

密 级_____



桂林电子科技大学
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

硕 士 学 位 论 文

(全日制专业学位硕士)

题目_____自然场景图像中的中文文本规范字检测算法研究

(英文)_____Text Detection Algorithm in Natural Scene Images

研 究 生 学 号:_____1403304004

研 究 生 姓 名:_____刘水清

指导教师姓名、职称:_____缪裕青（副教授）

申 请 学 位 类 别:_____工程硕士

领 域:_____图像处理

论 文 答 辩 日 期:_____2017 年 6 月 8 日

独创性(或创新性)声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得桂林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做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：刘永清 日期：2017.6.9

关于论文使用授权的说明

本人完全了解桂林电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属桂林电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为桂林电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。(保密的论文在解密后遵守此规定)

本学位论文属于保密在____年解密后适用本授权书。

本人签名：刘永清 日期：2017.6.9

导师签名：刘永清

摘 要

随着社会的进步和科技的发展,图像的获取变得异常便捷。图像中包含了丰富的信息,譬如颜色、形状、纹理和文字等。对图像中文本信息的检测,在基于内容的图像检索、视频分析、视觉辅助、汽车辅助驾驶、智能交通管理等领域有着广泛的应用。在自然场景图像中,当背景比较复杂时,文本检测是较困难的。本文重点研究自然场景图像中的文本检测技术,主要研究工作包括:

(1) 针对目前缺乏公开的完全基于中文文本的实验图像数据集的问题,建立了一个纯中文文本的实验图像数据集。使用高分辨率的智能手机和数码相机拍摄学校内和学校附近的自然场景图像,选择其中符合条件的 403 幅左右的图像,将之构建成图像数据集。

(2) 针对现有的自然场景图像中背景复杂的中文文本检测效果差的问题,提出了一种基于启发式规则和汉字结构特征的中文文本检测算法 TDSI。首先使用一系列启发式规则分别改进 MSER 算法和 SWT 算法,根据笔画宽度值的标准差将非文本区域过滤掉。然后根据汉字的结构特征,即候选区域的质心、重合区域等特征将文本区域聚集成汉字,最后将之聚集成文本行。实验结果表明,与 Busta、Chen 算法相比,TDSI 算法取得了较好的准确率、召回率和 F 值。

(3) 针对现有场景文本检测算法忽略颜色特征导致检测结果不佳的问题,提出了基于颜色矩的场景文本检测算法 TDSI-C。首先使用改进的 MSER 和 SWT 算法从图像中提取文本候选区域。然后根据图像的几何特征、笔画宽度特征、颜色矩特征将非文本区域过滤掉。最后根据汉字的结构特征将文本区域聚集成单个汉字,再聚集成文本行。实验结果表明,与不使用颜色矩特征的算法相比,TDSI-C 算法获得了较好的准确率、召回率和 F 值。

关键字: 最大稳定极值区域; 笔画宽度变换; 场景文本检测; 汉字结构; 颜色矩

Abstract

With the development of science and technology, images' acquisition is becoming extremely convenient. There are many information in images, such as color, shape, texture, text and so on. The study of the text detection in natural scene images has great practical value in the following fields, such as content-based image retrieval, video analysis, computer aided image system, safety driving assist, intelligent traffic management, and so on. In natural scene images, when the background of images are very complex, the text detection is very hard. This paper will improve the situation. The main work is as follows:

(1) Because of lacking public natural scene image database based on pure Chinese characters, a natural scene image database for pure Chinese characters is built. The images in the database are taken from within and around school by high resolution of smart phone and digital cameras, besides the Internet. There are almost 400 images in total in the database.

(2) Because of the higher false-positive rate of text detection in complex background, presenting a Chinese text detection algorithm in natural scene images (TDSI). Firstly, a series of heuristic rules are used to improve MSER algorithm and SWT algorithm. The standard deviation of stroke width value is used to filter non-text regions. Then Chinese characters structure is used to gather candidate character regions into single Chinese character. The location of candidate characters' center and overlap regions are used to estimate whether two candidate character regions are the same character or not. Experiments show that using self-built image database, TDSI algorithm has higher accuracy rate, recall rate and F value for the natural scene images in more complicated Chinese environment.

(3) Because of the poor results of existing scene text detection algorithm ignoring color features, presenting a new natural scene text detection algorithm for color moment. First of all, candidate character regions are extracted by MSER algorithm. Then the candidate character regions are simulated by SVM algorithm using some features, such as stroke width, color moments, the geometrical characteristics of an image and so on. These features can be better distinguished between text regions and non-text components. Finally, the Chinese character are aggregated by Chinese characters structure, and then the Chinese text lines are aggregated. Experiments show that compared with do not using the color features of the algorithms, TDSI-C algorithm has higher accuracy rate, recall rate and F

value for the natural scene images in more complicated Chinese environment.

