电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER DISSERTATION

论文题目 基于几何约束的笔划宽度变换（SWT）算法及其字幕文本定位应用

学科专业	模式识别与智能系统
学　　号	**201221070509**
作者姓名	袁俊淼
指导教师	程　洪　　教授

分类号 _____ 密级 _____

UDC [注1] _____

学　位　论　文

基于几何约束的笔划宽度变换（SWT）算法及其字幕文本定位应用

（题名和副题名）

袁俊淼

（作者姓名）

指导教师　程洪　教　授

电子科技大学　成　都

（姓名、职称、单位名称）

申请学位级别　硕士　学科专业　模式识别与智能系统

提交论文日期　2015.05.05　论文答辩日期　2015.05.18

学位授予单位和日期　电子科技大学　2015 年 6 月

答辩委员会主席　赵辉

评阅人　杨路　郝家胜　高斌　金卫

注 1：注明《国际十进分类法 UDC》的类号。

STUDY ON STROKE WIDTH TRANSFORM WITH GEOMETRIC CONSTRAINTS AND APPLICATION IN NEWS CAPTION TEXT LOCATION

A Master Dissertation Submitted to

University of Electronic Science and Technology of China

Major: **Pattern Recognition and Intelligent System**

Author: **Junmiao Yuan**

Advisor: **Prof. Hong Cheng**

School: **School of Automation Engineering**

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：_____ 日期：　　年　　月　　日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期：　　年　　月　　日

摘　要

随着移动智能设备的发展，对自然场景图像中文本的提取和理解的需求越来越大，其结果可以广泛的应用于社会机器人导航和交互、图像检索等领域，然而传统的 OCR 技术只能分析扫描文档，无法直接应用于自然场景，关键原因就在于自然场景图像文本定位的问题，在自然场景图像中文本和非文本混合在一起，而且文本区域出现的位置随机，这给文本定位带来了很大的挑战。本文研究了自然场景图像文本定位技术，利用全局性特征度量和几何约束改进了传统笔划宽度变换算法，实现了文本定位整套系统，并在新闻视频字幕文本定位中进行了应用。

本文的主要贡献有以下三个方面：

首先，针对自然场景图像文本区域通常具有高视觉显著性、富边缘密度和颜色一致性的特性，本文提取了候选字符和文本行的全局显著性和边缘密度特征。此外，将文本区域的位置和尺寸信息也转换成了全局性的度量，弥补笔划宽度变换算法对局部噪声敏感的缺陷。

其次，本文在传统的笔划宽度变换算法基础上进行改进，利用几何约束降低了候选字符黏连的情况。本文采用的的几何约束规则融合了笔画的宽度、颜色和方向特征，避免在笔划射线查找过程中由于边缘部分缺失而连接到非对称性边缘点上，从而形成黏连字符的情况。相比于传统的笔划宽度变换算法，对由于噪声、模糊、低对比度等造成的边缘缺失的情况下，几何约束笔划宽度变换提取出的笔划特征值比原始的笔划宽度变换算法更准确，形成的候选字符质量更高。由于几何约束减少了无效笔划射线的数量，使得候选文本像素更多的集中在文本区域，减少了非文本像素形成的候选字符区域，降低了字符水平和文本行水平过滤的复杂度，同时也避免了误判。

最后，基于上述研究成果，结合新闻视频图像的特点和字幕文本排列的规律，本文将基于几何约束笔划宽度变换算法的文本定位在新闻视频字幕文本定位系统上进行了应用。

关键词：自然场景，几何约束，笔划宽度变换，文本定位

ABSTRACT

With the development of mobile intelligent device, the demand of text information extraction in natural scene images increases sharply. The text information in natural scene images can be greatly helpful to social robots navigation and interaction, image retrieval and other fields. However, traditional OCR engine can only handle scanned documents and the results are bad when applied to natural scene images. The key issue lays on the text location, which is very challenging because of the mixture of text and non-text regions and the randomness of location of text. This paper studies the text location techniques in natural scene images. We improve the traditional stroke width transform with geometric constraint and global feature measurement of text area. On the basis of text location, we build a news captain location system.

The main contributions of our work are as following:

Firstly, based on the properties that text regions regularly present high visual saliency, abundant edges and consistent color. We detect the salience, vertical edge density to calculate the feature of candidate components and text lines in the whole image. Besides, we also convert the position and scale information of candidate text area to global measurement to compensate original SWT sensitiveness to local noise.

Secondly, we utiliz geometrical constraints which combines the feature of the width, color and direction characterization of stroke to improve the tranditional SWT. In the process of SWT, we limit the formation of stroke, avoiding incorrect links between asymmetrical edge points. Compared with the traditional stroke width transform, our method performs better when edge is partly missing and is more robust to noise, blur and low contrast in natural images. By the means of geometrical constraints, our method clusters the more candidate text pixels in text regions and reduces the production of the non text components, which reduces the complexity at the character level and text line level filtering and avoids misjudging the ambiguous candidates.

Based on the above research results, we apply the geometric constraint stroke width transform in the news caption location system in conjunction with the video image properties and the distribution rules of news caption text.

Keywords: natural scene, geometric constraint, stroke width transform, text location

目 录

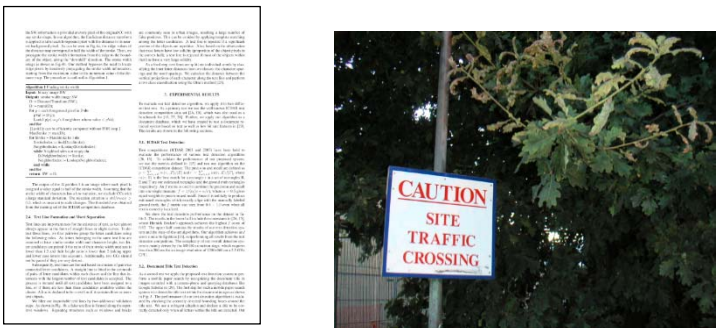
第一章 绪 论

1.1 研究背景及意义

文字作为人类的一项重要发明，在人们日常生活中发挥了重要的作用。自然场景图像中的文字，是人类思想抽象的简要概括，富含丰富的语义信息，是一个十分重要的信息源。

随着数字媒体技术，尤其是移动智能设备的快速发展，越来越多的文字信息以图像和视频的形式呈现，图像和视频中的文本包含了十分有用和重要的信息，这些信息可以很好地帮助我们理解图像和视频的内容，进而广泛地适用于图像和视频的自动分类、图片搜索[1]、机器人自动导航[2]和目标定位[3]等领域。例如自然场景中的指示牌、路牌等，通过手机拍照获取后，经过准确地定位和识别提取出文字信息，再结合机器翻译和语音合成等技术可以极大地方便人们在异国的旅行；数字图书馆的图像和视频中的文字包含了丰富的信息，如果能够自动地提取这些文字信息，并进行索引，则极大地提高人们对数字资源的利用率。总之，自然场景图像中的文本包含了丰富和精确的高层语言描述，提取这些信息对人们的生产和生活意义十分重大。然而现状是一方面我们面临着海量的图像和视频信息中的文字等待处理，另一方面却受限于现有的文本识别技术对自然场景图像中的文本识别效果太差的现实，因此如何解决这两者的矛盾，吸引了无数研究者投入到这个工作中来。

经过近些年来的发展，传统扫描文档的信息提取问题已经得到了很好的解决，也催生了很多基于扫描文档文字信息提取的应用，而自然场景图像的文本信息提取依然是一个很大的挑战。自然场景图像文本定位是自然场景图像文本信息提取的基础和关键环节，在基于内容的信息检索应用中极其重要，自然场景中的文字信息为理解多媒体内容的高层语义提供了重要的线索，其应用前景[4]十分广阔。最近十年在学术界和工业界也受到了越来越多的关注。虽然在这方面很多研究者都做出了很大的努力来解决这个问题，然而自然场景图像文本定位依然未得到很好的解决。问题的关键在于文本的表现形式更加复杂，具体体现在文本的多样性和背景的复杂性两方面。与传统的扫描文档相比，自然场景图像中的文本在字体的尺寸、形状、文字出现的位置和文字背景上更多样、更复杂。例如：不同场景中文字在字体、尺寸、形状上差异很大，在很多情况下由于拍摄的角度问题还可能发生扭曲和遮挡。此外，文字背景上的一些相似文本的区域，比如砖墙、窗户、长线条和树叶等也给自然场景图像文本定位带来了很大的困难。



(a) (b)

图 1-2 自然场景图像和传统的扫描文档图像：(a)扫描文档图像；(b)自然场景图像

1.2 文本定位面临的挑战

自然场景图像下的文本定位问题是计算机视觉和数字图像处理领域十分具有挑战性的课题，主要体现在以下几个方面：

传统的扫描文档布局相对稳定、背景单一、亮度均匀、字体规整，和传统的扫描文档图像相比，自然场景文本图像具有以下特点：

（1）文本多样性

自然场景图像中的文本在字符的尺寸大小、语言、字体，方向和颜色变化大，存在着艺术字体等情况，在一些自然场景图像中字体尺寸的差异甚至达到几十倍，此外也存在着多种语言的文字在一幅自然场景图像中同时出现的情况，另外由于拍摄的角度有时并不是正视，因此自然场景图像中还广泛存在文本字符的透视变形问题，如图 1-3 所示。

图 1-3 自然场景文本多样性

（2）背景的复杂性

自然场景图像的背景中有时存在着诸如树叶、窗户、草地等类似文本的区域，视觉上很难将其和真实文本区域进行区分，一些在玻璃橱窗上出现的文字由于和背景重叠，定位难度更大。另外自然场景中还存在复杂的文本极性判断这个特殊情况，例如暗文本在亮背景上，亮文本在暗背景上，如图 1-4 所示。

图 1-4 自然场景文字背景的复杂性

除了以上自然场景图像中文字和背景复杂多样的情况，在图像处理领域由于拍摄设备和环境条件不佳而造成的光照不均匀、噪声严重、图像模糊、低对比度、部分遮挡等问题在自然场景图像中依然存在。此外，自然场景图像本身质量比较差也给文本定位带来了很大的困难。如图 1-5 所示。

图 1-5 自然场景图像中存在的传统图像处理问题

1.3 国内外研究现状

1.3.1 研究现状

文本定位是确定字符、单词、文本行或者文本块等文本区域在图像中的位置，根据定位的精度选择不同的文本区域的外接矩形来表示。自然场景图像文本定位是自然场景文本信息提取系统中十分关键的一步，定位精度的高低直接关系到后面文本分割和文本识别的结果。也有一些文献将文本定位细分成文本检测和文本定位两个部分，然而无论是几部分，自然场景图像文本信息提取系统必须首先在自然场景图像中确定文字区域的位置，但文本定位存在诸多的难点，比如语言的种类、文字的颜色、分辨率、背景的复杂程度、光照的变换及倾斜和扭曲等都对文本定位有很大的影响。

自然场景图像文本定位的研究，尤其是复杂背景图像下文本定位研究是一个比较新领域，但由于其存在巨大的理论研究意义和商业价值，吸引了许多科研机构和商业组织。自然场景图像文本定位的研究最早从 Zhong 等[5]在杂志封面上进行文本定位研究开始，他们采用边缘对的方法得到候选文本区域，再采用颜色聚类的方式得到文本区域，这种方法在简单的杂志封面上应用效果可以，但在低对比度、复杂布局和颜色变化很大的图像中该方法失效。

国际文档分析识别大会（ICDAR）在 2003 年举办了首届专门的文本定位评测比赛，在此之前，各个研究者采用的数据集和方法都不一样，缺乏统一的数据集和评价方法，难以衡量。国际文档分析识别大会（ICDAR）从 1991 年开始，由国际模式识别协会（IAPR）举办，是自然场景文本检测领域中最重要的会议之一，每两年举办一次。主要关注以下三方面的进展：页面分割（page segmentation）、手写识别（hand-writen recognition）和粗壮阅读（robust reading），其中粗壮阅读（robust reading）又细分为文本定位、文本分割、单词和字符识别等细项，该项挑战是用来衡量自然场景文本识别系统的最新进展。自 2003 年起该赛事已经举办了 4 届，分别是 ICDAR2003、ICDAR2005、ICDAR2011 和 ICDAR2013，其比赛的样本库和评价方法都在网络上进行了公布，参赛者可以根据自己的情况选择一个挑战项目或者多个挑战项目参赛。尤其值得一提的是在 ICDAR2013 Robust Reading Competion 中，北京科技大学的殷绪成博士团队获得了自然场景文本检测（Text Localization in Real Scenes）、网络图片文本检测（Text Localization in Born-Digital Images (Web and Email)）、网络图片文本提取（Text Segmentation in Born-Digital Images (Web and Email)）三个比赛项目的冠军，在性能方面取得了较大的提高。

总之，虽然这些年自然场景图像文本定位技术取得了很大的进展，但能满足

实用要求的高准确率和召回率并对不同场景具有较强鲁棒性的文本定位系统仍然没有开发出来，文本定位作为自然场景文本信息提取系统的基础和最重要的一环，其定位效果是文本信息提取系统的关键。近些年来，该领域在国内也蓬勃发展，一些研究学者也取得了不错的成绩。

1.3.2 研究方法

文本定位的基础是数字图像处理，涉及到模式识别、机器学习等多个学科，根据研究所选取的特征，文本定位主要分为三类：基于边缘的文本定位方法、基于纹理的文本定位方法和基于连通域的文本定位方法。近年来随着机器学习的兴起，还有一些文章中提出基于学习的文本定位方法，例如文献[6,7]，其本质是把机器学习的方法应用到文本定位中来，但选取的特征并没有改变，本文依然把其归为以上三种方法中。

（1）基于边缘的文本定位方法

为了便于人眼识别，自然场景图像中的文本和背景颜色对比度通常都比较高，因此文本和背景之间一般存在着较强的边缘。基于边缘的文本定位方法利用的是文本区域的高边缘密度特性来区分文本区域和背景区域，前期可以利用 Canny、Robert 和 Sobel 等边缘检测算子进行边缘检测，然后通过形态学等方法进行区域的分裂和合并提取出候选文本区域，最后通过先验知识或者基于机器学习的分类方法进行过滤。

Kim 等[7]提出的方法中首先提取图像的边缘特征，然后经过长直线剔除，接着使用形态学的开操作滤除细小的连通域，弥补连通域内部的间隙，形成初步的连通域，最后使用先验知识过滤得到最终的文本区域。Chen 等[8]采用类似的方法，但把机器学习的方法引入到文本定位中，他们先利用边缘和形态学的方法得到初步的候选区域，然后利用基线扫描方法将候选区域分成文本行，最后使用 SVM 分类器进行验证，得到最终的文本区域。

欧文斌等[9]首先利用金字塔对图像进行分解，并将分解的子图灰度化，然后在每幅灰度图像上利用垂直 Sobel 算子进行边缘检测，接着使用边缘密度函数粘结边缘形成候选连通域，最后使用连通域分析方法确定文本区域的位置。当自然场景图像中的文本行紧密分布时，该方法取得了不错的效果，但当自然场景图像背景区域的边缘特征比较丰富时，其定位效果比较差。

Gao 等[10]提出的算法首先进行边缘检测，然后利用边缘的上升和下降等属性及尺寸信息对连通域进行聚类，得到初步的文本区域，在他们的方法中，为了消除颜色渐变对字符抽取的影响，采用了最大期望（EM）算法对候选文本区域进行建

模分析，最后利用版面分析找回漏检的文本区域，剔除类文本区域，从而得到最终的文本定位结果。

实验表明，基于边缘的文本定位方法计算简单、实现相对容易，在背景比较简单的自然场景图像下取得了较好的效果，然而基于边缘检测的文本定位算法也有着致命的缺陷：第一，当字体尺寸比较大时，尤其是在一张图片上字符的尺寸变换范围很大，文字和背景的边缘的密度信息区别就没有那么明显，会大量漏检文本区域，此外边缘特征对于不同语言的的不同字体也比较敏感，缺乏适应性；第二，背景复杂的情况下，文字区域和背景区域都有丰富的边缘，这样很容易把背景区域误检为文字区域。

鉴于以上的考虑，基于边缘的文本定位方法往往不单独使用，而是作为文本区域的初步筛选，然后结合其他的特征信息完成最终的文本定位。

（2）基于纹理的文本定位方法

基于纹理的文本定位方法把文字区域当成一种特殊的纹理，通过用滑动窗扫描图像，然后根据事先设计好的窗描述器的滤波结果判断当前滑动窗内是文本区域还是非文本区域，最后通过区域分裂和合并等操作得到最终的文本区域。基于纹理的方法的思想是基于一系列的特性，比如高边缘密度、文本和非文本的低梯度及高灰度方差、小波变换的分布和离散余弦变换的系数等的差异区分文本区域和非文本区域。这种方法的局限性在于因为要在多尺度上扫描图像，计算复杂度高，对多尺度上得到的信息融合缺乏固定的模式，此外这种方法对变形文本字符检测效果很差。基于纹理的文本定位方法采用的常见的纹理特征有角点，统计特征中直方图、方差、梯度分布、频谱分布，几何结构特征中的宽高比等。在基于纹理的定位方法，描述器的设计非常关键。例如方向梯度直方图（Hog），该描述器已经在人脸识别和行人检测上取得了较好的效果，然而与人脸和行人检测相比，自然场景图像中的文本则具有更大的复杂性，例如旋转、透视、伸缩变换等，使得一般的描述器很难区分文本区域和背景区域。这样如何设计一个合适的描述器来区分文本区域和背景区域就成为了自然场景文本定位是否能够成功的关键。此外滑动窗的方法由于要对整个图像进行滑动扫描，当为了增强对多尺度文本定位效果的时候，往往采用金字塔分解的方法，在多个尺度上进行扫描，因此具有很大的时间复杂性。对一张具有 N 个像素的图像，窗口的数量达到了N^2个，这使得在实际应用中变得不可能。基于纹理的定位方法相对于基于边缘的文本定位方法优点是在复杂背景下也能很好地检测出文本区域，但当背景区域和文本区域具有相似的纹理的时候，该方法失效。另外由于该方法的数据处理量大，实时处理效果不好，很难在移动设备上使用，只能作为辅助检测手段。

