# 第一章 绪论

## 1.1 研究背景及意义

文字是人类用来记录特定的事物，简化图像而形成的一种书写符号。文字的发展经历了形意文字，意音文字和拼音文字的时代。文字的出现使历史事件不在是口口相传,而是有确切的记录，通过文字不仅能够使我们了解前人的思想、成就、经验，还能够使提高信息的传播效率，大大的加快了人类的文明发展进程。

最初的字符识别研究是随着计算机的发展而产生的，虽然文字记录相对于口语相传有一定的优势，但是纸质资料容易损毁，传输不便。字符识别的目的就是为了克服这些困难，它能够将大量的文字记录转成可修改，易保存并能够在计算机上处理的电子资料。

光学字符识别技术，指对文本资料的图像文件进行分析识别处理，获取文字及版面信息的过程，其解决了高速的计算机处理和低速的人工输入之间的矛盾，实现了文档的自动电子化。不管是在计算机还没出现的过去，还是电子设备层出不穷的现在，大量的信息都还是在印刷品或纸上，人工将这些信息输入到电子设备中是个困难的工作。光学字符识别有效的化解了信息化发展的瓶颈。经过几十年的发展，如今的光学字符识别技术已经能够处理大量扫描清晰，版面整洁，分辨率高的文档。但在复杂的情况下，如扫描时文档倾斜，夹杂了背景，文档分辨率低，受到了污染，退化等，传统的方法不能够得到很好的转化率。而这些状况在现实中是容易遇到的，研究更好的方法完善目前的光学字符识别方法有重要的现实意义。

最近几年，随着智能手机等移动设备的普及，人们可以通过智能设备自带的摄像头随时随地保存身边场景的图像信息，图像信息中通常存在大量的文本，这些文本作为场景中重要的信息来源对场景的认知和描述有重要的价值，所以这就极大的扩展字符识别研究的潜在需求。

自然场景中的文本有着天然的复杂和低质量，首先获取文本图像的设备多样，分辨率参差不齐，其次场景的多变导致背景变化多样，最后是拍摄的条件导致的光照，形变等。自然界的文本随处可见，如海报，包装，广告牌，路牌，商店名称，车牌等。其应用场合也是丰富多彩的，比如盲人导航通过手持设备，识别街道上的路标，门牌，再转化为语音来为盲人提供帮助。通过拍摄自然场景中的文字图像，可把图像中的文字转换成可编辑的文本资料，使它成为可交流的语义信息。

语义信息对于场景的认知具有极其重要的意义，而通过场景中的文本识别，我们可以轻易获得场景的语义，比如驾车通过一个学校区域，场景中有学校的建筑，大门，树木等目标，但通过识别这些目标是无法确定这是一个学校区域的。而如果识别了学校区域内的文本标识如学校名称，那么具备这个语义信息就能够很容易确定这是一个学校区域，这一信息对于我们驾车指导是有重要的价值，比如需要注意减速。利用智能设备或计算机自动识别场景中的文本信息，可以对智能交通管理，汽车自动驾驶，视觉感知辅助，危机预防和司法鉴定识别提供极大的帮助。从理论研究和实际的应用前景来看，复杂情况下的文本字符识别的研究有着重要的理论意义和实用价值。

## 1.2 文本定位国内外研究现状

在光学字符识别中，文本定位并不作为一个独立的步骤被提出，因为一般文档图像都是经过扫描仪得到的，图像中很少存在大范围的非文本区域。但在场景图像中，文本定位的目的是在场景图像中找到文本的位置，文本定位作为文本识别的先决步骤，在识别中占据重要的地位。类似我们人类，当我们关注场景时，并不是直接识别出一个个字符的，而是眼睛注意到一个整块的文本区域，然后对这一个区域进行字符的识别。文本定位的难点在于场景的复杂性和文本的多样性。场景的复杂性表现在场景中存在各种干扰，如窗子，草地，这些干扰通常和文本有着相同的纹理。另外场景中的光照变化也会导致文本区域光照不均，这会导致文本区域无法成为一个统一的区域。文本的多样性表现在文本的颜色，大小和形状不同。