Keywords: scene text detection; maximally stable extremal region; stroke width transform; Chinese characters structure; color moment

目 录

摘 要	I
Abstract.....	II
目 录	IV
第一章 引言	1
§ 1.1 场景文本检测的研究背景与意义	1
§ 1.1.1 研究背景	1
§ 1.1.2 研究意义	3
§ 1.2 国内外研究现状	4
§ 1.2.1 文本候选区域提取方法	4
§ 1.2.2 文本候选区域验证方法	7
§ 1.3 场景文本检测存在的问题	9
§ 1.4 研究内容	9
§ 1.5 论文组织结构	10
第二章 自然场景文本检测相关介绍	11
§ 2.1 自然场景图像的特点	11
§ 2.2 公开的自然场景文本数据集	13
§ 2.2.1 ICDAR 竞赛及数据集	13
§ 2.2.2 其他数据集	14
§ 2.3 实验图像数据集构建	15
§ 2.4 文本检测算法评价方法	17
§ 2.5 本章小结	19
第三章 自然场景图像中背景复杂的中文文本检测算法	20
§ 3.1 文本候选区域的提取	20
§ 3.1.1 MSER 算法原理	20
§ 3.1.2 基于启发式规则的 MSER 算法改进	21
§ 3.2 文本候选区域的验证	22
§ 3.2.1 SWT 算法原理	22
§ 3.2.2 基于启发式规则的 SWT 算法改进	23
§ 3.3 基于汉字结构的算法改进	24
§ 3.4 中文文本检测算法及描述	25
§ 3.5 实验结果与分析	26

§ 3.5.1 实验结果	26
§ 3.5.2 结果分析	28
§ 3.6 本章小结	29
第四章 基于颜色矩的自然场景中文文本检测算法	30
§ 4.1 颜色空间	30
§ 4.1.1 RGB 颜色空间	30
§ 4.1.2 CMYK 颜色空间	31
§ 4.1.3 YUV、YCrCb 颜色空间	31
§ 4.1.4 CIE $L^*a^*b^*$ 、CIE $L^*u^*v^*$ 颜色空间	31
§ 4.1.5 HSV (HSB) 颜色空间	31
§ 4.2 颜色特征提取方法	32
§ 4.2.1 颜色直方图	32
§ 4.2.2 颜色熵	33
§ 4.2.3 颜色聚合向量	33
§ 4.2.4 颜色相关图	34
§ 4.2.5 颜色矩	34
§ 4.3 颜色距离计算	35
§ 4.3.1 基于 CIE $L^*a^*b^*$ 颜色空间的色差计算	35
§ 4.3.2 欧式距离	36
§ 4.3.3 基于 HSV 颜色空间的欧式距离	36
§ 4.4 基于颜色矩的文本检测算法	36
§ 4.5 实验结果与分析	39
§ 4.6 本章小结	42
第五章 总结与展望	45
§ 5.1 研究总结	45
§ 5.2 工作展望	46
参考文献	47
致谢	52
攻读硕士学位期间的主要研究成果	53

第一章 引言

§ 1.1 场景文本检测的研究背景与意义

§ 1.1.1 研究背景

随着科技的发展，高分辨率的手机和数码相机的大量使用，图像和视频信息的获取变得异常便捷。如何处理和分析这些图像和视频数据成为一个意义重大而又充满挑战的课题。

图像中包括了很多重要的信息，譬如颜色、形状、纹理和文字等。图像中的文字包括了很多语义信息。对图像中文字信息的检测，在基于内容的图像检索、视频分析、网络不良信息过滤、视觉辅助、汽车辅助驾驶、智能交通管理等领域都有着十分重要的应用。

图像中的文本检测是研究如何从图像中得到文本所在的区域。对图像中的文本信息的检测一般分为三类：文档图像的文本检测、原生数字图像的文本检测和自然场景图像的文本检测^[1]。

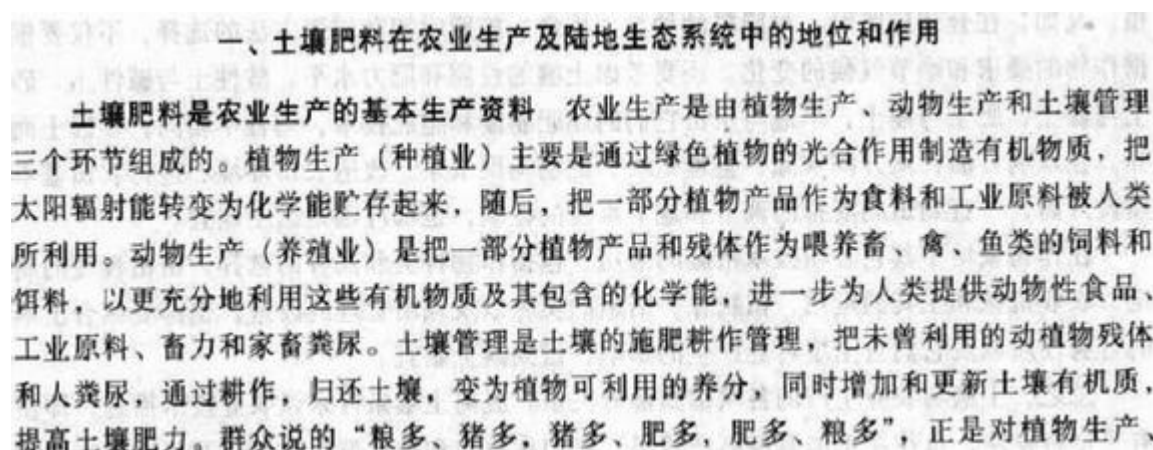


图 1.1 文档图像示例

其中，文档图像是指通过扫描仪等成像工具将书籍、期刊和杂志等扫描成的图像，如图 1.1 所示。文档图像大多是由文本组成的，其中的文本具有高分辨率、高对比度、背景比较干净等特点，检测难度较小。经过国内外各学者数十年的研究，到目前为止，对文档文本的检测技术已经十分成熟，并且广泛运用到人们的日常生活中。现有的 OCR（Optical Character Recognition，光学字符识别）软件对文档文本的识别率已达