Zhong 等[11]提出了一种彩色图像文本定位算法，该算法利用水平空间差分来粗定位文本，然后在定位到的文本区域内进行颜色分割得到文本字符。该方法对字体颜色和背景颜色对比度高且字体颜色一致的自然场景图像表现效果较好，但由于只采用了颜色特征，缺乏鲁棒性。

Li 等[12]提出了一种针对视频的文本检测和跟踪算法，他们利用固定尺寸的滑动窗扫描图像，提取滑动窗内的灰度均值和一阶中心距和二级中心距等统计特征训练人工神经网络，利用训练好的模式筛选候选文本区域，然后利用连通域分析技术得到文本定位的结果。Mao 等[13]提出的算法也是采用纹理特征，只是用小波分解的能量特征代替中心距特征，他们首先抽取小波分解后的局部能量变化特征，然后进行二值化处理得到二值图，接着使用连通域分析技术得到文本定位的结果。

Liu 等[14]提出利用笔划滤波的方法进行文本定位，其思想是文本由不同的笔划组成。他们设计了对条状对象敏感的笔划滤波器，其滤波响应包含纹理、边缘和角点特征等信息，能够较好的反映文本的本质特征，然而该方法对笔划宽度的设置很敏感，不同的字体尺寸需要设置不同的笔划宽度，缺乏适应性，另外该方法仍处于起步阶段，存在不少问题。

Chen 等[15]为了解决基于纹理的定位方法时耗特别严重问题，实现了一种快速的文本检测器，该检测器采用级联的 AdaBoost 分类器，每个弱分类器从灰度方差、梯度直方图等特征集中选择特征进行训练，实验表明，该算法的效率得到了很大的提高，但其定位的精度在自然环境中没有达到理想的效果。

Lyu 等[16]为了解决多语言文字的定位问题，提出了一种由粗到细的多尺度搜索算法，该算法使用了边缘强度和文本区域的高对比度特征来区分文本区域和非文本区域，此外该文献中还提出了一种局部自适应二值化算法来分割定位到的文本图像，然而和其它基于纹理的方法相似，该算法需要人工设定很多的规则和参数，很难推广到复杂的自然场景中使用。

（3）基于连通域的文本定位方法

基于连通域的文本定位方法[17-19]的思想是把呈现出一定相似特性，例如相似颜色的像素聚合在一起形成候选字符，接着通过一系列过滤规则的限制滤除非文本字符。基于连通域的方法非常具有吸引力，因为它能检测任意尺度的文本而且对文本方向的要求也不局限于水平方向。基于连通域的方法相对于基于边缘和纹理的方法更鲁棒、计算速度快，而且具有尺度和方向不变性。

基于连通域的文本定位方法的基本流程如下：首先使用低水平的滤波剔除大量的背景像素，然后在剩下的像素中使用一系列启发式学习方法，例如笔划宽度的连续性、颜色的同质性等，来构建候选文本区域，最后用启发式方法和几何特

性等特征过滤非文本字符。这种方法的优势在于大大地降低了时间复杂性。此外检测到的字符也提供了直接进行字符分割的信息，这对进一步应用文字识别提供了极大的便利。然而基于连通域的文本定位方法有着明显的缺点：在低水平的滤波对图像噪声和字符扭曲很敏感，易导致错误的候选字符聚类。其次，使用启发式的学习方法涉及到一系列的手动参数调节，可能造成在一种测试数据库上取得很好的效果的方法却在另外一个数据库上表现欠佳。最后，启发式的方法对相似文本区域的区分性不强，很容易造成误检，因此往往不单独使用，而是作为初步过滤使用。

Jain 等[18]利用颜色聚类把图像分解成不重叠的候选区域，然后通过连通域分析技术把候选区域聚合成文本行，接着通过几何规则等先验知识过滤掉非文本字符，由于需要手动设置很多参数，该方法在复杂的自然场景图像中的检测效果并不好，但该文章提出的思想对以后基于连通域的文本定位算法影响很大。

一些文献基于同一文本字符颜色相似的假设，首先将颜色空间量化，如江斌等[20]采用图理论方法在颜色空间进行聚类，章东平等[21]则通过寻找彩色图像直方图峰值的方法进行颜色空间量化，然后在颜色量化后的空间平面进行连通域分析，最后融合各个子空间的定位结果得到原始图像中的文本区域。这些方法使用范围广泛，但在低分辨率低、复杂背景下的自然场景图像上，定位效果不好，而且这些算法基于单一的文字颜色一致性的假设，在自然场景图像中有时并不满足。Zhou 等[22]提出的算法中首先使用颜色聚类得到候选文本区域，然后利用连通域分析技术对每个连通域进行验证，剔除非文本区域，最后通过连通区域合并算法将经过验证后的小区域合并成文本行，其缺陷和江斌等人的相似，由于只采用了颜色特征，对复杂背景图像下的文本定位效果不好。

Epshtein 等[23]提出的基于笔划宽度的自然场景图像文本定位方法，其思想是基于文本字符笔划的宽度一致性的特点，通过检测边缘和计算笔划宽度特征，然后采用自底向上的聚合和自顶向下的剪枝方法进行文本定位。该方法对语言的种类、字体的颜色、尺寸都具有较好的鲁棒性，在自然场景文本检测上取得了较好的效果，但在边缘缺失的时候，很容易造成字符黏连，导致部分文本字符被剔除，对最后的文本行聚链产生严重影响。

Neumann 等[24]提出了一种基于极值区域（Extremal Regions）的自然场景文本定位方法，利用极值区域对图像模糊、光照不均、低对比度和颜色和纹理渐变的鲁棒性，把自然场景文本检测问题转换成从极值区域集中按顺序选择的问题。其原理是先利用极值区域提取出候选文本区域，然后利用穷举搜索算法把每个通道的极值区域合并成最终的文本行。由于需要处理多通道问题，复杂度相对较高。

Yao 等[25]在笔划宽度变换的基础上提出了多方向聚链的方式，能够检测任意方向的文本，虽然取得了较好的效果，但其采用的仍然是局部的笔划宽度特征，对噪声的抵抗力不强，在边缘缺失等情况下容易形成字符黏连，导致候选文本字符的生成质量不高，进而对整个文本定位结果产生很大的影响。

Yin 等[26]利用剪枝算法计算出图像的最大稳定极值区域（Maximally Stable Extremal Regions），然后利用最小化正则变分的策略得到文本字符，接着通过自学习距离方阵得到聚类的单链聚类算法的距离权重和阈值，最后利用单链聚类算法将候选字符聚成文本行。

类似的基于连通域的文献还有文献[27,28]。

由于自然场景文本的复杂性，单独利用一种特征来区分文本和非文本虽然可以在特定的场景下取得比较好的效果，但缺乏推广能力。其发展趋势是偏重于多特征融合。

Tekinalp 等[29]结合图像的颜色、纹理和对比度等信息分别进行文本定位操作，最后利用规则整合三种定位结果得到最终的文本区域位置，该方法在融合多特征上做了重要的尝试，在简单背景上取得了较好的效果。

Chen 等[30]首先通过对文本区域的统计分析确定哪些特征更能体现文本区域的属性并具有对所有文本区域的响应相似的低熵属性，集成学习的方法进行特征选取。然后通过有限元概率分析选取的特征在文本和非文本上的响应，然后将选取的文本特征作为 AdaBoost 分类器的输入进行训练。该文献提出的方法对特征的选择具有重要的参考价值。

Kim 等[6]提出的算法中分别利用灰度和彩色边缘、颜色聚类方法进行文本的初定位，然后整合文本定位的结果对重叠的部分保留，不重叠的部分利用 SVM 进行分类，最终得到所有的文本区域。和 Kim 等人的方法相似，但 Liu 等[19]利用弹性边缘检测算法得到所有可能是文本区域的边缘像素，然后利用梯度和轮廓的几何特性来得到候选文本区域，最后通过纹理分析的方法剔除非文本区域。

Neumann 等[31]提出了一套自然场景文本检测和识别系统，该方法结合了基于滑动窗扫描的纹理的检测方法和基于连通域的检测方法的优势，该方法认为字符区域是在相对位置上具有特定方向的笔划的图像区域。而笔划的获得在该文献中采用了图像梯度域和一些固定方向的条状滤波器卷积得到。该文献对文本区域的特征设计给出了重要的参考，但条状滤波器宽度的选取需要大量的手动调节，缺乏推广能力。

本文借鉴基于多特征融合进行文本定位的经验，针对传统笔划宽度变换中存在字符黏连和局部敏感的缺陷问题进行改进，在选取文本固有的笔划宽度一致性

特征基础上将笔划的宽度、颜色和方向特征相结合，利用几何约束对在笔划线的生长过程进行限制，从而提高候选字符生成质量，使得文本像素更多的集中在文本区域，同时消除字符黏连等情况。在弥补笔划宽度变换的局部敏感性上，把文本检测的结果转换成文本区域的全局性度量，最后结合连通域分析、字符聚链等方法完成自然场景图像文本定位。

1.4 研究内容和章节安排

经过对国内外的研究现状进行调查发现，现有的自然场景图像文本定位算法中更多关注字符水平的分类，很少涉及文本行的过滤，对像素水平的过滤研究则更少。针对以上的问题，在广泛阅读文本定位的相关文献和认真对比分析各种特征提取算法后，本文分析了传统笔划宽度变换算法的优缺点和像素水平过滤的思想，在此基础上融合几何约束的规则，对像素水平的过滤进行了改进，该方法继承了基于纹理的定位方法对背景复杂的鲁棒性和基于连通域分析的方法的计算高效性的优势，同时克服两者的一些缺陷。此外，本文结合新闻视频字幕文本定位的具体应用，在基于几何约束笔划宽度变换的文本定位的基础上，设计和实现了新闻视频字幕文本定位系统。

本文的结构安排如下:

第一章：介绍自然场景图像文本定位的研究背景和意义、面临的挑战和国内外的研究现状及采用的主要研究方法，最后对本文的研究内容和章节结构安排进行了简要的介绍。

第二章：介绍了自然场景图像的显著性分析、边缘密度分析和利用 MSER 特征提取候选字符的方法，然后介绍了采用投影分析及窗扫描的方法进行候选文本行生成的方法，最后重点介绍了本文如何将文本检测的结果转换为候选字符和文本行的全局性的特征，弥补笔划宽度算法的局部敏感缺陷。

第三章：首先介绍了传统的笔划宽度变换算法的原理和处理流程，在分析传统的笔划宽度算法的缺点后，提出了改进的基于几何约束的笔划宽度变换算法。在完成了基于笔划宽度变换的字符水平的像素过滤后，重点讲解了候选字符的生成及字符水平和文本行水平的过滤算法，最后介绍了利用随机森林进行字符水平和文本行水平过滤的思想，并给出了实验结果。

第四章：在文本定位的基础上，结合新闻视频字幕文本定位的具体应用，设计和实现了新闻视频字幕文本检测系统。

第五章：总结全文，得出结论并对下一步的工作进行展望，最后对文本定位和进一步的自然场景图像文本信息提取应用前景进行了展望。

1.5 本章小结

本章首先介绍了自然场景图像文本定位的研究背景和意义，接着指出了自然场景图像文本定位面临的挑战，然后介绍了国内外在自然场景图像文本定位方面的研究现状和采用的主要的研究方法，最后针对目前存在的问题，最后对文本的研究内容和章节结构安排进行了简要的介绍。

第二章 基于文本检测的全局特征提取

文本检测在整个自然场景图像文本定位框架中一般作为粗过滤阶段，功能是大致确定文本区域的位置，其结果往往作为文本定位的输入。传统的方法通常是在文本检测的基础上再精定位，但本文创新地将文本检测的结果转换为一种全局性的特征度量应用于候选文本字符和候选文本行的筛选过程中，此外，本文利用最稳定极值区域（MSER）生成候选字符，利用窗扫描和文本块合并技术得到候选文本行，分别作为第三章基于几何约束笔划宽度变换的文本定位字符水平和文本行水平的输入。文本检测在整个文本定位系统框架的功能如图 2-1 所示。

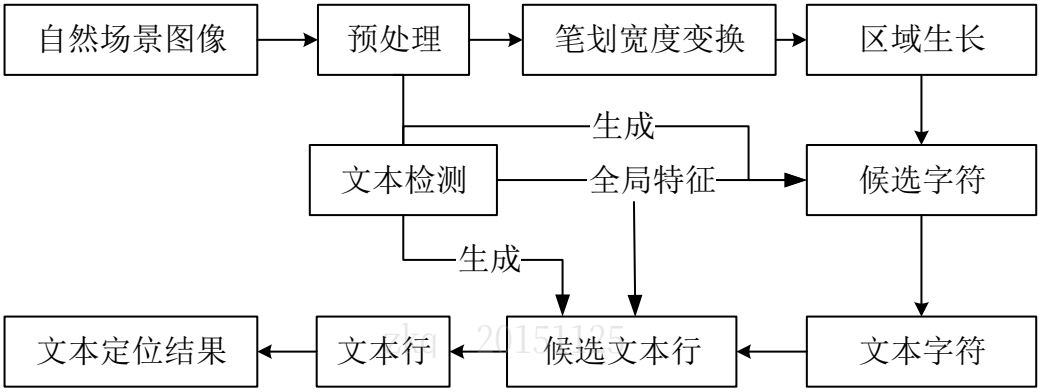


图 2-1 文本检测在文本定位系统中的功能

2.1 文本检测

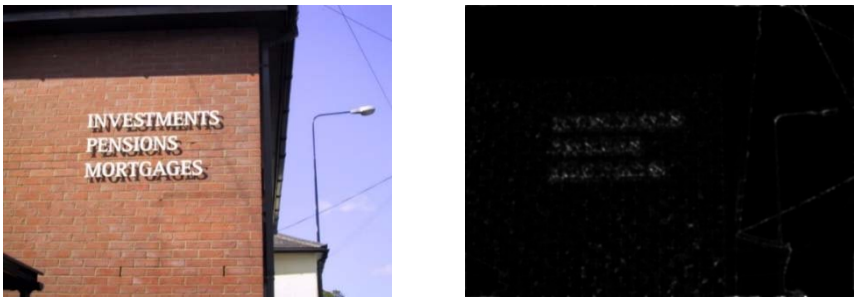
2.1.1 全局特征提取

（1）显著性分析

显著性是一种模拟生物视觉机制的注意力建模模型，代表性的是 Itti[32]提出的自顶向下的显著性计算框架。显著性分析很适合自然场景下的图像处理，尤其是在复杂背景下表现相对较好。显著性的计算往往是从像素点与周围背景在亮度值、颜色值、方向和纹理等方面的差异出发。而在自然场景中，为了突出文本区域和便于识别，往往满足这样的对比度差异，显著性检测主要包含特征提取和显著图生成两个步骤。

综合考虑效果和计算效率，本文采用的 Hou 等[33]提出的显著性模型计算方法，该模型从信息论角度出发，将图像信息分为不变部分和易变部分，利用生物体对易变部分更敏感的视觉机制，通过频谱分析得到显著性区域。由图 2-2（b）可知，

文本区域是图像的高显著性区域，非文本区域的显著性相对较低。

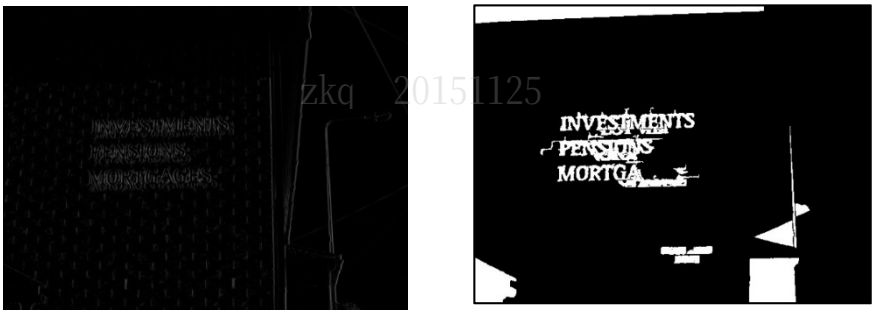


（a） （b）

图 2-2 显著性检测：（a）原始图像；（b）显著性特征图

（2）边缘密度分析

文本区域往往存在丰富的边缘，尤其是英文文本，其垂直边缘特性十分显著，尤其是文本以文本行的形式出现的时候，其边缘密度更加丰富，非文本区域的边缘密度往往较低，因此文本区域和周围的背景区域在边缘密度上差别很大。通过提取边缘密度，经过平滑滤波等预处理，然后用投影分析、窗扫描或者形态学操作等方式，可以快速的进行文本检测，对文本进行粗定位，得到候选文本区域。



（a） （b）

图 2-3 边缘密度和 MSER：（a）垂直边缘密度特征；（b）MSER 区域

（3）MSER 分析

最稳定极值区域即 MSER 是由 Matas 等[34]提出的一种特征检测算子，其基于分水岭算法，最早是用于解决宽基线场的匹配问题，由于 MSER 具有：1）尺度不变性；2）仿射不变性；3）稳定性：所有相同阈值范围内的区域都会被选择等诸多优点，现在广泛应用到图像检索、图像分类等领域。

Davide Nister 等[35]改进了传统的 MSER 提取算法，提出了 Linear Time Maximally Stable Extremal Regions 算法，大大提升了计算速度。其基本原理是对一幅灰度图像采用递增梯度阈值进行二值化处理，在得到的二值化图像序列中，存在这样的一些区域，满足在一定的阈值变换范围内，其连通域的面积变化很小，甚至没有发生变化，我们称这些区域为最大稳定极值区域。其数学定义如下：

$$q(i) = |Q_{i+\Delta} - Q_{i-\Delta}|/|Q_i| \quad (2\text{-}1)$$

其中，Q_i表示灰度阈值取 i 时某一连通域的面积，Δ表示灰度阈值的微小增量，q(i)表示灰度阈值取 i 时区域 Q 的变化率，当q(i)是局部极小值时，对应的区域 Q 是最大稳定极值区域（MSER+），将原图像的灰度值反转，重复以上的操作，获取最小稳定极值区域（MSER-），最后整合 MSER+和 MSER-，得到最终的 MSER 特征提取结果。