从文档图像分析到场景图像分析，文本的定位也受到很多关注，从最先的CD，杂志封面文本定位到智能交通系统中的车牌定位，视频中的字幕提取，再到限制条件少，复杂背景下的场景文本定位，文本定位算法的鲁棒性越来越高，适用的范围也越来越广泛。文本定位的方法一般可以分为三种基于连通域的，基于学习和两者结合的方式。基于连通域的流程一般是提取候选文本区域，采用先验信息滤除部分非文本区域，然后根据候选文本字符间关系构造文本词。在提取候选区域方法上经历了2004年文献【1】提出的简单的连通域分析到2005年文献【2】提出基于canny的边缘提取，接着是2010年文献【3】提出笔画宽度变换（SWT：Stoke Width Transform），再接下来的2011年文献【4】【5】提出局部梯度特征和改进的SWT。滤除非文本区域的方式大部分采用的是先验信息如文本的高，宽，高宽比，占有率，致密性，边缘梯度等。构造文本词则是文献【1】利用的是字符间水平间距和垂直间距，文献【4】在水平和垂直距离的基础上增加了质心和颜色之间的限制，文献【6】则是采用聚类的方式，但是参数的设置是通过学习来得到的。基于学习的方式则是不同特征提取方法如纹理，小波，笔画以及分类器的使用如SVM（Support Vector Machine），AdaBoost。两者结合的方式一般在提取阶段采用连通域的方式，但是滤除阶段是通过训练样本学习分类器来实现非文本的滤除。

依据这种分类方式，对近几年的方法和文献做个介绍

1）基于连通域的方式

连通域分析在场景文本图像二值化后进行的，所以开始的研究集中在于场景文本图像的预处理，增强和二值化，文献【1】先将图像灰度化，然后在灰度图像上估计背景区域和前景区域，根据背景区域和前景区域的关系确定最后的二值化阈值，在得到二值化图像后，通过升采样的方式来增强图像，接着在增强图像上进行连通域分析，最后根据连通域的位置关系将文本进行定位。这一方法可以克服场景图像的光照不均，阴影，低对比度等情况，但对于复杂情况不能够得到较好的效果，因为这一方法没有考虑到场景文本特性。而在场景图像中，不管文本有不同的颜色，亮度，布局和方向，边缘是文本稳定的一个特征。文献【5】利用边缘的强度，密度和方向方差来生成特征映射图，特征映射图中的每个值代表的是文本的可能性。最后利用一个简单的全局阈值来分割特征映射图，这样就生成了一个文本区域为前景，其他为背景的二值图了。虽然边缘是文本的一个稳定特征，但是有很多干扰如树木，草地也同样存在着丰富的边缘。文献【3】提出了一种更加有效的边缘相关特性—笔画宽度，不像之前的方法一样，这一特性并不是存在于单个像素中而是存在一组像素中，梯度方向相反的两个像素间的所有像素都拥有相同的特征。相比边缘，这一特性能够大大的减少引进的非文本区域。基于连通域的方式很少需要在多尺度上进行操作，所以运算时间较快，但存在需要大量先验信息的弊端。

2）基于学习的方式

基于学习的方式一般流程是先将图像分割成一个个窗口，提取窗口中图像的特征，然后利用一个训练好的分类器来将窗口分成文本和非文本，最后文本区域连成一个文本行。文献【7】采用统计的方式分析了哪些图像的特征可以用来表征文本，然后利用这些文本的特征响应来得到一系列的弱分类器，这些弱分类器作为AdaBoost学习算法的输入来训练一个强分类器。作者采用了79个特征来训练4个强分类器。这79个特征包括4个亮度均值特征，12个亮度标准差特征，14个直方图特征和25个基于边缘链接的特征。文献【8】同样是采用AdaBoost来学习，但是在检测时采用的是融合了梯度方向直方图和多尺度局部二值模型的特征池替换上述的79个特征。文献【9】相当于结合了WaldBoost和CRF（Conditional Random Field）两个分类器来进行文本定位。在区域检测阶段，对于每个尺度的每个局部区域，提取HOG（Histograms Of Oriented Gradients）特征，利用WaldBoost得到分类结果，然后将分类结果转换成后验概率，依据这个后验概率生成文本置信度映射。采用全局阈值将映射图分割成连通区域图，对于连通区域图中的每个区域，采用CRF模型来标记为文本和非文本。文献【10】采用的文本的纹理特征，并利用SVM来进行分类。首先将图像分解成一系列固定大小的片段（像素），然后基于Haar小波分解的方式来提取纹理特征，滤除非文本片段采用的是SVM分类器，最后将文本片段构建成文本区域。基于学习的方式，计算量大，一般都在多尺度上处理，而且需要先准备好训练的数据来训练分类器。