到 99%，例如我国自主研发的清华 TH-OCR 和汉王 ORC 在对文档图像的文本检测已经达到了很高的性能，并且能够满足实际的生产和应用需要。

原生数字图像是通过计算机软件合成的图像，并不是根据真实场景拍摄的，这些图像大多存在于网络上，如图 1.2 所示。这些图像具有前景或背景较复杂、压缩损失较多、文字边缘柔化现象较严重等特点。原生数字图像中的大多数文字都是用以辅助人们理解该图像所表现的真实内容，针对原生数字图像中文本的检测对网站内容的获取、分析、检索和过滤有着重要意义。为了推动原生数字图像文本检测技术的发展，满足原生数字图像语义分析、信息过滤等方面的需求，ICDAR(International Conference on Document Analysis and Recognition，文档分析与识别国际会议)会议于 2011 年首次推出了原生数字图像文本提取竞赛。

销售场景演练



图 1.2 原生数字图像示例

自然场景图像指的是通过数码相机、智能手机等电子设备拍摄的真实、自然、未经加工的图像，如图 1.3 所示。原生数字图像是由计算机软件合成的，而自然场景图像表征的是真实世界中存在的事物。自然场景图像中的文字存在的主要目的是提醒、警告和提示等，多用于给人们的生活带来便利。因此真实世界存在的文本一般较为醒目，与背景能较明显的区分开。可是因为在拍摄过程中各类人为因素的影响，使得从自然场景图像中检测出文本的难度大大提升。自然场景图像中的文本一般具有光照不均匀、背景复杂、文字的排列方式无固定模式和文字易出现变形失真等特点，检测难度较大。虽然在场景文本检测方面的研究较多，也涌现了很多优秀的科研成果，但还没有完全解决这些难题。论文将针对背景复杂的场景图像中的文本检测技术做深入研究。



图 1.3 自然场景图像示例

§ 1.1.2 研究意义

场景文本检测技术能将文本从图像中检测出来，由此可以实现计算机自动读取图像中文本的功能，有助于对图像内容的理解。该技术在众多领域得到了广泛应用。

在基于内容的图像检索中，对比图像的其他特征，诸如颜色、纹理和边缘等，图像中的文字相对较直观的反映了图像要表达的信息。检测图像中的文字，可以辅助理解图像所要表达的内容，提升图像检索的准确性。

在视频分析技术中，分析视频数据中每一帧图像的字幕、图像中出现的文字，结合视频中的语音信息，使用多模态分析方法，能更好的理解视频内容，对基于视频的内容提取、分类和分析起到了重要作用。

在视觉辅助技术中，使用视频设备获取周围环境中的文字信息，比如道路提示信息、店铺名称、物品标签和标语等，并将之转换为语音信息，可以帮助有视觉障碍的残疾人丰富其业余生活，为他们的出行和生活提供更多便利。而对图像中的车辆牌照进行检测，可应用于电子收费、车辆监控等方面，大大减少了人力成本。另外，在各大旅游景区，检测并识别道路指示牌、商店名、商品名等信息，将之转换为游客的母语并输出为语音，可以为游客提供便利，改善游客的旅游体验。

在汽车辅助驾驶中，场景文本检测技术可用于识别交通标识牌中的路名、方向和距离等，便于计算机自动调节驾驶速度、驾驶方向和驾驶路线等信息实现自动驾驶。同时，也可以通过计算机自动识别高速公路上的交通标志牌^[1]，提醒高速公路上疲劳驾驶的司机们小心驾驶，降低交通事故发生的频率。

§ 1.2 国内外研究现状

自然场景图像中的文本检测技术一般包含提取文本候选区域和验证文本候选区域两个步骤^[3]。提取文本候选区域是指从图像中提取具有相似特征的候选区域作为文本候选区域。验证文本候选区域则是指按照一系列相关算法验证从上述阶段获得的文本候选区域是否为文本区域。

§ 1.2.1 文本候选区域提取方法

提取文本候选区域的方法多种多样,但其关键思想都是一致的,即通过不同的算法将图像中具有相似特征的区域提取出来,作为文本候选区域。常见的提取文本候选区域的方法包含基于边缘的方法、基于纹理的方法、基于深度学习的方法和基于连通域的方法等。

(1) 基于边缘的方法

一般来说,不同类型的物体具有完全不同的边缘。自然场景图像中的大多数文本信息都具备标识作用,如商号名称、广告标语等。这些文字包括了大量的边缘特征。通过提取图像的边缘特征,可以将图像中有类似边缘的相邻像素点聚集成文本候选区域。

由于边缘是区分文本和背景的一个重要特征, Youbao Tang 等^[4]提出一种新的基于边缘特征和其他混合特征的场景文本检测算法。该算法主要分为两部分:文本候选区域的提取和分类两部分。在提取文本候选区域时,先从源图像中提取边缘,然后根据颜色特征分割边缘图像。