由于自然场景下，文本往往具有颜色一致性，文本区域和和背景区域的对比度也相对明显，因此可以使用 MSER 作为文本检测的手段。

由图 2-3（b）可知，MSER 特征提取的结果可能由于图像本身的低对比度、模糊等原因造成文本字符的 MSER 区域黏连的情况，此外也存在着因为文本字符内部的噪声或者颜色不一致而使得一些文本字符没有被提取成MSER区域，如图2-4所示。但整体上 MSER 的结果很好的包含了可能是文本字符的位置，是文本信息的一种比较好的描述，因此本文将 MSER 区域作为第三章基于几何约束笔划宽度变换的字符水平过滤阶段的补充输入。

（a）　　（b）

图 2-4 MSER 分析：（a）原始图像；（b）MSER 区域（彩色标注）

2.1.2 潜在文本块标记

（1）投影分析

通过前面的边缘密度和显著性特征提取，大量的背景区域被滤除，只保留可能是文本区域的显著性水平和边缘密度等特性，而且自然场景图像中的文本往往以行或列的方式分布，因此其显著性特征值和边缘特征值在行或者列的积累会大于背景区域，满足投影分析的条件。投影分析广泛地运用于文本定位，其原理是将整个图像在竖直或者水平方向进行投影，即像素值按列或者按行进行累计，由于候选文本区域在显著性水平上具有高显著性特征，在边缘密度分布上具有高密度特征，因此其在竖直或者水平方向的投影的累加值也往往处于较高水平，因此可

以通过投影分析进行文本的粗定位，从而得到候选文本行。

图 2-5 显著性检测投影分析

图 2-6 垂直边缘检测的投影分析

由图 2-5 和图 2-6 可知，对显著性特征图和垂直边缘密度特征图进投影分析的结果大致可以确定文本区域的位置，但由于图中存在着灯杆和竖直墙壁等竖长状物体，对文本定位的效果有很大影响，在第三章基于几何约束笔划宽度变换的文本定位中需要采取进一步的操作进行剔除。

（2）窗扫描

为了得到近一步的结果，对边缘密度特征图进行窗扫描，具体流程如下：1）用5×5方框滤波进行平滑处理；2）局部二值化；3）纵向和横向的小矩形窗扫描；4）形态学开操作；5）矩形窗扫描。其处理效果如图 2-7 所示：

（a） （b） （c）

图 2-7 窗扫描：（a）局部二值化；（b）边缘连接；（c）窗扫描

（3）文本块合并

文本块合并的目的是通过对文本特征显著性的点或边缘进行统计和几何处理，得到最终的文本区域。根据文本的排列规则，具体的合并算法如下：1）根据投影分析分别计算横向和纵向峰值宽度的均值；2）分别在横向上和纵向上在均值范围内进行文本块搜索，若搜索到满足条件的文本块，则合并在一起形成一个文本块。通过投影分析和文本块合并得到的候选文本区域经常存在非文本区域，如图 2-8 所示，因此还需要进一步的过滤，而传统的基于连通域的方法，因为提供的信息量太少，对非文本区域的过滤能力很差。因此本文将通过投影分析和文本块合并得到的文本区域作为第三章基于几何约束笔划宽度变换的文本定位的文本行水平过滤阶段的输入，利用笔划特征的细分能力对其进行过滤。

图 2-8 文本块合并效果

2.2 文本区域全局特征计算

传统的基于显著性、边缘密度和 MSER 区域的文本定位方法都是在文本检测的基础上进行更进一步的精定位操作，但由于显著性、边缘密度和 MSER 的方法产生的候选文本区域存在着诸多的问题，导致定位精度不高，如在显著性特征图和垂直边缘密度特征图的基础上，利用投影分析和窗扫描确定文本区域的位置都破坏了原始字符，只能产生一个可能的候选区域，要完成后续的文本识别还需要进一步的文本分割等处理步骤，而在自然场景中文本分割是一个很难的挑战。基于 MSER 的方法对参数的控制需要大量的经验，很容易造成候选字符彼此黏连在一起的情况，单独使用效果很差。

本文将在第三章介绍的几何约束笔划宽度变换是一种局部操作子，在计算的时候无法统计出文本区域的全局信息，但其具有在文本定位的同时进行文本初步分割的功能，而且其最终的定位结果经过孔洞修补等操作后，作为前景的文字区域突出，而背景噪声很少，已是高质量的二值图像，可以直接输入到文本识别模块，具有很大的优势。

鉴于以上的考虑，本文提出了一种将文本检测结果转换为文本区域全局性度量的方法，为第三章基于几何约束笔划宽度变换的文本定位算法在字符水平和文本行水平过滤阶段提供全局性的度量特征，弥补传统的笔划宽度变换算法只采用文本区域的局部特征而缺乏对候选字符和文本行全局性特征度量考虑的缺陷。

（1）候选字符或文本行的显著性度量

首先计算候选字符和文本行在显著性特征图中对应区域的显著性特征值的和，然后计算候选字符面积与显著性均值的乘积，最后将两者的比作为候选字符的显著性程度度量，其计算公式如下：

$$\mathrm{sat(c)} = \frac{\sum_{i,j} I_{sat}(i,j)}{a(c)*m(I_{sat})}, (i,j) \in c \,. \quad (2\text{-}2)$$

其中$a(c)$代表候选字符或文本行内所有文本字符的面积，$m(I_{sat})$代表显著性特征图像的均值，$\sum_{i,j} I_{sat}(i,j)$代表候选字符或文本行在显著性特征图上对应区域的显著性特征值的和。

（2）候选字符或文本行的边缘密度度量

首先计算候选字符和文本行在边缘特征图中对应区域的边缘灰度值和，然后计算候选字符面积与边缘图像灰度均值的乘积，最后将两者的比值作为候选字符和文本行的边缘密集度度量，其计算公式如下：

$$eden(c) = \frac{\sum_{i,j} I_{edge}(i,j)}{a(c)*m(I_{edge})}, (i,j) \in c \,. \quad (2\text{-}3)$$

其中$a(c)$代表候选字符或者文本行内所有文本字符的面积，$m(I_{edge})$代表垂直边缘密度特征图均值，$\sum_{i,j} I_{edge}\,(i,j)$代表候选字符或文本行在垂直边缘密度特征图上对应区域的垂直边缘特征值的和。

（3）候选字符或文本行位置和尺寸的全局性度量

此外，本文受文献[25]字符水平特征设计的启发，采用文本区域在整幅图像中出现的位置和尺寸信息的四维向量$[x', y', w', h']$作为全局性的度量，其中x'代表候选字符或文本行最小外接矩形左上角的横坐标和图像宽度的比，y'代表候选字符或文本行外接矩形左上角的纵坐标和图像高度的比，w'代表候选字符或文本行外接矩形的宽度和整幅图像宽度的比，h'代表候选字符或文本行外接矩形的高度和整幅图像高度的比。

2.3 本章小结

本章通过提取图像的显著性特征、边缘密度特征和最稳定极值区域特征，然后利用投影分析、窗扫描和形态学处理等操作，快速高效地得到了文本区域的大

致位置。但本文创新性地将文本检测的结果转换为一个全局性的度量，为基于几何约束笔划宽度变换的文本定位在字符水平特征和文本行水平特征选取阶段提供输入，弥补传统的笔划宽度变换算法的局部敏感缺陷。

第三章 基于几何约束笔划宽度变换的文本定位

文本定位是自然场景文本信息提取的第一个环节，其目的是在自然场景图像中将文字区域的位置标记出来，剔除非文本的草坪、道路、树木、建筑等背景区域，这是因为在进行文本识别的时候，识别引擎只对文本区域进行识别，对非文本区域的识别毫无意义。文本区域定位的越精确，包含的非文本区域越少，对提高后续文本识别的效果越有利。

本章的内容组织如下：首先介绍了传统的笔划宽度变换算法和其处理步骤，分析传统笔划宽度变换算法的缺点后，重点阐述了本文改进的基于几何约束的笔划宽度变换算法，然后详细细阐述了基于几何约束笔划宽度变换进行文本定位的流程，最后在公开的评测的数据集上验证了基于几何约束的笔划宽度变换算法的效果，并分析了失败案例的原因。

基于几何约束笔划宽度变换的文本定位部分主要由图像预处理、特征选择和提取、候选字符生成和过滤、文本区域生成和过滤四个部分组成，蕴含了从底向上的聚类思想和自顶向下的剪枝思想,其框架如图 3-1 所示。

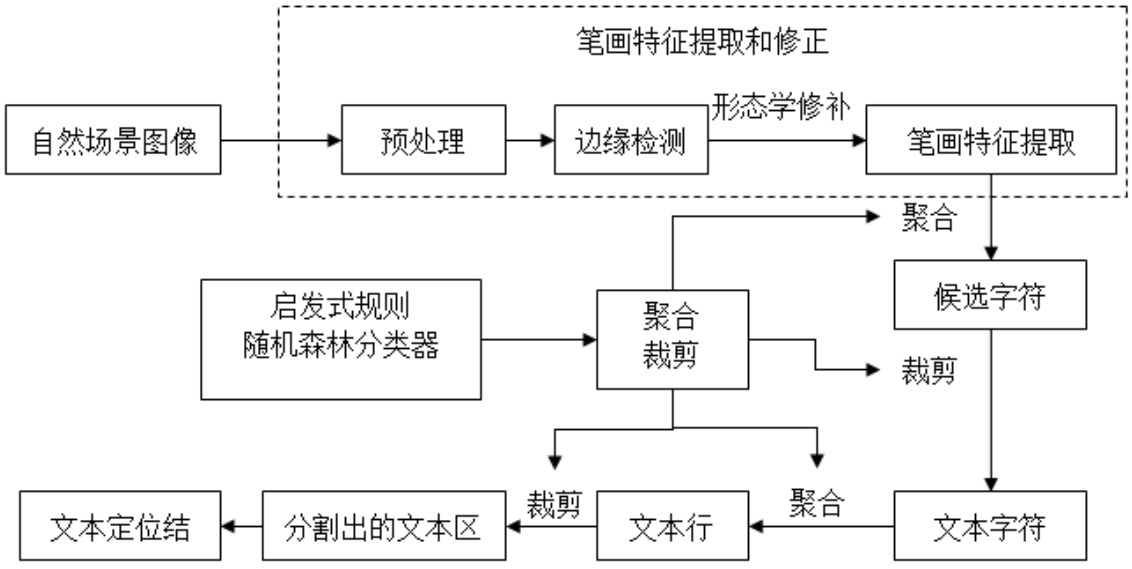


图 3-1 基于笔划特征的文本定位算法流程图

在图像预处理阶段，主要包含图像的二值化、直方图均衡化、边缘保持滤波、边缘检测和梯度计算等，目的为了得到具有丰富文本边缘，同时具有较少背景边缘的边缘图像，方便特征提取阶段笔划宽度特征的准确计算。

特征选择和提取阶段，主要包含基于几何约束的笔划宽度的计算和修正、字

符水平和文本行水平特征设计和计算。

候选字符生成和过滤阶段是利用区域生长算法得到候选字符，然后分别利用启发式规则和基于机器学习的分类器进行双阶段过滤，剔除非文本字符，保留文本字符。

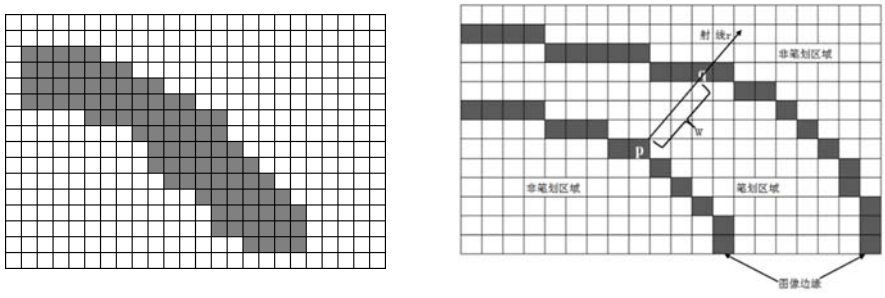
文本行生成和过滤阶段，则是把文本字符的布局特性聚成文本行，然后利用启发式规则和基于机器学习的分类器进行过滤，剔除非文本行，保留文本行。这样做有两个方面的目的：一是在文本行水平上过滤掉那些在字符水平上无法剔除的非文本字符，二是确定文本区域的位置，为文本图像的二次分割提供输入。

下面将分别对基于几何约束笔划宽度变换的文本定位的各个阶段进行详细的介绍。

3.1 笔划宽度变换算法

笔划宽度变换算法由 Epshtein 等[23]在 2010 年提出，是一种图像局部描述子，如图 3-2 所示，字符笔划宽度的定义为垂直字符边缘的线段长度。由于笔划宽度特征很好地描述了文本的本质特性，即文本字符具有相对稳定的笔划宽度，而非文本字符的笔划宽度往往变化很大，因此在文本字符和非文本字符之间具有很高的区分度，要进行笔划宽度变换首先要进行边缘检测，然后求取边缘点的梯度方向，按梯度方向进行查找寻找出另一个满足一定条件的边缘点，则形成点对，然后计算边缘点对间的线段长度，将其线段长度作为该线段上所有像素点的笔划宽度值。

下面先介绍传统的笔划宽度变换算法的处理流程，然后介绍本文在此基础上改进的基于几何约束的笔划宽度变换算法。



（a） （b）

图 3-2 笔划宽度变换：（a）笔划；（b）笔划宽度的定义

几何约束笔划宽度变换是本文三层过滤思想的第一层，即在像素水平上区分文本像素和非文本像素。经过几何约束笔划宽度变换后，文本像素区域被赋予正的笔划宽度值，其值代表该像素点最可能属于的笔划段的宽度值，而非文本像素区

域的笔划宽度值为-1。

3.1.1 传统笔划宽度变换算法

传统的笔划宽度变换算法主要由三步组成：（1）边缘点对查找；（2）笔划宽度赋值；（3）拐角处像素点笔划宽度修正。

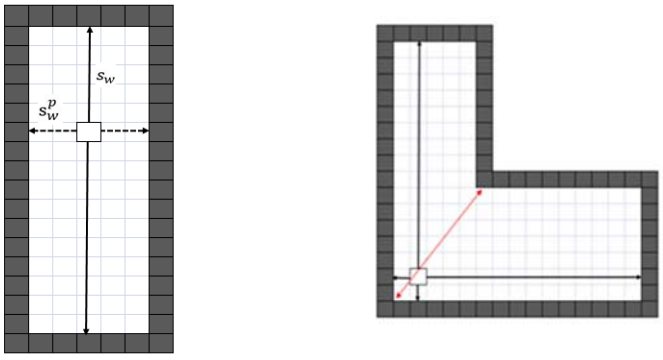
（1）边缘点对查找

在经过图像预处理后得到的边缘图像上，如图 3-2（b）所示，假设 p 是图像边缘上的一个像素点，d_p是利用 Sobel 算子得到的该点的梯度方向，从 p 点开始沿着d_p方向作射线$r = p + n \cdot d_p (n > 0)$进行生长直到第一次寻找到另一个边缘像素点 q，则终止查找，记终止点 q 的梯度方向为d_q，若d_p和d_q方向大致相反，即满足$d_q \leq -d_p \pm \frac{\pi}{6}$，则此次射线寻找有效；否则，该次射线寻找无效，重新选择新的边缘像素点p′继续以上的查找过程。

（2）笔划宽度赋值

在射线查找有效的情况下，计算查找起始点 p 和终止点 q 之间的线段长度$s_w = \|\overrightarrow{p-q}\|$，遍历射线上 p 和 q 之间的所有点，若该点没有被赋予过笔划宽度值，则赋予笔划宽度值s_w，若该点已经被赋予过笔划宽度值s_w^p，则比较其当前笔划宽度值s_w^p和s_w的大小，如图 3-3（a）所示，取其较小者作为该点的当前笔划宽度值。

重复以上操作，直到所有有效射线上的像素点都被赋予了代表该点笔画宽度的值，没有笔划宽度值的被赋值为-1，如图 3-4 所示。



（a） （b）

图 3-3 笔划特征计算:(a)笔划宽度赋值规则；（b）拐角笔划宽度修正

（3）拐角处像素点笔划宽度修正

经过第二步的笔划宽度赋值规则，在图 3-3（b）中得到的笔划宽度是经过该点的水平笔划线或垂直笔划线的宽度较小者，但其笔划宽度值并不能很好地表征其笔划宽度特征属性，其真实的笔划宽度值是图 3-3（b）中红色斜线段的长度所

表示的。因此需要对其笔划宽度进行修正，修正方法如下：

重新遍历所有有效的笔划射线，计算每一条有效笔划射线上所有像素笔划宽度的中值，重新设置该射线上笔划宽度值超过笔划宽度中值的像素点的笔划宽度特征值为笔划宽度中值。

经过以上的笔划宽度变换后，图像上的每一个像素点都被赋予了代表其笔划宽度属性的值，如图 3-4 所示。

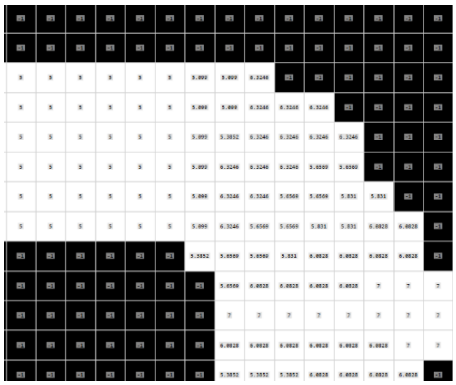
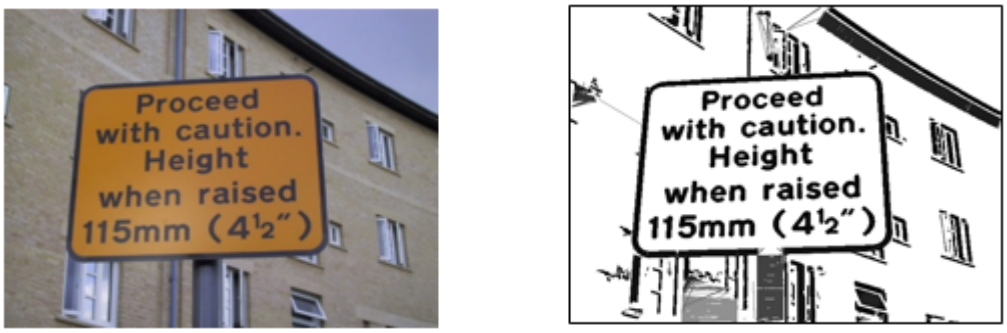


