-157.

[13] T. Dao, M. Le and Q. Nguyen, "Indoor Localization System Using Passive UHF RFID Tag and Multi-Antennas," 2014 International Conference on Advanced Technologies for Communications (ATC 2014), Hanoi, 2014, pp. 405-410.

[14] Tian Q , Salcic Z , Wang I K , et al. A Multi-Mode Dead Reckoning System for Pedestrian Tracking Using Smartphones[J]. IEEE Sensors Journal, 2016, 16(7): 2079-2093.

[15] Aharon M, Elad M, Bruckstein A. -SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311-4322.

[16] J. Jiang, D. Wu and Z. Jiang, "A correlation-based bag of visual words for image classification," 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, 2017, pp. 891-894.

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[C]. International Conference on Neural Information Processing Systems. Curran Associates Inc, Lake Tahoe, 2012: 1097-1105.

[18] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[19] S. Kido, Y. Hirano and N. Hashimoto, "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)," 2018 International Workshop on Advanced Image Technology (IWAIT), Chiang Mai, 2018, pp. 1-4.

[20] H. Xue, L. Ma and X. Tan, "A fast visual map building method using video stream for visual-based indoor localization," 2016 International Wireless Communications and Mobile Computing Conference (IWCMC), Paphos, 2016, pp. 650-654.

[21] J. Hlubik, P. Kamencay, R. Hudec, M. Benco and P. Sykora, "Advanced point cloud estimation based on multiple view geometry," 2018 28th International Conference Radioelektronika (RADIOELEKTRONIKA), Prague, 2018, pp. 1-5.

[22] E. Deretey, M. T. Ahmed, J. A. Marshall and M. Greenspan, "Visual indoor positioning with a single camera using PnP," 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Banff, AB, 2015, pp. 1-9.

[23] Yang J, Li H, Campbell D, et al. Go-ICP: A Globally Optimal Solution to 3D ICP

Point-Set Registration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016: 1-1.

[24]李鹏,张洋洋.采用辅助靶标的移动机器人立体视觉定位[J]. 红外与激光工程, 2019, 48(S1): 110-119.

[25]牛家旭, 孟真. 基于卷积神经网络的室内机器人视觉定位算法[J]. 信息技术与信息化, 2019(03): 65-67.

[26]Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science, 2006, 313(5786): 504−507.

[27]Dollar P, Appel R, Belongie S, et al. Fast Feature Pyramids for Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[28]Girshick R, Donahue J, Darrell T, et al. Region-based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[29]Csurka G, Perronnin F. An Efficient Approach to Semantic Segmentation[J]. International Journal of Computer Vision, 2011, 95(2): 198-212.

[30]An Z, Xu X P, Yang J H, et al. Design of Augmented Reality Head-up Display System Based on Image Semantic Segmentation[J]. Acta Optica Sinica, 2018, 38(7): 0710004.

[31]M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 3213-3223.

[32]X. Wang, A. Shrivastava and A. Gupta, "A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3039-3048.

[33]K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

[34]P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with Convolutional Networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1713-1721.

[35]Zagoruyko S, Lerer A, Lin T Y, et al. A MultiPath Network for Object Detection[J]. 2016.

[36]Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4): 834-848.

[37] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng and T. S. Huang, "Revisiting Dilated Convolution: A Simple Approach for Weakly- and Semi-Supervised Semantic Segmentation," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7268-7277.

[38] P. Wang et al., "Understanding Convolution for Semantic Segmentation," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 2018, pp. 1451-1460.

[39] J. Dai, K. He and J. Sun, "BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1635-1643.

[40] W. Ren et al., "Deep Video Dehazing With Semantic Segmentation," in IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1895-1908, April 2019.

[41] Y. Wei, J. Feng, X. Liang, M. Cheng, Y. Zhao and S. Yan, "Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6488-6496.

[42] 林志玮, 涂伟豪, 黄嘉航, 丁启禄, 刘金福. 深度语义分割的无人机图像植被识别[J]. 山地学报, 2018, 36(06): 953-963.