Meng 等^[5]提出了一种两步边缘分析(边缘合并和分类)的自然场景文本提取算法。该方法首先通过图像的边缘检测算法将图像分成若干部分,接着将具有相似笔画宽度的相邻边缘进行合并,即得到的文本区域。

Yu Chong 等^[6]于 2015 年提出了一种改进的笔画宽度变换算法,该算法包括两个阶段的边缘分析,分别是候选区域的边缘重组和边缘分类。该算法将字符和字符串的特征都用于字符串的分类。

2016 年, Yu Chong 等^[7]为了分类更精确,首先提取图像的边缘和区域的特征放入特征池中,然后使用线性 SVM 分类器从特征池中选择更有效的特征来训练分类器。

(2) 基于纹理的方法

由于图像中的文本与图像中的其他背景具有完全不同的纹理特征,通过纹理特征可以粗略地将图像中的文本区域从图像背景中区分出来。

在 Minetto 等^[8]提出的方法中,使用 HOG 描述子作为文本描述和检测的有效工具。

基于纹理的 HOG (T-HOG) 描述子把图像分割成拥有渐进边界的重叠的水平格, 用以刻画自然场景图像中的文本行。

Huang 等^[9]综合使用基于纹理的滑动窗口分类算法和 MSER (Maximally Stable Extremal Region, 最大稳定极值区域) 算法, 来减少 MSER 算法中出现字符与字符、字符与背景间的粘连现象的可能性。

Wibowo 等^[10]使用弹性、曲率和纹理三种特征的算法检测自然场景图像中的文本。其中用到的弹性特征和曲率特征分别指每幅图像文本区域中每个像素的一阶导数和二阶导数, 纹理特征则是文本区域的 4 向链码 (chain code)。

Fabrizio 等^[11]提出了一种综合的多尺度文本检测算法。首先在对图像进行分割之后, 使用改进的小波描述子对其进行分类, 并用新的图形建模方法获取文本候选区域。然后使用基于纹理的方法把非文本区域过滤掉。最后通过图像二值化方法将检测到的文本区域精准的切割出来。

针对自然场景图像中的文本存在字迹模糊、笔画断裂等问题, 尹芳^[12]提出了一种鲁棒性较强的提取 Gabor 小波特征的方法来提取文本区域。首先在原始的 Gabor 小波变换的基础上进行滤波方向的选择分类, 然后使用有方向选择性的小波变换来提取文本候选区域。

Busta 等^[13]提出一种易于使用的新的笔画探测器。该算法改进了传统的 FAST 角点算法, 使其更有利于自然场景图像中文本的检测。该算法检测文本候选区域的速度比 MSER 快 4 倍, 且检测效果比 MSER 好。但当图像对比度低、图像背景复杂时, 文本检测的准确率不高。

(3) 基于深度学习的方法

由于在图像分类中的优异表现, 深度学习近几年已成功应用到了很多图像处理领域, 并获得较好的成绩。近年来基于深度学习的场景文本检测算法也层出不穷:

颜建强^[14]提出了一种基于深度学习的场景文本检测算法。该算法先利用 Gabor 滤波器获取纹理特征, 然后综合运用 DBN (Deep Belief Network, 深度信念网络) 对文本候选区域进行分类。

孙雷^[15]提出了一种基于颜色增强的 CER (Contrasting Extremal Region) 和浅层神经网络的全自动文字检测系统, 在特征层直接利用神经网络从文本候选区域对应的二值图像中学习有用的特征。该方法避免了人工提取特征的过程中的信息损失, 同时降低了算法的计算量。

Jianqi^[16]等提出一种旋转区域建议网络 (Rotation Region Proposal Networks), 可以生成文本倾斜角度信息。该角度信息通过适应边界框回归使文本区域检测更准确。该方法可以对自然场景图像中任意方向的文本行进行检测。

Liao M^[17]等提出了一种端对端的可训练的快速场景文本检测器，在单层网络中就可以准确而有效的检测自然场景图像中的文本。

张树业^[18]提出了一种在级联分类框架下的相似字符挖掘算法，构建了基于自然场景文本检测的字符候选区域网络，还设计并改进了场景文本识别过程中的特征表达和学习算法。

基于深度学习的方法虽然能较准确的检测场景文本，但其需要的硬件要求较高，普通计算机仍无法实现。

（4）基于连通域的方法

基于连通域的方法是指将提取到的图像中的连通域（Connected Components）作为文本候选区域。通过各类算法提取到的连通域和图像中的文本区域一样，都具有相同或相似的图像特征，因此提取到的连通域可以作为文本候选区域。常用的连通域提取算法包括基于区域生长的算法和基于 MSER 的算法。

区域生长算法是最经典的图像分割算法之一，Basavanna 等^[19]提出一种基于自适应直方图分析的滑动窗口方法来进行文本直方图分析。该方法使用区域生长的方法，根据单词间的空格来分割字符，但算法准确率不高。