图 3-4 笔划宽度特征值

需要指出的是，该笔画特征计算的过程是针对暗文本在亮背景上的情况，对亮文本在暗背景上的情况，则在进行射线查找的时候，需要修改搜索方向为$-d_p$，其它操作不变。由图 3-5（b）可知，笔画宽度变换大幅度地剔除了非文本像素，使得文本区域更为突出，很好的完成了三层过滤思想中像素水平过滤的功能。然而，通过观察发现，笔划宽度变换的结果也存在着候选字符的生成质量不高，大量的非文本像素被误判为文本像素，由于边缘缺失导致边缘点梯度方向对称性破坏而出现的字符黏连等情况，这将会增加字符水平和文本行水平过滤的难度，尤其是候选字符黏连的情况，会严重影响文本定位的效果，本文针对以上的情况，利用几何约束进行改进。



（a）　　（b）

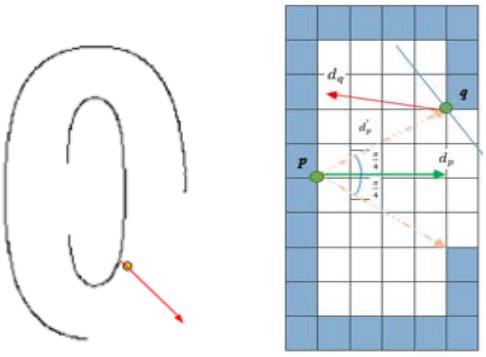
图 3-5 笔画特征变换效果：（a）原图像；（b）传统笔划宽度变换效果

3.1.2 基于几何约束的笔划宽度变换

传统的笔划特征是局部操作子，对噪声、图像模糊、低对比度等问题造成的边缘失准和边缘缺失情况容错能力差，易造成字符缺失或者黏连，进而导致后续的候选字符生成、过滤和文本行聚合阶段失败。此外，由于在笔划线的查找过程中仅仅利用了边缘像素点梯度方向的对称性，而在一些自然场景图像中很多非文本区域满足这个条件，此时传统的笔划宽度变换导致大量非文本字符的形成，使得候选文本像素在文本字符的分布并不是很集中，形成的候选字符质量不高。在分析了传统的笔划宽度算法笔划线查找的缺点后，本文对传统的笔划宽度变换算法提出了以下的改进。

（1）笔划射线生长查找约束

针对传统笔划宽度变换算法在边缘缺失造成的边缘点梯度方向对称性破坏情况，本文采用在沿着垂直笔划方向的梯度方向进行射线寻找的时候加入颜色距离的限制性约束进行改进。其具体思想是在笔划射线生长过程中，当笔划射线的长度超过了上一条有效笔划长度的 1.5 倍仍未找到满足梯度方向对称性的边缘点，则开始判断当前生长点与查找起始点 p 的颜色距离，若当前点和查找起始点的颜色距离小于 80，则继续查找，当射线生长长度超过了上一条有效笔划长度的 3 倍时候，仍然未找到对应的边缘点，则此条射线查找无效，重新选择新的起始点进行查找。



（a） （b）

图 3-6 条件约束：（a）射线生长约束；（b）射线方向约束

由于在射线生长查找约束的过程中使用到了颜色距离，而传统的 RGB 颜色空间受光照不均条件的影响很严重，但在自然场景图像中却大量存在着光照不均匀的问题，同样的颜色在不同的光照条件下差异很大，采用传统的 RGB 颜色空间效果很差。因此本文采用的是 HSV 颜色空间，HSV 颜色空间中颜色的色调和亮度彼此分离，更接近于人眼对颜色的感知，降低了光照变化的影响。

（2）笔划射线方向查找约束

针对局部光照不均、部分轻微遮挡造成的边缘缺失情况，本文通过对笔划射线的查找方向进行约束来达到避免字符黏连的目的。如图 3-6（b）所示，其具体思想是在笔划射线生长过程中，当当前笔划射线寻找的长度超过了上一个有效笔划长度的 2 倍，仍未找到满足条件的边缘点情况下，则重新从当前笔划射线的起点开始沿着与起点的梯度方向呈正负 45 度偏差的方向，即$d_p' = d_p \pm \frac{\pi}{4}$进行二次搜索，此时边缘点梯度方向对称性需要满足的条件放宽为$d_q = -d_p \pm \frac{\pi}{2}$，若在上一个有效笔划射线长度 3 倍范围内仍未找到满足梯度方向对称性的边缘点对，则重新寻找新的起始点进行笔划射线查找。

（3）笔划宽度特征异常值处理

在笔划宽度变换的时候，因为射线查找是按点阵像素查找的，再加上字符内部的噪声等原因，有效笔划射线并不能覆盖所有的文本像素区域，这样有时会造成字符内部出现孔洞的情况，使得孔洞所在处的像素笔划宽度值失真。此外由于受噪声等影响，候选字符的边缘也会出现缝隙，当孔洞和缝隙的数量较多的时候，会影响对这个候选字符的判决，导致文本字符被误判为非文本字符。因此需要对孔洞和缝隙进行修正，此外孤立的笔划点和候选字符之间的粘连也需要进行修正，不然会对字符水平非文本字符的过滤产生严重的影响。

首先，孔洞和缝隙修补。

形态学处理是指用数学形态学的方法从图像中提取边界、角点、轮廓和连通域等更具代表性的特征，其典型应用包括形态学细化、形状检测、滤波和修剪毛刺等。基本操作是膨胀和腐蚀，及基于基本操作的开运算和闭运算。灰度图像的闭运算是先膨胀后腐蚀，可以有效地填充笔划宽度特征图像内部的细小孔洞，弥合外围细小的缝隙。

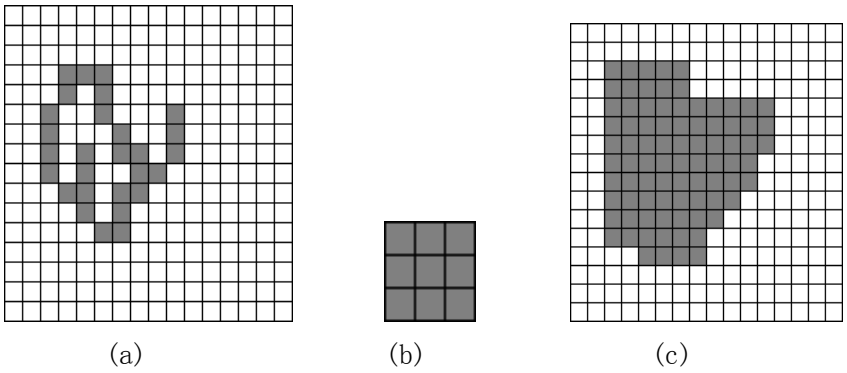


图 3-7 形态学闭运算效果图：（a）原图像；（b）结构元素；（c）闭运算效果

此外，剔除孤立笔划点和短直线。字符的笔画像素往往不单独出现，而小的噪声区域也可能形参类字符，通过对孤立笔画点的去除，可以很好地改善笔画特征。这里我们设置孤立笔画点的去除规则如下：统计有效笔画点周围5×5范围内有效笔画点个数 n，若 n 小于 10，则置该范围内所有像素点为无效笔画像素点（笔画宽度值为-1）。

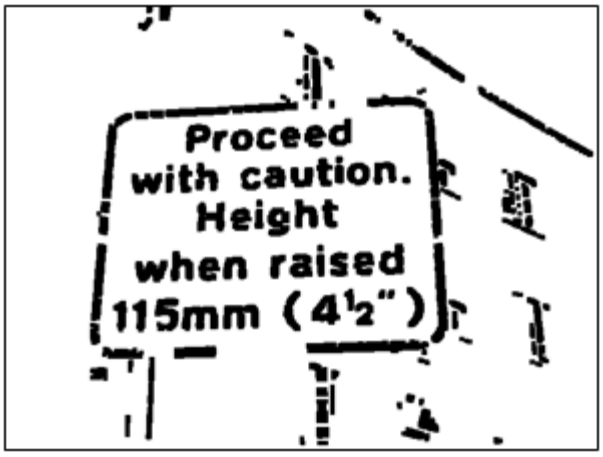


图 3-8 基于几何约束的笔划宽度变换效果

3.2 基于区域生长的候选字符生成和启发式规则的验证算法

3.2.1 基于区域生长算法的候选字符生成

传统的连通域生成算法主要是考虑的是连通规则的制定，具体说就是传统二值图像的连通域分析需要具备在灰度值上的相似性和空间上的连续性：1)空间上的连通性：4 邻域和 8 邻域和在此基础上衍生出的 4 连接、8 连接和 m 连接；2)灰度值的相同性。

区域生长算法是经典的图像分割算法之一，其基本思想是基于像素点或子区域之间在颜色、灰度、纹理等特征上的相似性，从选择的生长点开始，按照预先指定的连通性规则不断将和生长点相似的像素或区域合并到生长点所在的区域，然后更新生长点，不断重复这个生长过程，直到满足预先相似性定义的像素点或区域不存在为止，此时停止生长。

这里我们采用类似的思想，但对生长规则进行改进，另外选择笔划宽度变换算法形成的连通域的轮廓作为生长点，基于笔划宽度和颜色的区域生长的候选字符生成算法如下：

笔划宽度特征图上用一个 4 维向量$\{s_w, I_r, I_g, I_b\}$表征一个像素点，s_w表示该像素点的归一化的笔划宽度值，I_r、I_g、I_b分别表征该像素点的归一化的 R、G、B 通道的

颜色值，利用区域生长算法得到候选文本区域，区域生长的相似性评判准则采用以下公式:

$$S_{pq}=\sqrt{0.6|s_{wp}-s_{wq}|+0.3|I_{rp}-I_{rq}|+0.2|I_{gp}-I_{gq}|+0.1|I_{bp}-I_{bq}|}\ . \qquad (3\text{-}1)$$

区域生长的终止条件是$S_{pq}\geq 0.5$。

3.2.3 连通区域分析

连通域分析分析的目的是为了区分字符连通域和非字符连通域，剔除非字符连通域。在这部分利用的是三层过滤思想的第二层的第一阶段，即在字符水平进行基于启发式规则的过滤，第二层的第二阶段，即基于机器学习方法的过滤将在3.4.2节中进行介绍。启发式规则能快速高效地剔除明显的非文本字符，计算简单但缺乏泛化能力，而基于机器学习方法的过滤具有很好的推广性，但计算复杂。因此双阶段过滤很好地结合了启发式规则的快速高效和基于机器学习方法精度高的优点。实验表明，双阶段过滤思想在候选文本字符的过滤上取得了很好的效果，本部分先介绍基于启发式规则的连通区域分析。

基于启发式规则的连通域分析是字符水平双阶段过滤思想的第一阶段，目的是利用简易的统计特性和几何规则快速剔除那些明显不是文本字符的连通域。

启发式规则主要包含候选字符的统计特性和几何特性，这些特性易于计算，对类字符具有很强的过滤作用，可以在很大程度上过滤掉非文本字符。对候选字符 c，我们主要计算字符的以下基本特征：1)最小外接矩形$br(c)$；2)笔划宽度均值$\mu(c)$；3）笔划宽度方差$\sigma(c)$；4)笔划宽度中值$m(c)$。

在此基础上我们选取以下特征作为字符的过滤条件：

（1）候选字符笔划宽度方差[23]

文本字符由于其笔划宽度具有一致连续性，其笔划宽度方差和均值的比较小，而非文本字符一般较大，本文采用的候选字符的笔划宽度方差定义为：

$$\delta(c)=\frac{\sigma(c)}{\mu(c)}\ . \qquad (3\text{-}2)$$

其有效范围为[0,0.6]。利用笔划宽度方差的启发式规则可以去除一些类似文本的字符比如草地、窗户和森林等，因为这些区域一般都具有较大的笔画宽度方差，而文本字符因为笔划宽度具有一致性，其字符笔划宽度方差一般较小。

（2）候选字符宽高比

文本字符为了便于识别，其外形一般都在合适的范围内，本文采用宽高比作为字符形状的一种描述，其定义为：

$$apect(c) = min\left\{\frac{w(c)}{h(c)}, \frac{h(c)}{w(c)}\right\}. \quad (3\text{-}3)$$

其有效范围为[0.1,1],其中$w(c)$和$h(c)$分别为候选字符 c 的最小外接矩形的宽度和高度。通过这个约束可以过滤掉自然场景图像中的细长物体，如木棍、电线杆等形成的狭长连通域。

（3）候选字符像素占空比[25]，其定义为：

$$o(c) = \frac{q(c)}{area(c)}. \quad (3\text{-}4)$$

有效范围为[0.1,1]，$q(c)$为字符 c 字符像素的个数，$area(c)$为字符 c 最小外接矩形$br(c)$的面积，即$area(c) = w(c) \times h(c)$，其中$w(c)$、$h(c)$分别为候选字符 c 的最小外接矩形的宽度和高度。这是因为在连通过程中可能会形成一些随机的的孤立直线，而在连通区域形成的时候可能把这些区域合并成连通域。如果面积比太大，或者太小，则剔除。

（4）候选字符所有像素笔划宽度最大值与笔划宽度中值$m(c)$的比，其有效范围为[0.1,5]。

（5）候选字符面积限制

为了便于识别，文本字符一般都具有合适的大小，面积小于 25 像素和大于图像面积 1/3 的候选连通区域被当成噪声剔除。

3.3 基于聚链的候选字符合并和启发式规则的验证算法

3.3.1 基于聚链的候选字符合并

一些候选字符在字符水平无法区分是文本字符还是非文本字符，比如因为随机出现的噪声和混乱背景形成的类文字字符，需要在更高的文本行水平上进行过滤，因此需要将单个的文本字符合并成文本行。

（1）首先将文本字符合并字符对，其合并规则如下：

- 相似的字符尺寸：考虑到字符的大小写混合情况，这里要求组成字符对的两个字符的外接矩形框的高度比例小于 2。
- 相似的字符笔划宽度：要求组成文本对的两个字符的平均像素笔划宽度值的比例小于 2。
- 相似的颜色：要求组成文本对的两个字符在 HSV 颜色值的上下误差为 80。
- 空间上相邻:字符中心的距离不超过候两个选字符尺寸较大高度值的 3 倍。

（2）文本对形成文本行

基于以下的准则进行合并文本对：

- 至少共享一个终端

共享一个终端指文本对 A 和文本对 B 至少有一个字符是重合的，包含以下四种情况:

a) 文本对 A 的起点是文本对 B 的起点。

b) 文本对 A 的终点是文本对 B 的起点。

c) 文本对 A 的起点是文本对 B 的终点。

d) 文本对 A 的终点是文本对 B 的终点。

- 具有相似的方向和尺寸，这里我们采用 Yao 等[25]提出的方法。

文本对的方向相似性计算方式如下：

$$S_o(C_1,C_2)=\begin{cases}1-\frac{\gamma(C_1,C_2)}{\pi/2} & if\ \gamma(C_1,C_2)\le\pi/6\\ 0 & otherwise\end{cases}. \quad (3\text{-}5)$$

其中$\gamma(C_1,C_2)$为文本对C_1,C_2的方向夹角。

文本对的尺寸相似性定义为：

$$S_s(C_1,C_2)=\begin{cases}1-\frac{|n_{c1}-n_{c2}|}{|n_{c1}+n_{c2}|} & if\ \gamma(C_1,C_2)\le\pi/6\\ 0 & otherwise\end{cases}. \quad (3\text{-}6)$$

文本对的相似性定义为：

$$S(C_1,C_2)=\omega S_o(C_1,C_2)+(1-w)S_s,\ \omega\in[0,1]. \quad (3\text{-}7)$$

利用以上的合并规则，合并文本对成文本行，直到没有文本对可以合并。

3.3.2 文本行验证

文本行验证是三层过滤思想的第三层，包括基于启发式规则的文本行验证和基于机器学习的文本行验证，本部分介绍基于基于启发式规则的文本行验证，基于机器学习方法的文本行验证将在 3.4.3 节中阐述。

为了意义的完整表达，自然场景中的文本字符通常不单独出现，而是以单词或句子的形式，而且文本行在外形尺寸上也满足一定的要求，因此本文采用以下的文本行验证方法：

（1） 候选文本行包含的文本字符个数≥ 3。

（2） 候选文本行的外接矩形的长宽比≤ 15。

为了增强对文本行内部偶然出现的噪声的鲁棒性，若候选文本行满足以下条件中的一个，则该候选文本行仍保留。

（1） 候选文本行内字符面积与非字符区域面积的比例大于 3。

（2） 候选文本行内字符的置信均值大于 0.3。

3.4 基于随机森林的候选字符和文本行分类

3.4.1 随机森林原理介绍

随机森林[39]是一种统计学习方法，由 Leo Breiman 和 Adele Cutler 在 2001 年提出，它具有训练速度快，预测阶段耗时少，能够不做特征选择的处理高纬度数据，能给出特征的重要性程度，对噪声和坏点不敏感和不容易过拟合等优点，在经济学、医学等领域取得了广泛的应用。

随机森林的基本思想是随机生成多棵决策树来建立一个森林，决策树之间是没有关联的，然后从原始样本中抽取多个样本分别训练每个决策树分类器，分类阶段让森林中的每棵树都对其输入的训练样本进行完全分裂，直到叶子节点不能再分裂，或者叶子节点中的所有样本都属于同一类，最后组合多棵决策树的分类结果，预测样本的最终分类结果。虽然随机森林中的每一颗决策树的分类能力很弱，但通过 bagging 的方法将每颗决策树组合在一起就成了强分类器。