3) 混合方式

两者结合的方式，一般都是分为两个阶段，第一个阶段是提取候选的连通区域，这个阶段一般采用的是连通域分析的方法。阶段二是文本区域和非文本区域的分类，这一阶段一般是采用分类器的方式实现的。文献【11】首先用局部自适应的方式来将图像分解成白、黑和灰三个层次，但是只在白和黑两个层次上获取候选的字符区域，对于每个候选的区域提取12个特征如边缘特征，笔画宽度特征等，SVM的训练数据包含4374个字符样本和4374个非字符样本，最后候选区域被分成字符和非字符。文献【12】采用最大极值稳定区域来提取候选的字符区域，候选区域如果有相同的几何特性就将其组合一个更大的区域，对于大区域提取水平垂直方差，笔画宽度，颜色等特征来创建AdaBoost分类器进行分类。两者结合的方式虽然没有众多的参数设置，但还是无法摆脱训练数据的限制。

## 1.3 文本识别国内外研究现状

文本识别是定位的最终目的，只有将文本识别后，才能够提供场景的一些信息，并且我们可以利用识别的结果来做更多事情。目前不管是企业还是学术机构都有很多团队在作相关方面的研究，在国外如google，国内如百度，腾讯。百度在百度翻译和百度识图中都有相应的应用模块，当然识别效果还是有待提高。

追踪相关的研究团队和文献，下面着重介绍几个团队的研究现状和成果，这些团队都是花费数年时间在这个课题上进行研究并取得一些不错的成果。一般的文本识别的研究可以分为三种，一是单个字符的识别，这时场景中的文本都已经分割成单个字符，目的是正确的字符分类，这一研究的主要目的是测试特征提取和分类器的有效性。二是词识别，这个阶段的识别是在文本定位的基础上进行的，主要是分割算法的研究和词典信息的利用。三是端到端的识别，这种研究没有明显的定位过程，定位和识别都在一个框架中完成的。

1）端到端文本识别

文献【13】【14】，作者建立了SVT（Street View Text）的数据集，这些数据集收集自Google地图的街景图，数据集中的文本图像存在很大的变化，并且分辨率都不是很高。同时作者也将数据集和其他数据集进行比较并提出一种新的词识别方法，这个方法采用的是一般的物体识别技术框架。主要包括两个流程分别是字符识别和词配置，字符识别采用的是多尺度滑动窗口的方式，提取的是每个字符的HOG特征，分类器采用简单的随机蕨。对于需求的大量训练数据，采用人工添加一定数据的高斯噪声和随机仿射变形的方式来得到的。利用图结构模型来进行词识别，将字符识别阶段得到的分数和位置输入图模型中，通过优化图模型得到最佳词配置。

后续的很多研究也都基本建立在这个流程上，文献【15】【16】采用自顶向下和自底向上的两个线索来识别场景中的文本。自底向上的信息是由图像的单个字符检测得到，自顶向下的信息则是基于词典的先验信息。这两者通过一个CRF模型来进行综合的。不同于前述方法，对于滑动得到的窗口，采用基于分类器分数和训练数据学习到的字符长宽比来滤除明显非字符的部分，另外词典信息的先验信息和字符的分类器分数，位置是通过创建和优化CRF模型来进行融合的。另外字符识别的分类器利用的是SVM。