MSER 算法最早由 Matas 等^[20]于 2002 年提出，是目前场景文本检测技术中应用最广泛的算法之一。在 Nister 等^[21]提出的改进的 MSER 算法中，其时间复杂度接近线性 $O(n)$ 。2011 年 ICDAR 大赛组委会将其列为最有希望的文本检测基础算法。在自然场景图像中，文字大多具有较明显的对比度和统一的颜色，这正符合 MSER 算法的要求。很多算法^[22-25]都选择 MSER 算法作为场景文本检测中进行图像二值化的方法，也即文本候选区域的提取方法。

Chen 等^[26]提出了一种新的文本检测算法，该算法使用 MSER 做预处理以改进 SWT 的性能。该算法结合了视觉搜索系统，能较精确的提取最大稳定极值区域。但当图像背景较复杂时，其文字检测准确率不高。

Liu 等^[25]提出一种多粒度的锐化模糊方法来检测场景文本。首先对输入图像使用 USM（锐化算法）和双边滤波算法对图像进行锐化模糊。然后使用 MSER 算法从原始图像和处理过的图像中提取连通域，通过距离函数对图像特征进行加权来构建文本候选区域。最后，通过一个字符选择器来估计文本候选区域，并将非文本候选区域过滤掉。

§ 1.2.2 文本候选区域验证方法

在自然场景文本检测算法中，当提取到文本候选区域后，需对其进行验证，确认其是否为文本区域。验证文本候选区域的方法主要包括基于分类器的方法和基于特征的方法。

（1）基于分类器的方法

验证文本候选区域的实质是一个二分类问题，即将文本候选区域分为文本区域和非文本区域两类。基于分类器的方法即通过分类器将文本候选区域中的非文本区域去掉，留下文本区域。

Qingqing Wang 等^[27]提出一种包含两个步骤的非文本过滤的场景文本检测算法。首先使用 MSER 算法得到文本候选区域，然后综合使用随机森林(Random Forest, RF)和条件随机场(Conditional Random Fields, CRF)过滤非文本区域。最后，通过边缘削减算法将被标注为文本的区域聚集成文本行，并且用基于 HOG 特征的分类器计算准确率。

Xuejian Rong 等^[28]提出了一个新的场景文本检测框架，包括用于字符预测的场景文本字符的特征表示模型和用于单词配置的 CRF 模型。然后使用密度采样的 SIFT 描述符和费舍尔向量提取自然场景图像中的字符特征。最后还提出了一个获取文本信息的自然场景视频数据集。

Yanna Wang 等^[29]提出一种针对文本图像的二值化马尔科夫随机场(Markov Random Field, MRF)框架。该框架融合颜色、纹理、上下文特征来实现图像二值化。算法的主要创新点包括：将图像分割成不同的子图，使用笔画特征自动获取种子像素的前景和背景；根据位置信息，不同的权重和种子像素可以通过聚集加权种子获得高度信任聚类中心的子图像。

Xiaolong Liu 等^[30]提出一种新方法：使用原生数字模板匹配来检测场景字符，使用混合的局部特征和场景字符的全局空间结构的 MRF 框架来提高检测准确率。基于 MRF 框架的字符检测方法的最终结果主要通过混合的一对一分类器投票决定。

Wen-Hung Liao 等^[31]提出了一种中英文环境下的场景文本检测算法。该算法主要包括四个阶段：（1）使用双边滤波器对场景图像进行预处理；（2）使用 Canny 边缘算子提取候选文本边缘，使用 MSER 算法提取文本候选区域，然后合并这两种特征以获得更好的结果；（3）连接候选字符：考虑文本的水平方向和垂直方向，使用几何特征将候选字符聚集成候选文本；（4）最后使用支持向量机(Support Vector Machine, SVM)分类算法对文本区域和非文本区域进行区分。

Yajie Chen 等^[32]提出一种检测和识别都市场景下的交通标志牌的算法。该算法使用面向梯度的直方图和线性 SVM 算法检测交通标志牌。交通标志牌上的文本和标志通过连通域分析的方法做分割。最后使用视觉词袋模型检测交通标志牌上的标志。

Jiakai Gao 等^[33]提出了一种新的基于颜色还原和 Adaboost 分类器的文本定位方法, 该方法适用于提取背景复杂的自然场景文本图像中的文本区域。

(2) 基于特征的方法

文本区域和非文本区域通常具有完全不同的特征, 根据该特点很多算法使用各种不同的图像特征来验证文本候选区域是否为文本区域。

Houssem Turki 等^[34]首先使用集中领域的方法提取文本候选区域, 该方法使用遮罩的方法过滤复杂背景, 即使用不同尺度的 Otsu 算法和金字塔模型的边缘增强算法。其次使用基于亮度的 MSER 算法检测字符候选区域。最后根据 SIFT 特征和 HOG (Histogram of Oriented Gradient, 方向梯度直方图) 特征的 DWT (Discrete wavelet transform, 离散小波变换) 算法将非文本区域过滤掉。

Jufeng Liu 等^[35]提出了一种在不同天气状况下检测都市道路交通牌上的交通标语的算法。首先使用颜色分割算法和形态学方法提取文本候选区域, 主要关注文本候选区域的轮廓。然后提取 EOH (Edge Orientation Histogram, 边缘方向直方图) 的形状特征, 使用线性 SVM 分类器对文本候选区域进行分类, 将非文本区域过滤掉。