随机森林的关键点是决策树和信息增益，前者是随机森林的基本组成单元，是对空间的一种划分，其划分的子空间互不相交，后者是决策树选择属性进行分枝的依据。

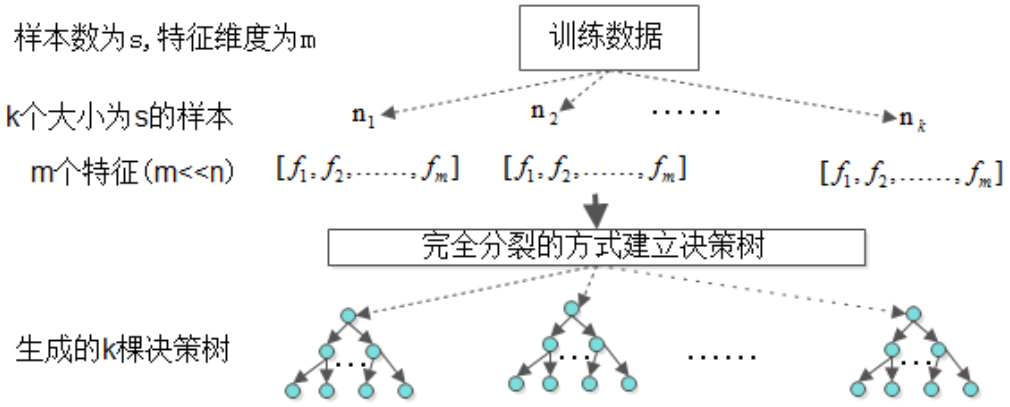


图 3-9 随机森林的生成过程

随机森林的训练样本满足以下两个条件：

（1）样本的选取：采用有放回的方式，随机样 N 个样本（N 为输入样本的个数），共采用 K 次，得到 K 个随机样本簇，其中每个样本簇的样本个数为 N.这样保证每棵树的训练样本都不是完整的样本。

（2）特征的选取：从 M 维样本特征向量中，随机选取 m（n<<M）个组成决策树训练样本的特征向量$[f_1, f_2, \ldots\ldots, f_m]$。

以上样本随机采用和特征随机采样保证了随机性，是的随机森林算法不像其它基于决策树的算法，即使不剪枝也不会出现过拟合现象。

信息增益是信息熵的差$\Delta H = H' - H$,其中 H 代表信息熵，定义如下：

$$H = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \ . \qquad (3\text{-}8)$$

其中 n 代表类别数，$p(x_i)$代表第 i 类样本的概率。

选择属性f_i进行分枝，如果此时分枝规则是$x_i \geq f_i$，则，满足此条件的被分到一个分支，不满足的被分到另一个分支，则这两个分支中至少包括两个类别，若继续分枝则可能出现更多的类别，此时计算这两个分支的信息熵H_1、H_2，然后计算此时的总信息熵$H' = p(x_1)H_1 + p(x_2)H_2$。

3.4.2 字符水平特征设计

基于几何约束的笔划宽度变换算法有效地剔除了非文本像素，保留了候选字符的文本像素，区域生长算法将文本像素聚成了候选字符。然而为了达到文本定位的目的，即确定文本区域的位置，虽然 3.2.1 和 3.2.2 节中利用启发式规则过滤掉了一部分的非文本字符和非文本行，然而基于启发式的规则并不精确，缺乏泛化能力，只能作为粗略的筛选。基于统计特性和几何特性的过滤是硬性过滤，不能剔除那些类文本连通域，往往有一些满足这些条件的候选字符依然不是文本字符，因此需要采用更更加智能和精确的过滤，这部分我们采用基于机器学习的连通域过滤，首先提取能够很好地区分文本连通域和非文本连通域的特征，然后通过随机森林的方法进行学习，用得到的分类器判别每一个连通域是文本字符还是非文本字符。

基于旋转不变性、尺度不变性和低计算量的考虑，以下基本特征被用来计算字符水平的特征:1)候选字符的质心；2)候选字符尺寸。

基于以上基本特征，本文设计并计算了以下字符水平的特征：

（1）轮廓描述器

文字区域通常具有丰富的边缘特性，这个特性除可以用边缘密度描述其空间分布情况外，也可以通过计算边缘像素点的梯度强度和方向进行描述。方向梯度直方图 Hog[40]（Histogram of Oriented Gradient）已在行人检测、字符和人脸识别上取得了较好的效果，本文利用 Hog 描述候选字符的轮廓特征，边缘像素点的梯度强度和方向采用 Sobel 算子得到，然后将每个字符的轮廓图像划分成 9 个块，在每个块内计算 8 个方向上的分布情况，如图 3-10 所示,最后进行归一化来获得字符的轮廓描述器。因为字符具有一定的笔划宽度，其边缘常常以对称点对的方式出现，而且其笔划宽度具有相对一致性，所以其 Hog 响应和非字符具有区分性。

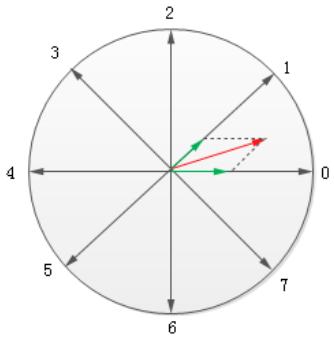


图 3-10 梯度方向直方图和分解示意图

（2）笔划宽度方差$\sigma(c)$

自然场景中的类文本区域，如草地、树叶、窗户等，这些区域通过前面的几何规则往往难以剔除，而采用笔划宽度方差却可以很好地区分，这是因为文本字符一般都具有相对恒定的笔划宽度，因此笔划宽度方差相对较小，而草地、窗户等区域的笔划宽度方差通常较大，其定义为：

$$\sigma(c) = \frac{\sum_{J(i,j)}(I(i,j)-u(c))^2}{q(c)} \quad . \tag{3-9}$$

其中$u(c)$为候选连通区域笔划宽度均值，$q(c)$为候选字符像素总数。

（3）字符尺寸纵横比$aspect(c)$

为了便于识别，文本字符的尺寸比例一般都在一定的范围内，而非文本字符，如细长的灯杆等其尺寸比例往往悬殊很大，因此可以提取外接矩形的宽高比的最小值，其定义为：

$$aspect(c) = min\left\{\frac{w(c)}{h(c)}, \frac{h(c)}{w(c)}\right\}. \tag{3-10}$$

其中$w(c)$和$h(c)$为候选文本字符 c 的外接矩形的宽度和高度。

（4）字符像素密度$d(c)$，其定义为：

$$d(c) = \frac{q(c)}{\pi \cdot S(c)^2} \quad . \tag{3-11}$$

其中$q(c)$为候选连通域内的像素个数，$S(c)$为候选连通区域的尺寸$S(c) = L(c) + l(c)$, $L(c)$和$l(c)$分别为候选连通域外接矩形的长边和窄边。因为文本字符一般都具有一定的内部结构，因此其字符像素密度比值会在一定区间内，超过这个区间的可以判为非文本字符。

笔划宽度特征的计算严重依赖于边缘检测的结果，在一些情况下，如图像模糊或者过度曝光等造成边缘缺失严重甚至检测不到边缘的时候，就不能准确地提取到像素的笔划宽度特征，为了进一步增强算法对噪声的鲁棒性，本文补充最稳定极值区域响应和笔划滤波响应两个字符水平的特征。

（5）最稳定极值区域（MSER）响应，其定义为：

$$mr(c)=\frac{q(c)-ma(c)}{q(c)}. \tag{3-12}$$

其中$q(c)$为候选字符前景像素区域面积，$ma(c)$为候选字符前景像素对应区域是MSER区域的区域面积。视觉上字符区域往往具有一致的颜色，因此可以用MSER极值响应作为其颜色一致性的度量。

（6）笔划滤波响应

笔划滤波响应[41]的原理是基于字符是由一些线段、笔划等基本结构组成，这些结构具有宽度一致性，而且方向相对固定，主要集中在0、pi/4、pi/2、3pi/4四个方向上,而非文本字符往往不具有这些特性。本文采用如图3-11所示的滤波器，该滤波器对条状物体具有很强的响应，而对非条状物体进行抑制，因此使得字符区域的响应较强，利于后面的进一步处理。

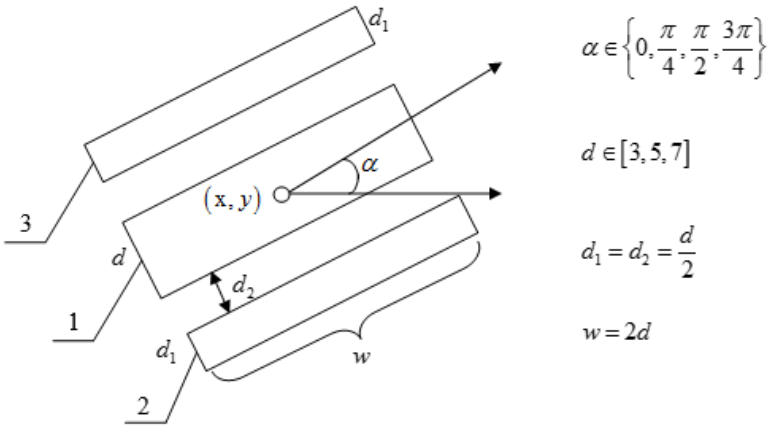


图3-11 笔划滤波器

笔划滤波器在像素点$p(x,y)$的滤波响应定义为：

$$R^B(x,y)=\max_{(a,d)} R^B_{a,d}(x,y)\ , \tag{3-13}$$

$$R^B_{a,d}(x,y)=\frac{2u_1-u_2-u_3-|u_2-u_3|}{\sigma}\ . \tag{3-14}$$

其中$R^B(x,y)$为像素点$p(x,y)$的笔划滤波响应值，$R^B_{a,d}(x,y)$为在方向为α,笔划滤波器中间瓣宽度值为 d时笔划滤波器的响应，α为笔划滤波器与x轴正向的夹角值，d为笔划滤波器中间瓣的宽度值，x为像素点$p(x,y)$的横坐标值，y为像素点$p(x,y)$的纵坐标值，B代表该笔划滤波器求的是灰度图像的亮度值，u_1,u_2,u_3分别表示笔划滤波器中间瓣（图3-11中1区域），左瓣（图3-11中2区域）和右瓣的面积（图3-11中3区域），σ表示笔划滤波器中间瓣的方差。不同于文献[41]中的是，本文不采用多尺寸d扫描的方式，而是将d设置为字符的笔划宽度均值，而笔划

宽度均值的设置很好地表征了文本字符的和非文本字符在笔划滤波器响应上的区别，克服了原始方法对滤波器中瓣宽度值 d 的设置的敏感性。

（7）候选字符边缘平滑度

候选字符边缘平滑度定义是候选字符边缘轮廓的长度和其凸包的长度之比。文本字符的轮廓边缘变化平缓，相对光滑，而非文本字符的轮廓往往变化剧烈且不规则，如树叶形成的锯齿状的轮廓。因此采用边缘平滑度特征可以有效的区分轮廓粗糙的的非文本字符。

$$sm(c) = \frac{cl(c)}{convl(c)}. \tag{3-15}$$

其中$cl(c)$代表候选字符 c 的外轮廓的长度，$convl(c)$代表候选字符 c 的外轮廓的凸包的长度。

（8）候选字符内部连通性

候选字符内部连通性定义是候选字符的膨胀图像和原图像的差，这是因为文本字符的内部空隙有一定的规则，膨胀操作之后不会出现大面积的黏连，而非文本字符内部的空隙可能杂乱无序，当进行膨胀操作之后容易黏连在一起，因此内部的连通性往往较差。

$$con(c) = \frac{n(d(c)-c)}{n(c)}. \tag{3-16}$$

其中$n(c)$代表候选字符 c 的像素个数，$n(d(c)-c)$代表候选字符 c 的膨胀图像和原始图像差的像素个数。

（9）候选字符紧凑度

候选字符紧凑度定义是候选字符连通域像素点数目与连通域轮廓周长的平方之比，这是因为非文本连通域的轮廓复往往复杂、周长较长，因此其紧凑度往往很小，而文本连通域的的字符笔划弯曲相对较少，因此其紧凑度会在合适的范围内。

$$t(c) = \frac{n(c)}{cl(c)*cl(c)}. \tag{3-17}$$

其中$n(c)$代表候选字符 c 的像素个数，$cl(c)$代表候选字符 c 的轮廓周长。

（10）颜色一致性

利用第二章基于 MSER 特征的文本检测算法得到的初步候选文本区域，计算候选字符和文本区域的重叠比，作为候选字符的颜色一致性度量，其计算公式如下：

$$ccon(c) = \frac{n(c\cap ma(c))}{n(c)}. \tag{3-18}$$

其中$n(c)$代表候选字符 c 的像素个数，$ma(c)$代表自然场景图像形成的 MSER 区域，

$n(c \cap ma(c))$代表候选字符 c 重叠的 MSER 区域的像素个数。

3.4.3 文本行水平特征设计

对于某些类似文字的字符，如和字符“I”相似的候选字符，很难在字符水平上进行区分，但根据自然场景图像文本的排列规律，可以通过分析候选字符周围的字符来进一步判断，即在文本行水平上进行分类，本文设计和计算文本行水平以下特征：

（1）文本行包含的文本字符个数 n。

（2）文本行的文本置信概率

在经过字符水平阶段的过滤后，每个字符都有其是文本字符的置信概率，文本行的文本置信概率定义为文本行内所有文本字符的字符置信概率值的平均值

$$p(l) = \frac{\sum_{i=1}^{n} p(c_i)}{n}. \tag{3-19}$$

其中$p(c_i)$代表字符c_i的字符置信概率，n 代表文本行 l 内的字符个数。

（3）文本行字符间距的一致性

文本行内的字符之间间距具有一致性，因此可以用这个特征作为候选文本行是否是文本行的度量，其定义为文本行内所有文本字符质心间距离的方差。

$$dc(l) = \sigma\big(dv(c_i)\big), i \in [1, n]. \tag{3-20}$$

其中$\mathrm{dv}(c_i)$代表字符文本行内相邻字符之间的距离构成的向量。

（4）文本行字符方向一致性

文本行内的字符常具有相对一致性的方向，因此可以用方向一致性进行区分，其计算方式为取文本行内的第一个字符和第二个字符质心的连线作为文本行的方向，然后依次计算相邻候选字符质心连线与文本行方向的余弦值，最后计算每个字符的方向方差作为文本行方向一致性的度量。

$$sc(l) = \delta(acos(v_i, v_1)), i \in (2, n). \tag{3-21}$$

其中v_i代表第 i-1 个字符的质心m_i和第 i 字符的质心m_{i-1}构成的向量，v_1代表文本行内第一个字符的质心和第二个字符质心构成的向量，$acos(v_i, v_1)$代表向量v_i和v_1的夹角。

（5）文本行字符尺寸一致性

文本行字符尺寸一致性定义为文本行内所有文本字符宽度和高度的方差。

（6）文本行字符笔划宽度一致性，其定义为文本行内所有文本字符笔划宽度均值的方差。

$$swc(l) = \sigma\big(swm(c_i)\big), i \in [1, n]. \tag{3-22}$$

其中$swm(c_i)$ 代表字符c_i的笔划宽度均值。

3.5 实验结果和算法性能分析

3.5.1 文本定位算法评价

（1）文本定位算法的评价存在以下难点：

标定难度大：标定是将文本区域用外接矩形包含。问题是不同的研究者标定的结果不同，一些研究者把一个文本区域分成几个，另一些研究者可能把多个文本区域合并成一个，因此标定的差异很大。

样本库不统一：缺乏自然场景图像文本定位的通用样本库，不同的算法在不同的样本库上进行评测，由于样本数量和复杂程度等因素造成评测结果各执其词缺乏统一的说服力。

定位的精度不一样：根据实际需求，一些要求定位到单个的字符或者单词，另一些只要求定位到文本行即可，对于前者需要在定位出的文本行的基础上进行进一步的细分割，不同算法的分割结果也差异很大，这也给文本定位的评价带来了难题。

（2）评价标准

自然场景文本定位属于目标检测的一种，因此广泛使用的评价目标检测效果的准确率、召回率和 F 测度依然实用。准确率的定义是正确检测的结果数目与所有检测结果数目的比，召回率的定义是正确检测的结果数目与真实的所有的结果数目的比，F 测度的定义是准确率和召回率的调和平均数。但结合自然场景文本检测的具体情况和本文采用的测试数据集，评价标准如下：

水平方向文本数据集的评价方法如下：

首先,定义检测到的文本区域质量定义为：

$$m(r;R) = max\{m(r,r')|r' \in R\}. \tag{3-23}$$

其中r代表检测到的文本区域的外接矩，r'代表真实的文本区域的外接矩形，R代表所有真实的文本区域的外接矩形，$m(r,r')$是检测到文本区域和真实的文本区域的交集和它们并集的比。

准确率的定义：

$$p = \frac{\sum_{r_e \in E} m(r_e;T)}{|E|}, \tag{3-24}$$

其中E代表检测到的文本区域集合，T 代表真实的文本区域集合，这里都以外接矩形的方式表示文本区域。

召回率的定义：

$$r = \frac{\sum_{r_t \in T} m(r_t;E)}{|T|}, \tag{3-25}$$

其中 E 和 T 的定义与准确率的定义中的相同。

F 测度的定义：

$$F = \frac{1}{\frac{\alpha}{p}+\frac{1-\alpha}{r}} . \tag{3-26}$$

非水平方向文本数据集的评价方法如下：

对于非水平方向文本使用轴对齐的外接矩形描述文本区域的位置往往不够准确，而其最小外接矩形能更好地体现文本区域的位置，Yao[44]等人提出的评价方法被广泛采用，其具体实现如下：

先求文本区域的外包络多边形，然后求外包络多边形的最小外接矩形，从图中可以看出最小外接矩形比文本区域区域的轴对齐外接矩形更确切地表现了文本区域的位置，然而这也造成了判断文本行是否被检测到的难题，如图，如果直接计算估计的文本区域位置 D 和真实的文本区域位置 G 的重叠率时间复杂度很高，我们通过计算旋转后的 G’和 D’来求其重叠率，其计算公式如下：

$$O(G,D) = \frac{A(G' \cap D')}{A(G' \cup D')} . \tag{3-27}$$

其中$A(G' \cup D')$表示G'和D'并集的面积，$A(G' \cap D')$表示G'和D'重叠区域的面积。

如果估计的文本区域的位置和真实文本区域的位置夹角小于 pi/8,并且重叠率超过 0.5，则认为估计的文本区域位置为正确的文本区域位置。

本文采用广泛采用的准确率和召回率和 F 测度，其定义如下：

准确率：检测出的确实是正确的文本区域/所有被判断为文本的区域

$$p = \frac{|TP|}{E}, \tag{3-28}$$

召回率：检测出的确实是正确的文本区域/图像中真实存在的文本区域

$$r = \frac{|TP|}{T} , \tag{3-29}$$

F 测度：准确率和召回率的调和平均值

$$F = \frac{2pr}{p+r} . \tag{3-30}$$

其中，TP 表示检测出的确实是真实文本区域的个数，E 表示所有被判断为文本区域的个数，T 表示真实存在的文本区域个数。

（3）数据集

本文采用 ICDAR2003[42]数据集数据集进行训练和测试，并在 MSRA-TD500[43]数据集上进行简单的测试，以衡量算法的通用性。ICDAR2003 数据集是国际文档分析和识别大会（ICDAR）组织的 Robust Reading Competition 竞赛公开的部分数据集，该数据集被广泛用来测试自然场景图像文本定位的结果。其中包含训练集和测试集图像共 509 张，其中 258 张图像用作训练，其余 251 张用作测试，测试集中图像的尺寸从 307×93 到 1280×960 像素变化。