滑动窗口的方式容易导致类间的混淆，字符的类数越多，相互间识别错误的可能性就越大，并且多尺度滑动窗口会是识别时间大大增加。文献【17】【18】实现了实时的文本定位和识别。同样将识别分成两个阶段，在第一个识别阶段，提取出图像的所有极值区域，通过提取区域计算量很小的特征如高宽比，水平跨越特征，致密性和欧拉数等判断是否是一个字符。第一阶段得到结果是每个区域为字符的可能性，而只有那些有最大可能性的极值区域被用在第二阶段。第二阶段提取的是耗费时间更长的特征，另外采用SVM分类器来进行分类。在得到为字符的极值区域后，首先要训练一个分类器，训练分类器的数据直接是来至于Windows操作系统上的字体。提取的是字符的极值区域边缘像素方向，对每个方向用25维的特征表示，总共采用8个方向，最后一个字符就表征为200维的向量。同样是用RBF（Radial Basis Function）核SVM来进行分类。

文献【19】【20】从另外一个角度出发，结合大规模，多层次的神经网络和非监督特征学习的方式来训练高精度的文本检测和字符识别模块。这里神经网络采用的是卷积神经网络。应用滑动窗口的方式可以得到一组文本的候选行。根据文本行上的检测响应，可以估计字符在文本行中的位置，然后结合字符位置和字符识别响应结果通过束搜索来得到最终结果。

2）裁剪字符识别

文献【21】【22】【23】关注的是文本识别中的裁剪字符识别，这些研究主要包括对目前的局部特征和不同学习方式进行评估。文献【21】作者构建了两个数据集分别是场景文本图像数据集和印刷手写数据集。并且在数据集上评估了SC (Shape Contexts)，GB（Geometric Blur），SIFT（Scale Invariant Feature Transform），Spin image等特征，另外也评估了三种分类策略NN（Nearest Neighbor），SVM，MKL（Multiple Kernel Learning）。文献【22】提出一种基于SIFT的改进特征SSIFT（Shape-SIFT），这个特征的特点在于结合全局的形状特征和局部的SIFT。文献【23】没有采用基于学习的方式，而是物体匹配的方式。这个主要包括模板图像库的建立，待识别图像特征的提取和模板图像的匹配，对于多个字符区域也是采用多尺度滑动窗口的方式进行的，最终结果是投票和几何验证产生的。

从前述方法中，我们看到虽然目前的方法能够识别场景文本，但大部分方法都是属于相同的框架的，大量词典信息的应用是提高正确率的关键，裁剪字符识别正确率不是很高，这说明识别正确率的瓶颈在于词图像分割成单个字符的过程，后续的研究就在于分割和裁剪字符的识别。

## 1.4 本论文的主要工作及组织结构

本文主要是研究如何将场景中的文本图像转化成机器可识别的语义信息。文章分为三个重点来进行研究分别为场景图像的分析和预处理，场景文本的检测和定位，场景文本的识别。

论文章节安排如下：

第一章，绪论。主要介绍课题研究的背景和意义,并从场景文本定位和场景文本识别两个方面来进行国内外研究现状的介绍。这包括研究的团队，相关的算法和数据集等的介绍。

第二章，复杂条件下的文本图像预处理。首先对场景文本图像从不同方面进行分析得到场景文本的特性，然后从颜色，噪声，角度，灰度等方面出发对复杂条件下场景文本图像进行预处理以便后续的处理。

第三章，自然场景条件下的文本区域定位。文本区域定位是文本识别的先决条件，在这一章中，主要是设计了一个文本区域定位系统，用不同的两种方法进行候选连通区域的提取，然后用视觉显著性的评估来滤除非文本区域，最后采用公开发表的数据集和评价标准对算法进行评估。这一章重点在于尽可能多的提取候选文本区域，对于引进的非文本区域要考虑先验信息需求少，计算时间短的方法。

第四章，复杂条件下的文本识别。这一章首先探讨了定位和识别的关系。为了充分应用文本定位的中间结果和简化识别流程，本章将文本识别分割成词图像分割，特征提取，分类器选择，后期融合等几个过程。对于每个过程都使用了不同的策略，评估选取最佳方式。这一章的重点在于如何简便的进行词分割和更好的表征字符以及合适分类器和参数的选择。

第五章，总结和展望。这一章主要是对本文工作进行总结包括本文的贡献，不足和对未来继续研究这一领域应该更加关注的问题。