Jiakai Gao 等^[33]从文本候选区域中提取 HOG 和 LBP (Local Binary Pattern, 局部二值模式) 特征, 然后使用 Adaboost 分类器来区别文本区域和非文本区域。

Zhang Yang 等^[36]使用 RF 与 SVM 分类器对文字候选区域进行验证, 并对比了实验结果, 认为 RF 分类器结合 EOH 与 MSLBP 特征可较好的提高检测文字区域的准确性。

图像中的笔画宽度特征是文本特有的特征, 不论是中文、英文还是其他语言, 都具有笔画宽度特征, 而其他物体比如树叶、栅栏、人物等却没有该特征。所以笔画宽度特征能较好的将文本区域和非文本区域区分开。

Huang Rong 等^[37]提出了一种新的基于边缘的场景文本检测算法, 称为边缘射线滤波器, 用来检测场景文本中的字符。该算法对图像对应的边缘图计算笔画宽度时, 充分利用图像的边缘特征, 而不是使用简单的空间分布来筛选文本区域。

Xu 等^[38]提出了一种新的文本检测算法, 主要包含两部分: 基于鲁棒的扩展 SWT (Stroke Width Transform, 笔画宽度变换) 算法和文本对象判别的深度信任网络。前者使用平滑边缘信息和梯度信息获取高质量的边缘图像。在 SWT 算法的基础上, 通过分析连通域来消除帧间字符和字符帧内的错误, 得到边缘信息。对于后者, 用深度信任网络学习如何有效识别字符和非字符候选区域, 从而提高检测精度。

Zhang 等^[39]提出了一种新算法，首先将 SWT 算法得到的结果作为基本的笔画候选区域。然后使用自适应的结构元素生成紧凑的字符，合并这些文本候选区域。最后使用 K 最近邻算法把单个字符组合成字符串，从而确定任意方向的字符串。

目前，大多自然场景文本检测算法都使用上述的两阶段框架，但还有部分学者尝试使用端到端的场景文本检测算法^[40-42]，这种方法在场景文本检测的基础上，添加了字符检测，实时的检测文本并将检测到的文本通过 OCR 软件显示出来。

§ 1.3 场景文本检测存在的问题

综上所述，虽然很多算法对场景文本检测技术进行了深入研究，但该领域还有很多具有挑战性的问题，比如：

(1) 缺乏纯中文文本的实验图像数据集。现有公开的场景图像数据集大多是基于英文文本或中英文混合文本的，暂时还没有纯中文文本的公开的场景图像数据集。

(2) 自然场景图像中的背景复杂，导致场景文本检测算法不容易将背景和文本区分开。背景中常包含诸如栅栏、树木、人物等干扰信息，这些干扰信息将会影响算法对文本的检测结果，使得目前文本检测算法的检测效果不理想。

(3) 缺乏针对中文文本的场景文本检测算法。现有的算法大多针对场景图像中的英文文本或中英文混合文本进行检测。而汉字结构的复杂性和多样性，导致对中文文本的检测与对英文文本的检测有很大的差异。使用专门针对英文文本进行检测的算法来检测图像中的中文文本，检测结果较差。

(4) 颜色特征对场景文本检测算法的结果影响较大，但大部分文本检测算法没有考虑这点。由于多半图像处理算法都要求待处理的图像是灰度图像，一般算法在对图像进行处理前会将原始图像转换为灰度图像。然而自然场景图像中的文本一般具有相同或相近的颜色值，如果完全忽略颜色特征，会导致场景文本检测算法的检测结果变差。

§ 1.4 研究内容

针对上述提出的问题，论文主要做了以下几方面研究：

(1) 研究构建纯中文文本的实验图像数据集。构建图像数据集时，既要寻找符合条件的图像，还要为数据集中的每幅图像做恰当的标注，以便后期实验的使用。论文将研究哪些图像适合放入实验用的数据集，以及如何快速有效的对每幅图像做标注。

(2) 研究自然场景图像中复杂背景下的中文文本检测算法。研究如何通过改进 MSER 算法和 SWT 算法来提高文本检测精度, 场景文本检测算法对中文和对英文文本检测结果不同的原因, 以及汉字的结构特征对场景文本检测算法的影响。

(3) 研究图像颜色特征对场景文本检测结果的影响。重点研究颜色矩特征是如何对检测结果产生影响的, 以及如何衡量颜色矩特征对检测结果的影响力大小。

§ 1.5 论文组织结构

全文共分为五章, 论文组织结构安排如下:

第一章, 绪论。大致介绍了场景文本检测技术的研究背景和意义、国内外研究现状、场景文本检测存在的问题、研究内容及文章组织结构。

第二章, 自然场景文本检测相关介绍。首先介绍自然场景图像中的文本所具有的一般特点。然后介绍 ICDAR 竞赛和相应的数据集、国内外其他几个公开的自然场景图像数据集、论文使用的实验图像数据集的构建。最后详细介绍文本检测算法常见的评价标准和论文采用的评价标准。