但 ICDAR2003 数据集中只有英文文本而且只有水平方向，相对自然场景复杂情况和多语言多文字的实际要求上还不能满足,因此本文另外采用了 MSRA-TD500 数据集进行简单，该数据集中共有 500 张图像，其中 300 张用作训练，200 张用作测试，数据集中图像的分辨率变化范围为 1296×864 到 1920×1280 像素。特别说明的是该数据集包含中文和英文两种语言，文本也不局限于水平方向，而且在字体和背景的变化上更加接近于自然环境，比如字体在字形、颜色、尺寸和方向上变化更复杂，背景中存在树叶、草地和窗户等类似文本的区域，因此更具有挑战性。

图 3-12 ICDAR2003 数据集样例

图 3-13 MSRA-TD500 数据集样例

3.5.2 实验结果和分析

（1）训练样本和测试样本

本文的训练数据主要来自 ICDAR2003 提供的训练数据库，这些训练样本在字形、颜色、排列方式、背景上都差异很大。字符水平的训练正样本来自经过启发式规则验证得到的和真实文本字符重叠率超过 60%的候选文本字符，不满足条件的候选字符作为负样本，另外一部分负样本来自对不含文本的如树叶、草丛等图

像上通过随机采样得到。在训练前首先对样本进行预处理，包含去噪、二值化等操作，将样本制作为白色背景黑色字体或黑色字体白色背景的两类标准样本后再训练。字符水平的正样本和负样本示例如图 3-14 和图 3-15 所示。

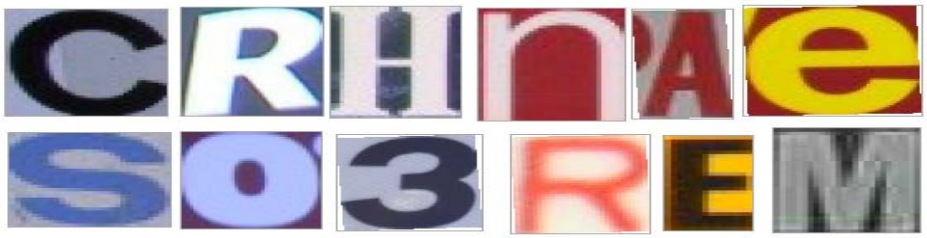


图 3-14 字符水平正样本示例

图 3-15 字符水平负样本实例

文本行正样本是 ICDAR2003 训练图像中的一个单词或者文本行等含有文字区域的图像，负样本的选取是从树叶、草丛等纹理丰富但是非文本区域的图像中随机选取。文本行水平正样本和负样本示例如图 3-16 和图 3-17 所示。

图 3-16 文本行正样本示例

图 3-17 文本行负样本示例

（2）实验结果

为了测试几何约束对候选字符黏连情况的影响，本文定义了候选字符黏连比，其定义为黏连的字符数和真实字符总数的比，同时本文也测试了候选字符生成所需要的平均时间。其测试结果如表 3-1 所示。由测试结果可知，几何约束有效地避免了字符黏连情况，另外由于几何约束在笔划射线的查找过程中提供了除边缘点梯度方向对称性之外更多颜色距离、方向等信息，避免了无效笔划射线查找的时间消耗，降低了候选字符生成的时间。

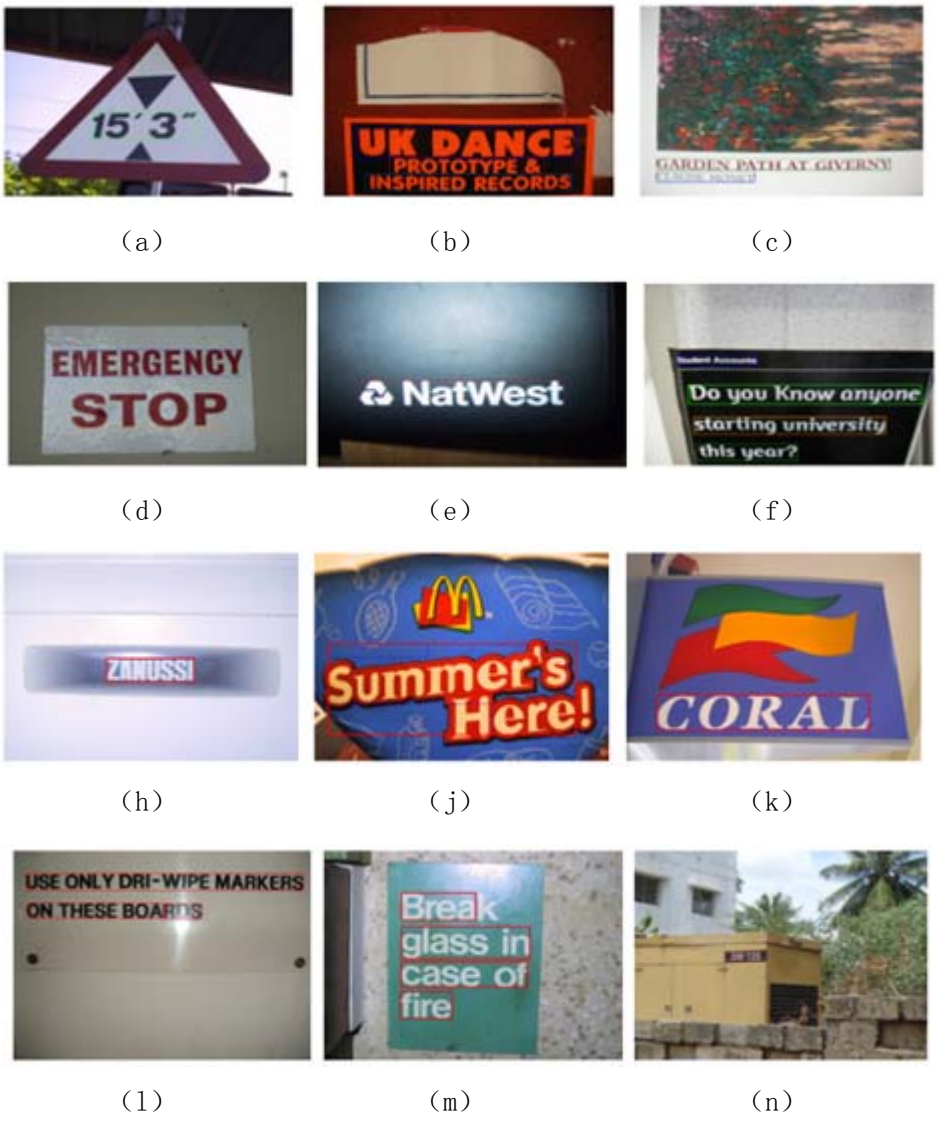
表 3-1 候选字符黏连情况测试

方法	字符黏连比(%)	生成候候选字符所需要的时间（s）
几何约束 SWT	3	2.3
传统 SWT	8	3.8

在 ICDAR2003 上的测试结果如表 3-2 所示，由表 3-2 可知本文改进的几何约束 SWT 在召回率指标上比传统 SWT 提高了 3%，但由于本文在候选文本区域生成阶段增加了文本检测的结果，造成了准确率有小幅下降，这是因为一些自然场景图像中由于拍摄的角度和背景复杂等问题，使得一些文本区域并不一定具有高显著性和高边缘密度特性，同时其文本区域的位置和尺寸分布也与数据集中绝大部图像不同。但在解决候选字符黏连问题上，本文提出的几何约束 SWT 得到了显著改善，改善了传统 SWT 边缘点对梯度方向的对称性条件，降低了对图像边缘的依赖程度，为进一步提高 SWT 在图像低对比度、图像模糊等情况造成的边缘缺失等情况下的效果具有重要的借鉴意义。而且由于笔划射线查找过程中几何约束的限制，避免了大量无效笔划射线的查找过程，使得候选字符生成时间消耗也降低了近 40%。

表 3-2 不同算法在 ICDAR2003 上的测试结果

方法	准确率	召回率	F 指标
几何约束 SWT	0.71	**0.63**	0.67
传统 SWT	**0.73**	0.60	0.66
Hinnerk Becker	0.62	0.67	0.62
Alex Chen	0.60	0.60	0.58
Wolf	0.30	0.44	0.35
Full	0.1	0.06	0.08



（a） （b） （c）

（d） （e） （f）

（h） （j） （k）

（l） （m） （n）

图 3-18 文本定位成功的例子

由图 3-18 可知，基于几何约束的笔划宽度变换的文本定位方法可以很好的抑制自然场景图像上树叶（图 3-18（a）和（n））、花丛（图 3-18（c））等易形成类似文字的噪声区域，此外也可以看到，对如图 3-18（a）中的三角形标志、图 3-18（k）中的旗帜、图 3-18（n）中的窗户、透气栅等本身和文本字符十分相似的区域也被准确地过滤掉。在光照不均匀（图 3-18（e）、（m）和（l））和文字的适度仿射和透视变形（图 3-18（a）、（f）、（j）和（k））情况下也能准确地定位出文本区域的位置。

为了进一步衡量基于几何约束的笔划宽度变换的文本定位效果，本文与公开的 Neumann[37]等人的结果做了对比。

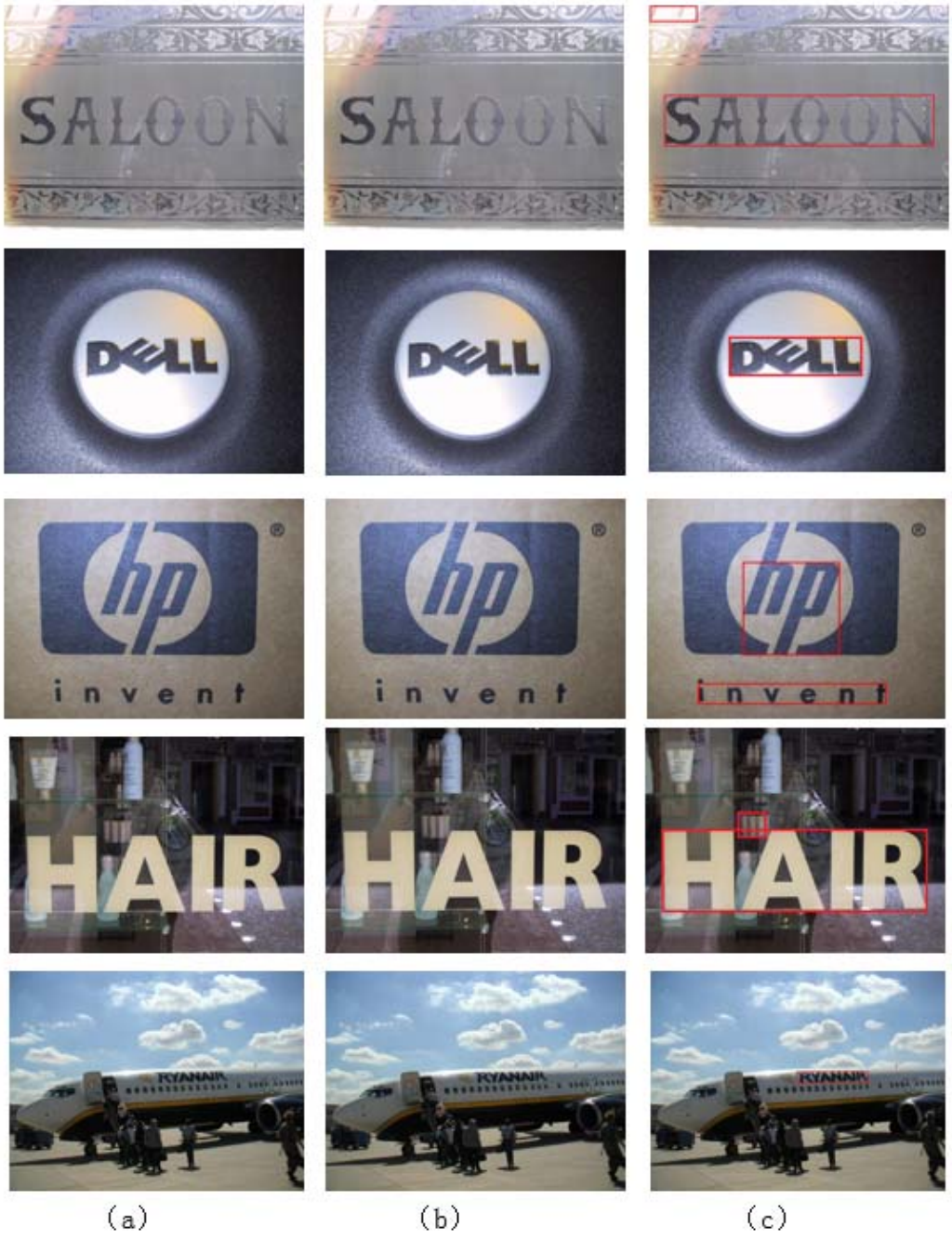


图 3-19 与 Neumann 公开结果的对比：

（a）自然场景图像；（b）Neumann 的定位结果；（c）本文的定位结果

由图 3-19 可知，本文提出的基于几何约束的笔划宽度变换算法，在低对比度（图 3-19 中第一行）、背景噪声复杂（图 3-19 中第二行）、字体差异巨大（图 3-19 中第三行）和背景纹理复杂（图 3-19 中最后一行）等情况下依然可以准确地定位

出文本区域的位置，而在这些情况下，Neumann 等人提出的算法却完全定位失败，没有检测出文本区域，但同时也可以看出，本文改进的基于几何约束笔划宽度变换的文本定位算法也有一定的误检，把一部分和文字十分相似的背景纹理判断为了文本区域，以后还需要进一步提升过滤准则，进一步提高算法的准确率。

（3）失败案例分析

虽然本文提出的基于改进的笔划宽度变换的在自然环境下的文本检测效果比较高，但其作为局部描述子具有双面性，既有精确的细分能力，也同时容易受噪声的影响，仍有很大的改进空间，如图 3-20 中的文本定位失败的情况:（a）对比度太低;（b）图像模糊情况严重;（c）光照不均匀情况严重且文字和复杂背景重叠;（d）字符尺寸太大;（e）笔划宽度特征不连续。在这些情况下，往往提取不到图像的边缘，或者出现重复边缘，进而造成提取的笔划特征不准确，候选字符生成失败等，影响了最终文本定位的结果。

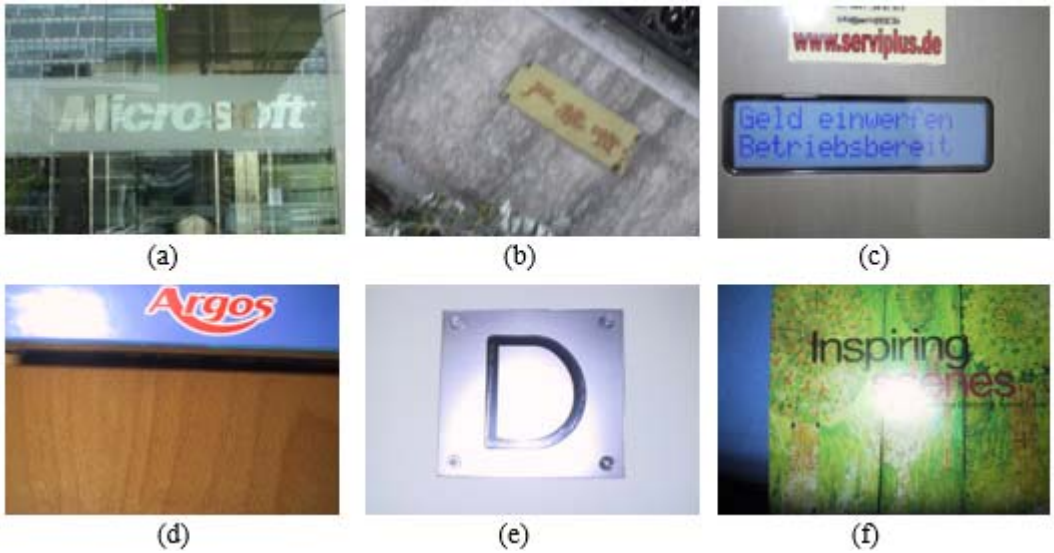


图 3-20 文本定位失败的例子

3.6 本章小结

本章利用基于几何约束的笔划宽度变换算法进行文本定位，首先介绍了传统的笔划宽度变换算法，并详细分析了影响笔划宽度变换的平滑滤波、边缘检测等各种因素之后，然后提出了改进的基于几何约束笔划宽度变换算法。然后在笔划宽度变换的基础上利用区域生长算法得到候选字符，在候选字符和文本行的筛选阶段，采用双阶段过滤机制，分别是基于启发式的规则验证和基于随机森林的分类器分类。在字符水平和文本行水平的特征设计上，本文创新地利用文本检测的结果转化得到的候选字符和文本行的全局描述特性，在一定程度上克服了笔划宽

度变换算法的局部敏感缺陷。此外，在文本行水平的特征设计上，又综合考虑了文本行内的字符在尺寸、方向上一致性特性，设计了全新的特征。最后分析了文本定位算法评价的难点，并介绍了本文采用的文本定位算法评价的方法，评测数据集及本文改进的基于几何约束笔划宽度变换的自然场景文本定位的实验结果并进行了分析。