[43] 徐谦, 李颖, 王刚. 基于深度学习图像语义分割的机器人环境感知[J]. 吉林大学学报(工学版), 2019, 49(01): 248-260.

[44] 苏健民, 杨岚心, 景维鹏. 基于 U-Net 的高分辨率遥感图像语义分割方法[J]. 计算机工程与应用, 2019, 55(07): 207-213.

[45] Luong Q T, Faugeras O D. The Fundamental Matrix: Theory, Algorithms, and Stability Analysis[J]. International Journal of Computer Vision, 2014, 17(1): 43-75.

[46] A. Shrivastava, A. Gupta and R. Girshick, "Training Region-Based Object Detectors with Online Hard Example Mining," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 761-769.

[47] P. Vikhar and P. Karde, "Improved CBIR system using Edge Histogram Descriptor (EHD) and Support Vector Machine (SVM)," 2016 International Conference on ICT in Business Industry & Government (ICTBIG), Indore, 2016, pp. 1-5.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

[1] Jin Dai, Lin Ma, Danyang Qin and Xuezhi Tan. High Accurate and Efficient Image Retrieval Method Semantics for Visual Indoor Positioning[C]. The 8th International Conference on Communications, Signal Processing, and Systems, 2019.

（二）申请的专利

[1] 马琳, 戴进, 谭学治. 何晨光. 一种基于语义和内容的快速图像检索方法及装置: 中国, 201910251034.0

致　谢

时间如白驹过隙，转眼间我在哈工大的六年学习生涯即将结束。如今再回顾这六年的点点滴滴，在将来我走向工作以及社会中，都是弥足珍贵的回忆。这六年的回忆可以让我明澈本心，砥砺前行。

首先我要感谢我的导师谭学治教授，谭老师是一个十分具有人格魅力的人，他教会了我许多在学校中应该去做的事，应该去合理利用的时间以及应该把自身主要研究重点、主要精力放在何处。在实习与工作选择方面也给了我很大帮助。

然后我要感谢马琳副教授，感谢马老师对我的科研和生活的耐心教导和鼓励，在马老师的项目组中，我学到了很多技术，也掌握了许多新的知识。在毕业论文的撰写过程中，马老师给我提出了许多有关于格式和内容的宝贵意见，您认真的工作态度给我留下了十分深刻的印象，是我今后学习的榜样。

感谢哈工大通信所 1221 实验室的全体师兄师姐，和你们相处并没有不同年龄以及不同阅历之间的隔阂，大家一起学习，一起娱乐，团结协作共同进步。感谢杨浩、李伦、谭竞扬师兄，你们在我的学习生活中给了我很多帮助，感谢赵航师兄与贾彤师姐，你们是我在实验室中的领路人，让我可以更快的融入实验室这个集体。感谢田润、殷锡亮两位博士师兄，虽然研究的不是一个方向，但是在我学习过程中给了我很大的帮助，在论文写作方面我有很多东西需要学习。特别感谢冯冠元师兄，是你在我接触课题之初给了我悉心教导，在我的课题遇到困难的时候给我悉心解惑，帮助我采集数据，处理问题。感谢哈工大通信所 1205 的所有师兄师姐，在我更换实验室的时候，给了我很多帮助，感谢贾爽师姐与杨松祥师兄，在一起做项目的时候一起讨论问题，一起解决问题，共同进步。感谢陈亮师兄，在我写文章的时候给了我很多帮助，传授给多很多经验，能让我写出更优秀的文章。

感谢通信一班的所有同学，在繁忙的学习之余，组织了许多活动，让大家一起放松，形成一个更加融洽的班集体氛围，让大家更加和谐相处。

感谢我的父母，感谢你们二十多年来的养育之恩。感谢你们在我忙于学业的时候做我最坚实的后盾，感谢你们给予我的无限阳光和力量，让我感受到了无私的父爱和母爱，我将带着你们殷勤期望和谆谆教诲，去迎接未来的挑战。