第三章, 背景复杂的自然场景图像中的中文文本检测算法。在现有的场景文本检测算法的基础上提出了一种基于复杂背景的自然场景中的中文文本检测算法 TDSI (Text Detection Algorithm in Natural Scene Images)。首先介绍 MSER 算法的基本概念和对 MSER 算法的改进, 然后介绍 SWT 算法的基本概念和对 SWT 算法的改进, 接着针对汉字结构的改进做了详细介绍, 其次介绍 TDSI 算法的整体流程, 最后对实验结果进行分析总结。

第四章, 基于颜色矩的自然场景图像中的中文文本检测算法。针对现有算法忽略颜色特征导致算法结果较差的问题, 提出了一种基于颜色特征的场景文本检测算法 TDSI-C。首先分别对常见的颜色空间、常用的颜色特征提取方法、颜色距离计算做了详细介绍, 然后介绍了基于颜色矩的算法改进, 其次介绍该算法的整体流程, 最后对实验结果做了详细介绍和分析。

第五章, 总结与展望。对论文做了简要的回顾, 总结了论文提出的两个算法和所做的相关改进, 并指出下一步需要研究的工作。

第二章 自然场景文本检测相关介绍

§ 2.1 自然场景图像的特点

文档图像背景简单，背景颜色一般为白色，文字颜色一般为黑色，由于文档图像相对较简单的结构，因此较容易进行二值化处理。相比文档图像，自然场景图像的背景较为复杂，结构多变，具有完全不同的颜色、亮度和对比度等信息。另外由于拍摄角度的不同，还可能导致图像发生扭曲变形等问题，因此自然场景图像往往不易进行二值化。自然场景图像主要包括以下特征：

（1）自然场景图像背景复杂

文档图像的背景一般都是白色，或者有一些较简单的修饰。其背景较简单，颜色一般只有一到两种。而在用手机或相机拍摄的自然场景图像中，背景复杂，可能包括树叶、草坪、栅栏、建筑等各种干扰信息。背景色彩斑斓，一幅图像中可能包括赤橙黄绿多种颜色。如图 2.1 所示。



图 2.1 背景较复杂的图像示例

（2）自然场景图像光照不均匀

文档图像一般通过扫描仪等仪器获取，这些仪器的光源光照分布相对较均匀，不易造成光照不均匀的现象。而在拍摄自然场景图像的过程中，如果在室内的人造光源下或在相机使用闪光灯的情况下，拍摄出的自然场景图像可能出现光照不均匀的现象，即拍摄的图像中有的部分光照较强，导致该部分图像曝光过度，无法看清图像内容，而有的部分光线较暗，导致无法清晰的观测到图像内容，如图 2.2。



图 2.2 光照不均匀的图像示例

(3) 自然场景图像具有完全不同的布局框架

文档图像主要用以传播文字信息，为了迎合人们的阅读习惯，一般将文本整齐划一的归类到规则的矩形区域中。而自然场景图像的主体是各不相同的景色和物体，其摆放排列完全没有秩序可言，不同的自然场景图像一般都具有完全不同的布局框架，如图 2.3 所示。即使是完全一样的场景，如果拍摄的角度、光照、天气状况不相同，拍摄出的图像也不完全相同。



图 2.3 具有特殊文本布局的图像示例

(4) 自然场景图像易发生变形失真

文档图像一般是将文档放在扫描仪中进行扫描，文档离镜头的距离是相同的，也就不会出现变形失真。但是在拍摄自然场景图像的时候，如果拍摄的角度不合适或者

不正确，将会导致距离镜头近的物体较大而距离镜头远的物体较小，如图 2.4 所示。如果拍摄时相机摆放时有倾斜角度，将会导致拍摄的图像出现倾斜的问题。拍摄时人的抖动还会导致图像出现大小不一的扭曲、变形和模糊等现象。上述这些问题都可能导致拍摄的自然场景图像发生变形失真问题，为场景文本检测带来困难。



图 2.4 存在变形失真的图像示例

§ 2.2 公开的自然场景文本数据集

§ 2.2.1 ICDAR 竞赛及数据集

ICDAR 文档分析与识别国际会议从 1991 年开始举办。该会议自 2003 年起，每届都会举办相应的竞赛。这些竞赛为各学者提供了一个展示各自算法和相互交流心得的平台，使得文本检测技术得到了突飞猛进的发展。商用的 OCR 软件在文档图像的检测上已经取得了高速度、高识别率的有效成果，但对于通过数码相机等设备拍摄的自然场景图像中的文本检测，暂时还无法得到令人满意的结果。

ICDAR 大赛组委会于 2003 年起开始举办鲁棒阅读竞赛（Robust Reading Competition），该竞赛公布了所用的自然场景图像数据集。该届竞赛将自然场景图像中对文本的鲁棒性识别按照处理过程分为文本定位、字符识别和单词识别三个不同的子阶段，并针对这三个阶段分别举行竞赛。



图 2.5 ICDAR 数据集图像示例

直至 2017 年, 该竞赛已包括了手写体历史文献的布局检测 (Handwritten Historical Document Layout Recognition)、手写体历史脚本分析 (Historical Handwritten Script Analysis)、字符/单词识别 (Character/Word Spotting)、手写体识别 (Handwriting Recognition)、文档图像二值化 (Document Image Binarization)、文档重建 (Document Reconstruction) 和鲁棒阅读竞赛 (Robust Reading Competitions) 等 12 项竞赛, 同时还提供了相应的图像数据集。这些数据集中收录的图像都是 24 位彩色图像, 包括了路牌、广告、商品名称、牌匾等不同类型的图像, 如图 2.5 所示。图像中包含的所有文本都是英文或阿拉伯数字, 而没有中文。