第四章　几何约束笔划宽度变换在新闻视频字幕定位中的应用

随着计算机技术及数字多媒体计算的发展，多媒体信息呈爆发式增长，人们迫切的需要对海量视频进行标记、分类和检索，传统的人工标注的方法效率低下且不准确，如何像基于文本的搜索引擎那样对多媒体信息进行检索和分类成为了人们的迫切需求。要实现对新闻视频的检索，就需要对视频建立索引，相对于视频图像内容和音频来说，视频中的文字信息不仅信息量大，而且更容易提取，例如在新闻视频中人工添加的字幕和标注是对视频内容的高层语义概括，新闻视频上的标题很好的概括了新闻发生的时间、地点和人物等关键信息，而且有时候这些信息并不一定会在音频中出现，也无法通过对视频内容的分析得到。对定位出的字幕和标注进行识别后，就可以像传统的搜索引擎那样建立检索，这无疑会大大提高人们对视频处理的便利性。

视频中的文本信息对理解视频的内容具有重要的意义，字幕文本定位及其进一步的新闻视频文本提取应用是面向新闻流数据的多媒体数据分析、建模和时空关联技术的重要组成部分。本文在基于几何约束的笔划宽度变换算法的文本定位的基础上，使用 OpenCV2.4.10 计算机视觉图像处理库、C++语言对新闻视频字幕文本检测系统进行设计并实现。

4.1 系统原理介绍

新闻视频字幕文本定位系统主要包含图像预处理、文本检测和文本定位模块，其系统流程如下。

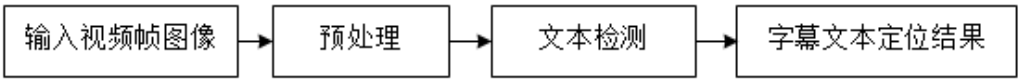


图 4-1 新闻视频字幕文本定位流程

预处理模块是为笔划宽度变换做准备，主要包括灰度化、图像去噪、图像增强、边缘检测和梯度计算等处理过程，为笔划宽度变换算法提供边缘和梯度信息。

文本检测模块的目的是初步确定文本区域的位置，本文将文本检测模块的结果转换为候选文本区域和全局性特征度量的方法已在第二章中已经详细阐述，此处不再介绍。

文本定位是确定视频中文本区域的位置，又可细分为文本检测和文本定位两个部分。文本检测和文本定位采用的方法大致相同，如在第一章介绍的基于边缘的方法、基于纹理的方法、基于连通域的方法。有很多文献将文本检测和文本定

位作为一个模块。经过文本定位这个步骤，将得到的文本区域的位置坐标输入到下一个步骤中的文本分割阶段。本文采用的文本定位方法已在第三章中进行了详细阐述，此处不再介绍。

下面以单帧新闻视频图像来介绍本文新闻视频字幕文本定位的流程：

图 4-2 视频帧图像

经过文本定位步骤，得到如下图所示的文本图像，得到如下的定位区域

（a） （b）

图 4-3 文本图像：（a）暗色文本在亮色背景上；（b）亮色文本在暗色背景上

为了将来进一步的文字识别应用，还需要对文本图像进行分割，文本分割的目的是进一步减少背景噪声，突出文本。经过文本分割步骤，得到如图 4-4 的的黑色字体在白色背景的待识别图像。

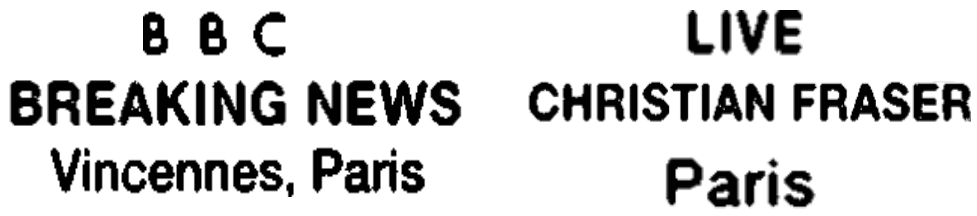


图 4-4 文本分割图像

由上图可知，文本分割图像中作为前景的文本区域突出，非文本的背景区域简单没有噪声，其文本分割的图像经过一些简单的处理就可以直接作为文本识别引擎

的输入。

4.2 系统模块设计

新闻视频文本定位系统主要有图像预处理模块、文本检测模块、几何约束笔划宽度变换模块、字符水平和文本行水平的文本区域生成和过滤模块组成。其中文本检测和几何约束笔划宽度等模块已在第二章中进行了详细介绍，在本章将结合新闻视频字幕的特点，介绍候选字符标记和过滤模块、 针对新闻视频的文本行生成和验证模块及为进一步文本识别所做的二次分割模块。

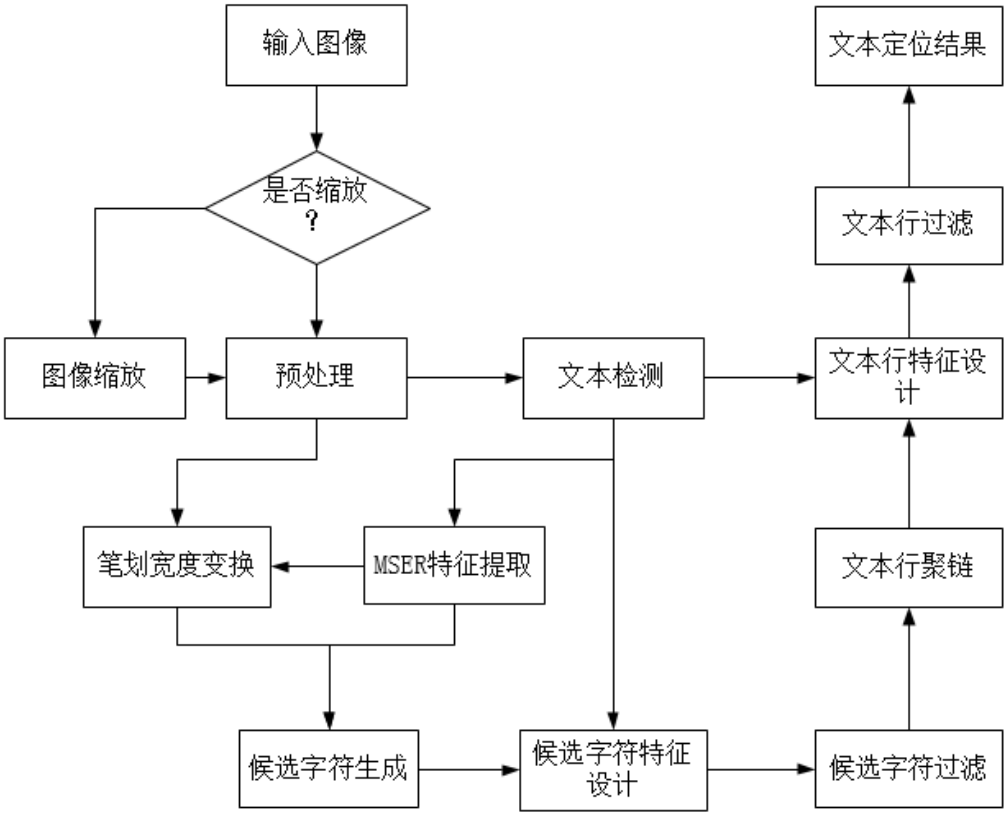


图 4-5 新闻视频字幕提取文本定位模块设计

文本定位模块的输入为一幅彩色的视频帧图像，对图像的格式要求不限，输出为定位得到的文本图像。但与很多基于纹理的和一些基于连通域的算法不同的是，本文文本定位的主要思想是由下而上在像素水平、字符水平和文本行水平实现文本区域由小到大聚合，然后是文本区域像素水平的筛选、候选字符生成和过滤、候选文本行生成的生成和过滤，最后得到文本区域。然后使用从上至下的过滤思想分别在文本行水平过滤掉在字符水平无法过滤掉的类文本字符，在字符水平过滤掉在像素水平无法过滤掉的非文本像素，因此其具有天然的文本分割优势。

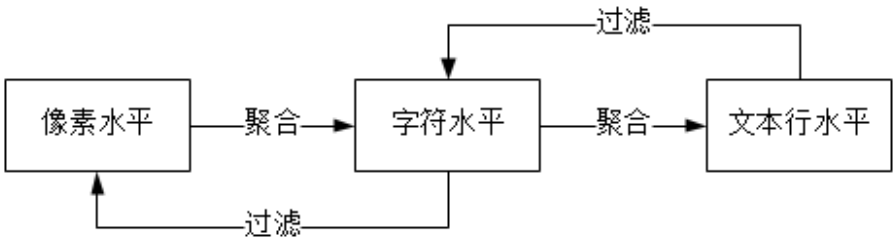


图 4-6 三层过滤思想

本文使用的笔划宽度变换算法和采用的 MSER 特征在像素水平上已经完成了文本的初步分割，另外本文提出的改进的笔划宽度变换算法对字符内部的孔洞和缝隙又进行了修补，因此，文本定位流程的输出的文本图像已是灰度图像，其文本区域突出，背景区域几乎没有噪声，对文本识别系统来说，已经满足条件。

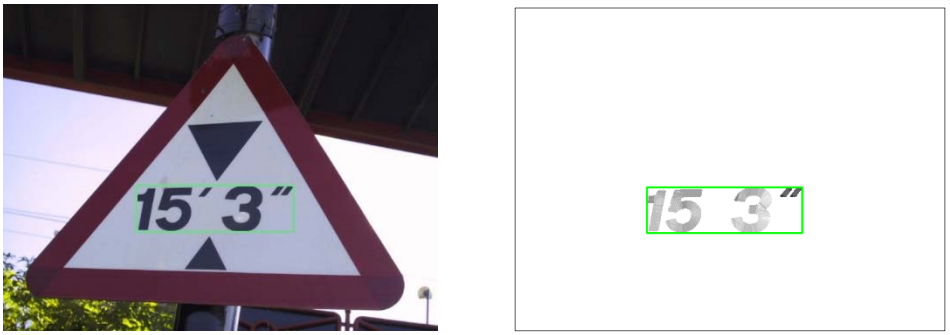


图 4-7 几何约束笔划宽度变换的分割效果

4.3 系统模块实现

4.3.1 预处理

预处理模块主要由灰度化、图像增强、平滑滤波、边缘检测和梯度计算等部分组成，其目的是为了得到笔划宽度变换算法所需要的边缘图像和边缘像素点的梯度方向信息，这些信息为下一步笔划宽度变换提供输入。预处理模块的流程如图 4-8 所示。

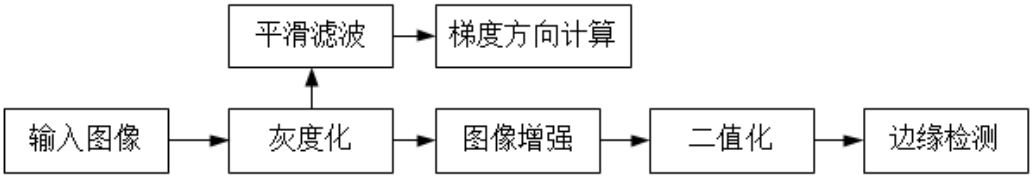


图 4-8 预处理模块设计

4.3.1.1 灰度化

为了减小计算量，预处理阶段首先对彩色图像进行灰度化处理。灰度图像与彩色图像相比，其边缘信息损失不大，但简化了后面的处理。本文中采用的彩色图像到灰度图像的变换公式如下：

$$G(i,j) = 0.299 \times G_r(i,j) + 0.587 \times G_g(i,j) + 0.114 \times G_b(i,j)\,. \qquad (4\text{-}1)$$

其中$G(i,j)$、$G_r(i,j)$、$G_g(i,j)$、$G_b(i,j)$分别代表灰度图像上的灰度值和彩色图像上的 R、G、B 颜色值。

4.3.1.2 平滑滤波

自然场景图像中往往包含各种噪声，为了提高图像边缘定位和梯度方向计算的准确性，需要首先进行去噪处理。但要求去除噪声的同时要求保持边缘的清晰。传统的加权滤波算法虽然具有较强的去噪能力，但降低了边缘的强度。中值滤波算法对椒盐噪声有较强的去噪能力，同时又能保持图像细节。其基本原理是用模板扫描图像，将模板中心与待滤波的像素重合，统计并排序模板覆盖下周围像素值，取其中值作为该像素点的中值滤波响应值。

虽然中值滤波在消除噪声和保持边缘之间取得了很好的平衡，但中值滤波的时间消耗一般是加权滤波的 5 倍以上，而且随着扫描窗口的增大，边缘强度信息减弱，时间消耗迅速增加。

针对传统中值滤波算法的不足，本文采用 Deng 等[45]提出的一种新的自适应加权中值滤波算法。该算法首先检测出图像中的噪声点，不同于传统中值滤波算法采用固定扫描窗口大小，而是在扫描的时候根据窗口内的噪声点数量自适应调整扫描窗口的大小，同时将窗口内的噪声点按照特定规则进行自适应分组，并根据相似度计算各组像素点的权重，最后利用加权滤波算法对检测到的噪声点进行滤波。

4.3.1.3 二值化

图像二值化是图像阈值分割的一种特殊情况，在二值化前还需要进行去光照不均处理、直方图均衡化等操作增强图像的对比度。

（1）形态学顶帽变换解决光照不均问题

自然场景拍摄的图像中往往由于光照的变化导致拍摄到的图片亮度不均匀，这对边缘的准确提取有很大的影响，因此首先需要尽可能减弱光照不均问题带来的影响。

通过观察发现，自然场景中的图像亮度分布情况呈圆形或扇形，因此本文采

用圆形的结构元素对灰度图像进行开运算，提取出亮度背景曲面，然后利用形态学顶帽变换减去亮度背景，最后调整图像对比度，得到背景亮度均匀且高对比度的灰度图像。

（2）直图均衡化

直方图是用像素灰度值的统计分布来表征图像，是数字图像处理领域广泛采用的特征。直方图均衡化是通过点运算修改原始图像的灰度值，使得新图像的灰度值分布范围更大且均匀，从而增强图像对比度。

（3）Niblack 二值化

二值化算法根据所使用的阈值的不同，可以分为全局二值化、局部二值化。全局二值化算法采用固定阈值，对图像上的所有像素进行同样的操作，其优点是处理速度快，但是效果没有局部二值化算法好，其主要应用于背景和前景灰度值差异比较明显，灰度直方图呈现明显的双峰模式的情况。典型的全局二值化方法如 Otus。局部二值化的阈值则是采用动态阈值，该阈值的选取取决于领域的像素值，典型的局部二值方法有 Niblack 方法、Bersen 方法、Sauvola 方法，Niblack 和 Bernesen 可以很好的分割开字符笔划与周围的背景，但对无文字区域的噪声和背景比较敏感，容易产生虚假尾影，Sauvola 方法在此基础上进行了改进。

自然场景图像由于背景的复杂性，往往不满足全局二值化的条件，因此本文在预处理阶段选用局部二值化的方式。

传统的局部自适应阈值分割方法的阈值一般是通过计算窗口内像素灰度均值，然后再叠加一个常量的偏移得到，而 Niblack 的阈值是通过是计算窗口内像素灰度均值，然后叠加一个窗口内像素灰度值标准差的n倍得到，即：

$$Threshold(i,j) = m(i,j) + n * \delta(i,j) \quad n \in [0,1]. \qquad (4\text{-}2)$$

其中$m(i,j), \delta(i,j)$分别表示窗口内所有像素的灰度均值和方差。经过试验对比，本文采用 Niblack 二值化方法进行处理。

4.3.1.4 边缘检测

边缘检测的依据是图像灰度的不连续性，图像的边缘是灰度值发生阶跃变换或者屋顶变换的位置集合，很好地体现了图像的结构特征，具有很强的显著性。在漫画和插图中，寥寥几条线就可以描述一个场景或物体。同样在图像中，人们对文字最直观的感受是文字的边缘和轮廓等结构信息，而不是颜色和纹理等特征。

边缘检测主要包含平滑滤波、锐化滤波、边缘判定和边缘连接 4 个基本步骤。进行平滑滤波的目的是减少噪声对梯度计算的影响，同时尽可能地保持边界的强度。锐化滤波突出了发生灰度局部变换的的像素点。边缘判定则是根据具体情况

对边缘强度进行阈值限定得到满足需求的边缘的同时尽可能减少噪声边缘。边缘连接是对边缘的修补和剔除操作，使得边缘更加完整和准确。

边缘检测可以大幅度地剔除不重要的信息，保留那些对能代表图像结构熟悉的边缘像素点。传统的基于笔划宽度变换的文本定位方法对边缘检测的准确度很敏感，直接关系到笔划特征的提取的好坏。

边缘检测根据所使用的算子阶数可分为一阶导数法、二阶导数法，此外还有Canny[38]等。一阶导数法有 Roberts、Sobel 和 Prewitt 算子等，二阶导数法主要是高斯-拉普拉斯算子。由于基于一阶、二阶导数方法的边缘检测算法对噪声很敏感，在边缘检测前首先要进行平滑处理，但平滑处理是把双刃剑，在抑制噪声的同时也损失了边缘强度，增强了边缘定位的不确定性。对比各种边缘检测算法的效果后，本文采用 Canny 算子，它在保证边缘定位精度的的同时尽可能增强对噪声的干扰能力之间取得平衡。

4.3.1.5 梯度方向计算

在笔划宽度计算的满足条件的边缘点对查找过程中，需要使用到查找点像素的梯度方向，这里我们采用 Sobel 算子和原始图像做卷积得到。Sobel 算子是一阶导数的离散型差分算子，其算法实现主要包含两组3×3的模板S_x和S_y作为卷积核与图像中的像素点做卷积和运算，分别得到纵向和横向的差分近似值G_x和G_y，$G_x = S_x * G, G_y = S_y * G$，其中 G 代表灰度图像，其梯度的大小为 G,梯度的方向θ。

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{4-3}$$

$$G = \sqrt{{G_x}^2 + {G_y}^2}, \tag{4-4}$$

$$\theta = artan\left(\frac{Gy}{Gx}\right). \tag{4-5}$$

经过以上的预处理后，得到的边缘图像减少了很多噪声，边缘像素更多的集中在文本区域，非文本区域边缘像素较少，减小了对后续笔划宽度变换时候的干扰。

4.3.2 候选字符标记和过滤

本文采用种子填充的方法进行候选字符标记，其实现如下：

（1）遍历笔划宽度变换特征图中的每一点，若遇到笔划像素点，则查找该像素点的四邻域点，否则，继续遍历。

（2）对笔划像素点的四邻域点进行检查，若是笔划像素点，则以继续该邻域

点为起点进行查找，若没有邻域点是笔划像素点，则终止查找，赋于标号值。

在候选字符标记的基础上，继续计算在 3.2.3 中定义的字符笔划宽度方差、字符宽高比、字符像素占空比、字符像素笔划宽度中值与笔划宽度最大值的比和候选字符的面积等特性，并根据设置的条件滤除不满足的非文本字符，完成字符水平层第一阶段的非文本字符过滤。

候选字符的过滤第二阶段采用随机森林的方法，其特征采用 3.4.2 中描述的字符水平的特征，训练样本的制作也按照 3.5.2 节中所描述的方法。

4.3.2 文本行生成和验证

本文在处理过程中假定文本行往往是一个单词，若单词之间间距很小，则文本行是一行单词，进行文本行生成的主要目的是过滤一些在字符水平无法剔除的非文本字符。文本行的生成算法和验证规则如下。

在文本字符基础上，首先进行文本对的生成，文本对是由两两字符组成的最小文本行，两个文本字符是否组成文本对由 3.3.1 中所设计的规则为准，对文本对聚合成文本行的情况，考虑到新闻视频字幕的水平分布情况，采用以下的准则进行聚合。

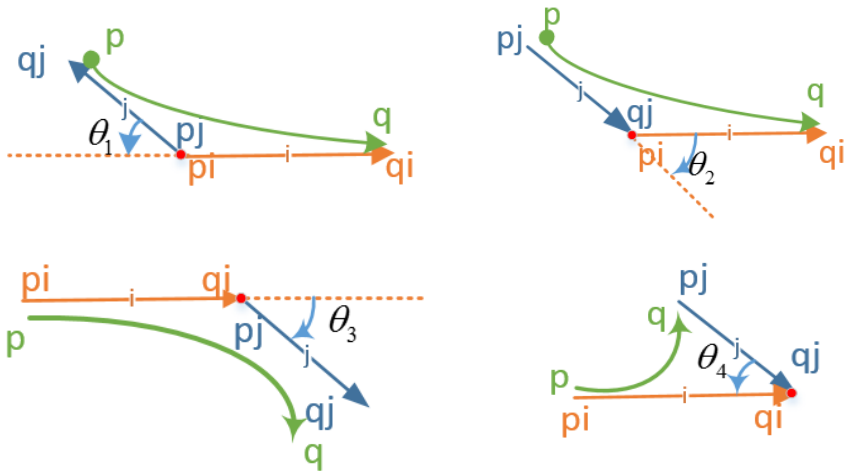


图 4-9 文本行聚链规则

对存在以上情况的字符对和文本行，若图中所标注的两个文本行之间的方向满足如上图所示的情况，p_i和q_i代表文本行 i 的行首和行尾，p_j和q_j代表文本行 j 的行首和行尾，图中标注的$\theta_1,\theta_2,\theta_3,\theta_4$代表本文判别两个文本行是否进行合并的方向控制角，p、q 代表聚合成的新的文本行的行首和行尾。

本文采用的文本行生成算法是为多方向的文本检测所设计的，考虑到新闻视频中的文本主要是水平方向的具体情况，因此在实际的参数设置中，本文设置

$\theta_1, \theta_2, \theta_3, \theta_4$均为$\pi/12$。

4.3.3 文本图像二次分割

为了解决部分文本字符被意外过滤掉的情况，本文在定位出的文本图像上再次进行分割，利用文本行内字符的紧邻特性，召回三层过滤思想中第二层字符水平过滤阶段被过滤掉的字符，这些被过滤的字符主要是因为和非文本字符高度相似而被过滤掉。

（a） （b）

图 4-10 文本图像二次分割：（a）原始图像；（b）二次分割

在上图中字符‘i’在字符水平的过滤阶段由于和竖状条高度相似而被意外过滤掉，当在文本行聚链生成阶段又重新被标定在文本区域内，但由于送给识别引擎的输出是文本行聚链的结果，其内部是已经被分割后的字符，而字符‘i’等不在分割后的文本行内，因此需要在定位出的文本区域内进行二次分割，由于基于笔划宽度变换的文本定位阶段已经把文本区域定位到了视频字幕区，而为了便于阅读，新闻视频中字幕区域的位置背景色相对单一，因此在二次分割阶段，本文采用漫水算法,得到的结果如图 4-10（b）所示，从而使得文本定位和分割模块的输出结果更准确。

4.4 本章小结

本章介绍了几何约束笔划宽度变换算法的文本定位在新闻视频字幕文本定位系统中的应用，结合新闻视频字幕文本的特点，对字符形成文本对和文本对形成文本行的算法进行了特殊处理，另外采用二次分割的方法召回被意外过滤掉的文本字符，为进一步的的文本识别提供较好的输入。

第五章 总结和展望

随着智能移动设备和数字技术的发展，图像中的文本将呈爆发式的增长，而图像中的文本信息，如路标、指示牌等对机器人导航、图片检索具有重要的意义。另外视频中的人工标注的文本和视频原生的文本等的提取对海量视频检索也提供了重要的依据。

但是目前的自然场景图像文本提取相关技术，如文本定位、文本分割和文本识别等还有很多的难题要攻克。作为自然场景图像文本信息提取的基础环节，文本定位具有极其重要的作用，本文在基于几何约束笔划宽度变换算法的文本定位方法的基础上，结合新闻视频字幕提取的具体应用背景，对新闻视频字幕文本定位的整套流程进行了实践。

5.1 总结

自然场景图像文本定位因为在文字、背景和图像本身的复杂性特点，是计算机视觉、图像处理和模式识别的交叉领域，目前仍然是一个十分具有挑战性的难题，文本在分析文本定位的研究现状和方法后，提出了一种基于几何约束的笔划宽度变换算法的文本定位方法，又结合新闻视频字幕检测的具体应用，完成了新闻视频字幕文本定位系统的设计和实现。

本文的主要工作和成果如下：

首先，在传统的笔划宽度变换算法基础上，本文利用几何约束在笔划射线的查找方向和生长过程进行了改进，有效地避免了字符黏连等情况，使得候选字符的生成质量得到了提高，此外对笔划宽度异常值的修正又进一步改善了候选字符的质量，为后面的字符水平和文本行水平阶段的处理提供了较好的输入。

其次，在字符水平和文本行水平中加入了文本检测阶段转换的文本区域显著性特征、边缘密度及尺寸和位置等全局性的特征度量，一定程度上弥补了传统的笔划宽度算法的局部敏感性，在剔除虚假字符的同时，也保留了真正的文本字符，和传统的笔划宽度变换算法相比，取得了较好的定位效果。

最后，在几何约束笔划宽度变换的基础上，结合新闻视频字幕的特点，设计和实现了新闻视频字幕文本定位系统。此外对文本定位的图像进行二次分割操作，召回在字符水平阶段那由于和非文本字符高度相似而被意外过滤掉的文本字符，增强了文本识别输入的字符质量，对提高文本识别结果有重要意义。

5.2 展望

虽然本文提出的基于几何约束的笔划宽度变换算法在自然场景图像文本定位中取得了较好的效果，但还有很多问题要解决，未来的工作主要从以下两方面展开。

（1）文本极性检测

由于本文在笔划宽度变换的时候使用的搜索方向是预先指定的，虽然本文对亮文本在暗背景和暗文本在亮背景的情况下分别进行了处理，然后融合两种结果，但这种融合算法增加了很多的不确定性。而现实世界中的情况千变万化，既有暗文本在亮背景上，也有亮文本在暗背景上，甚至在一张图像上这两种情况都存在，这给笔划宽度的计算带来了很大困难，如何自动的确定笔划搜索的方向对提高自然场景文本定位的结果，进而推动自然场景文本提取系统实用化具有重要的意义。

（2）利用识别结果的反馈修正文本定位

由于自然场景在光照不均、噪声、低对比度、抖动造成的模糊等方面的特殊情况，分割出的文本字符往往并不能很好体现文本字符的原貌，过度分割可能会把一些本属于文本字符的像素分为背景像素，欠分割可能会把一部分背景像素当成文本字符像素，无论哪种情况，都可能会对后续文本识别的结果造成很大的影响。如何根据识别的结果反馈指导文本分割的程度，形成一个文本定位和识别的闭环反馈系统，对提升文本定位和识别的结果的准确性都具有重要的意义。

复杂背景下的文本定位还是一个新生的领域，其研究对于移动互联时代人们的生活具有重大的影响，同时也具有很大难度和挑战性，虽然很多的研究者在文本定位方面做了不少的探索，但由于自然场景的复杂性，同时又兼具传统的图像分割和识别的难题，离真正的实用化还有很长的路要走，但其前景是十分广阔的。

（1）基于文字内容的视频和图像检索

随着 Internet 的发展，数字化图像信息爆发式增长，如何从海量的数据中快速找到人们所需要的有价值的知识已成为人们的迫切需要，大数据技术、搜索引擎和推荐系统等技术也为了减轻人们从海量信息中获取有用信息的难度而出现，然而这些技术大多是直接处理数字或文字信息，对于图像和视频信息的处理做得好的并不多，远不能满足人们对图像分类和检索的需求。因此，智能图像分类和智能图像检索技术成为人们迫切的需要，传统的基于内容的图像检索技术和基于内容的视频检索技术,都需要对图像和视频的高层特征进行提取研究，利用的信息大多是图像的纹理、颜色、角点等特征，然后进行建模和分类。而在一些情况下，图像中的文字信息，尤其是新闻视频中人工添加的字幕等文字信息，对图像的内

容进行了更准确的高层语义概括，提取和检索这些信息对图像和视频的理解具有很大的帮助。

（2）基于移动终端的智能图像分析系统

随着移动互联网的高速发展，越来越多的文字信息也以图像的方式呈现，各种基于相机的应用随之蓬勃发展。人们利用随身携带的手机拍摄自然场景中的路标、车牌等，如果这些图像中的文本能自动地被定位、分割和识别，再结合机器翻译和语音合成技术，则会极大的便利人们的生活。然而现有的文本定位方法相对于人们的需求，时间和空间复杂度都相对较高，对硬件的要求也使得当前的智能终端难以满足，虽然智能终端的处理能力近些年来得到了很大的增强，然而对于计算和空间复杂度较高的模式识别和人工智能方法仍然显得捉襟见肘，而近些年兴起的云计算则很好地解决了这些问题。

云计算是通过网络将复杂的处理任务交给服务器集群进行计算分析，然后将结果返回给用户，随着移动网络带宽的提高，人们使用移动网络的体验也更加流畅。例如，人们可以使用智能终端拍摄看到的文本图像，然后将图像发送到部署了文本检测、识别和翻译的云服务器上，由云服务器完成整套的信息处理，最后将处理结果返回给用户，增加交互功能，终端可以对云服务器识别的结果进行打分，云服务器据此反馈改进分类算法，这样不断通过在线训练，提高检测和识别的效果。这样的即时翻译系统可减轻用户在不同国家旅行的交流障碍。此外利用云的方法进行共享和存储这些信息，结合索引技术，相当于把现实世界中的文字迁移到了互联网上，极大地丰富了信息获取的渠道。

利用智能终端获取图像，然后利用云服务进行图像分析的组合来获得对场景的认知和理解有着广泛的应用前景。

致　谢

在论文完成之际，特向给予我支持、帮助和鼓励的师长和同学，以及亲友表示衷心的感谢！

首先感谢我的指导老师程洪教授，在研究生阶段，无论是在人生观的树立、还是科研方法的形成及科研态度的培养上，都得到了导师悉心的指导，在我论文的选题、研究、实验验证和撰写及排版的过程中均凝聚着导师大量的辛苦付出。程老师治学严谨、实事求是，对待我们遇到和提出的问题谆谆教导,将复杂难懂的问题清晰明了的阐述出来，给我留下了深刻的印象。

感谢杨路老师，杨路老师工作上勤奋努力，生活上对每一位同学都平易近人，在百忙之中仍然关心着我们科研论文中遇到的问题，并给出指导性的意见，让我少走了很多弯路，在研究生生活中受益良多。

此外，感谢我的父母和姐妹，正是他们的关心和支持才能让我成长至今，他们是我遇到困难时候坚持不放弃的动力和源泉，是他们二十多年来的辛苦付出让我顺利完成从小学到研究生的学业。尤其是我的父母，为我的成长付出了大量的心血，在此毕业之际，再次向他们表示衷心的感谢。

还要感谢教研室帮助过我的师兄和同学，特别是陈启明和周楠博士，在我科研的道路上给予了很大的帮助，在生活上也给了很多的建议。

最后，感谢参与评审的各位专家和老师，感谢大家在百忙之中对我论文进行评阅并提出宝贵的意见。

参考文献

[1] S. S. Tsai, H. Chen, D. Chen, et al. Mobile visual search on printed documents using text and low bitrate features[C]. In ICIP, 2011.

[2] G. N. DeSouza and A. C. Kak. Vision for mobile robot navigation: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 237-267.

[3] D. B. Barber, J. D. Redding, T. W. McLain, R. W. Beard, et al. Vision-based target geo-location using a fixed-wing miniature air vehicle[J]. Journal of Intelligent and Robotic Systems, 2006, 47(4): 361-382.

[4] K. H. Zhu, F. H. Qi, R. J. Jiang, et al. Automatic character detection and segmentation in natural scene images[J]. Journal of Zhejiang University -Science A, 2001, 8(1):63-71.

[5] Y. Zhong, K. Kalle, and A. K. Jain. Location text in complex color images[J]. Pattern Recognition, 1995, 28(10):1523-1535.

[6] H. P. Li, D. Doermann, and O. Kia. Automatic text detection in digital video[J]. IEEE Transactions on Image Processing, 2000, 9(1):147-156.

[7] K. C. Kim, H. R. Byun, Y. J. Song, Y. W. Choi. Scene text extraction in natural scene images using hierarchical feature combining and verification[C]. In ICPR, 2004.

[8] D. Chen, H. Bourlard, and J. P. Thiran. Text identification in complex background using SVM[C]. In CVPR, 2001.

[9] 欧文斌，朱军民，刘昌平．自然场景文本定位[J]．中文信息学报，2003.17(5):55-60.

[10] J. Gao and J. Yang. An adaptive algorithm for text detection from natural scenes[C]. In CVPR, 2001.

[11] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images[C]. In ICDAR, 1995.

[12] H. P. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video[J]. IEEE Transactions on Image Processing, 2000, 9(1): 147-156.

[13] W. G. Mao, F. Chung, et al. Hybrid Chinese/English text detection in images and video frames[C]. In PR, 2002.

[14] Q. Liu, C. Jung, S. Kim, Y. Moon, and J. Kim. Stroke filter for text localization in video images[C]. In ICIP, 2006.

[15] X. Chen and A. Yuille. Detecting and reading text in natural scenes[C]. In CVPR, 2004.

[16] M. R. Lyu, J. Song, and M. Cai. A comprehensive method for multilingual video text detection, localization and extraction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(2): 243-255.

[17] H. K. Kim. Efficient automatic text location method and content-based indexing and structuring of video database [J]. Journal of Visual Communication and Image Representation, 1996, 7(4): 336-344.

[18] A. Jain and B. Yu. Automatic text location in images and video frames[J]. Pattern Recognition, 1998, 31(12): 2055-2076.

[19] Y. Liu, S. Goto, and T. Ikenaga. A contour-based robust algorithm for text detection in color images[J]. IEICE Transactions on Information and Systems, 2006, 89(3): 1221-1230.

[20] 江斌，胡福乔. 基于图理论聚类的彩色图像文本提取[J]. 微电子学与计算机，2003, 20(8):89-93.

[21] 章东平，祝金标，刘济林. 自动定位彩色图像中的文本[J]. 浙江大学学报，2005, 39(2):229-233.

[22] J. Y. Zhou and D. Lopresti. Extracting text from WWW images[C]. In ICDAR, 1997.

[23] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform[C]. In CVPR, 2010.

[24] L. Neumann and J. Matas. Real-time scene text localization and recognition[C]. In CVPR, 2012.

[25] C. Yao, X. Bai, W. Liu, Y. Ma, et al. Detecting texts of arbitrary orientations in natural images[C]. In CVPR, 2012.

[26] X. C. Yin, X. Yin, K. Huang, and H. W. Hao. Robust text detection in natural scene images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 970-983.

[27] H. K. Kim. Efficient automatic text location method and content-based indexing and structuring of video database[J]. Journal of Visual Communication and Image Representation, 1996, 7(4): 336-344.

[28] L. Neumann and J. Matas. A method for text localization and recognition in real-world images[C]. In ACCV, 2010.

[29] S. Tekinalp and A. A. Alatan. Utilization of texture, contrast and color homogeneity for detecting and recognizing text from video frames[C]. In ICIP, 2003.

[30] X. Chen and A. L. Yuille. Detecting and reading text in natural scenes[C]. In CVPR, 2004.

[31] L. Neumann and J. Matas. Scene text localization and recognition with oriented stroke detection[C]. In ICCV, 2013.

[32] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.

[33] X. Hou and L. Zhang. Saliency detection: A spectral residual approach [C]. In CVPR, 2007.

[34] J. Matas, O. Chum, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004, 22(10): 761-767.

[35] D. Nistér and H.Stewénius. Linear time maximally stable extremal regions[C].In ECCV, 2008.

[36] H. Chen, S. S. Tsai, G. Schroth, et al. Robust text detection in natural images with edge-enhanced maximally stable extremal regions[C]. In ICIP, 2011.

[37] http://www.textspotter.org/

[38] J. Canny. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8:679-714.

[39] L. Breiman. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[40] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection[C]. In CVPR, 2005.

[41] C. Jung, Q. Liu, and J. Kim. A stroke filter and its application to text localization[J]. Pattern Recognition Letters, 2009, 30(2): 114-122.

[42] http://www.iapr-tc11.org/mediawiki/index.php/ICDAR_2003_Robust_Reading_Competitions#Robust_Reading_and_Text_Locating

[43] http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_ (MSRA-TD500)

[44] S. M. Lucas, A. Panaretos, L. Sosa, et al. ICDAR 2003 robust reading competitions[C]. In ICDAR, 2003.

[45] X. Deng, Y. Xiong, and H. Peng. A new kind of weighted median filtering algorithm used for image processing[C]. In ISISE, 2008.

攻读硕士学位期间取得的成果

[1] 程洪，袁俊淼，杨路. 一种基于笔划特征的自然场景文本检测算法[P]. 中国，发明专利（待授权）