§ 2.2.2 其他数据集

除了 ICDAR 提供的自然场景图像数据集外, 还有其他公开的自然场景图像数据集可供使用:

(1) COCO-Text 数据集。2016 年, 由 Veit 等^[43]提出的自然场景图像数据集, 一共包括 63686 幅自然场景图像、173589 个文本实例和三种细粒度的文本属性。文本实例主要包括手写版和打印版、清晰版和非清晰版、英文版和非英文版等不同版本。

(2) Synthetic Data for Text Localization in Natural Image (VGG)数据集。该数据集由 Gupta 等^[44]于 2016 年建立。该数据集中存放的都是复杂背景下人工合成的场景图像数据,主要用于场景文本定位。这些图像可以直接指定文字的标签信息以及文本在图像中的位置,从而省去了人工标注的过程。

(3) Synthetic Word Dataset 数据集。这是 Jaderberg 等^[45,46]于 2014 年创建的数据集。该数据集包括九百万幅涵盖了九万个英文单词的自然场景图像,主要适用于文本检测和文本分割。

(4) IIIT 5K-Words 数据集。这是 Mishra 等^[47]于 2012 年建立的数据集。该数据集中的图像都来自谷歌图像搜索,一共包括 5000 幅左右。这些图像主要包括了自然场景图像和原生数字图像。图像中包括广告牌、招牌、房屋号码、房屋铭牌和电影海报等内容。数据集分为训练集和测试集,该数据集还提供了 50 多万个词典单词。

(5) MSRA Text Detection 500 Database(MSRA-TD500)数据集。该数据集是由 Yao^[48]于 2012 年建立的。该数据集中的自然场景图像主要涉及不同尺度、颜色和语言种类的文本,同时还包括了大量发生仿射、投影和旋转变换的文本。数据集包括了 500 幅室内和室外的自然场景图像,图像分辨率范围在 1296*864 到 1920*1280 之间。

(6) Street View Text (SVT)数据集^[49]。该数据集中的所有图像都来源于谷歌街景 (Google Street View),图像分辨率较低,文字变化较大。图像都是 24 位彩色图像,像素大小是 1260*860,其存储格式为 JPEG。该数据集一共包括 350 幅自然场景图像,其中 100 幅用于训练,250 幅用于测试。每幅图像都用一个文本向量表示其中包含的文字信息。

(7) KAIST Scene_Text Database 数据集^[50]。该数据集包括 3000 幅室内和室外的自然场景图像,图像中包含了韩语、英语(数字)和混合(韩语、英语、数字)的语言。

(8) Chars74k 数据集^[51]。该数据集是一个经典的字符识别数据集,各类样本数量相对较均衡,主要包括英文字符和坎那达语(Kannada)字符。训练图像和测试图像各 920 幅,其中包含了自然场景图像和人工合成的图像。

(9) Oriented Scene Text Database(OSTD)数据集^[52]。该数据集收集的自然场景图像较少,只有 89 幅,但其中包括了排列方向各不相同的文字。

(10) USTB-SV1K 数据集^[53]。该数据集中的图像都来自谷歌街景。主要包括了 1000 幅多方向、多视角的自然场景图像。

§ 2.3 实验图像数据集构建

上节介绍的公开的自然场景图像数据集大多数是基于英文环境的,还有少数是基

于中英文混合环境的，但缺乏完全基于中文环境的自然场景图像数据集。刘晓佩^[54]建立了基于中文环境的自然场景图像数据集，但是还没有将之公开。为测试论文算法的性能，使用分辨率较高的智能手机和数码相机拍摄学校内和学校附近的自然场景图像，并加入部分网上找到的合适的自然场景图像，构建了纯中文文本的实验图像数据集。大多数图像是在光照较好的情况下拍摄的，也有部分图像是在晚上光照不好的时候拍摄的。该实验图像数据集中图像的文本内容主要涉及路标、交通警示语、标语、横幅和商店名称等，一共约 403 幅图像。这些图像分辨率大小不一，包括 3264*2448、750*562、640*360 等。数据集中大多数图像背景复杂，具有完全不同的颜色、字体、字号、光照、对比度等，很适合测试论文算法。数据集中的每幅图像都包括了至少一条文本行数据。数据集中的部分图像如图 2.6 所示。



图 2.6 实验图像数据集中的图像示例

根据 ICDAR2013 竞赛^[55]的要求, 为实验图像数据集中的每幅图像添加标注, 便于后期的实验处理与分析。每幅图像的标注内容主要包括图像的编号和能够将图像中的文本区域完全框起来的最小矩形的位置坐标, 形如“图像编号 矩形最左上角像素点的坐标的 X 值 矩形最左上角像素点的坐标的 Y 值 矩形最右下角像素点的坐标的 X 值 矩形最右下角像素点的坐标的 Y 值”。由于手动标注图像的工作量较大, 使用 C# 语言编写软件代替人工标注, 实现半自动化的图像标注, 提高了标注的效率。数据集中每幅图像的文本区域的标注信息如图 2.7 所示。

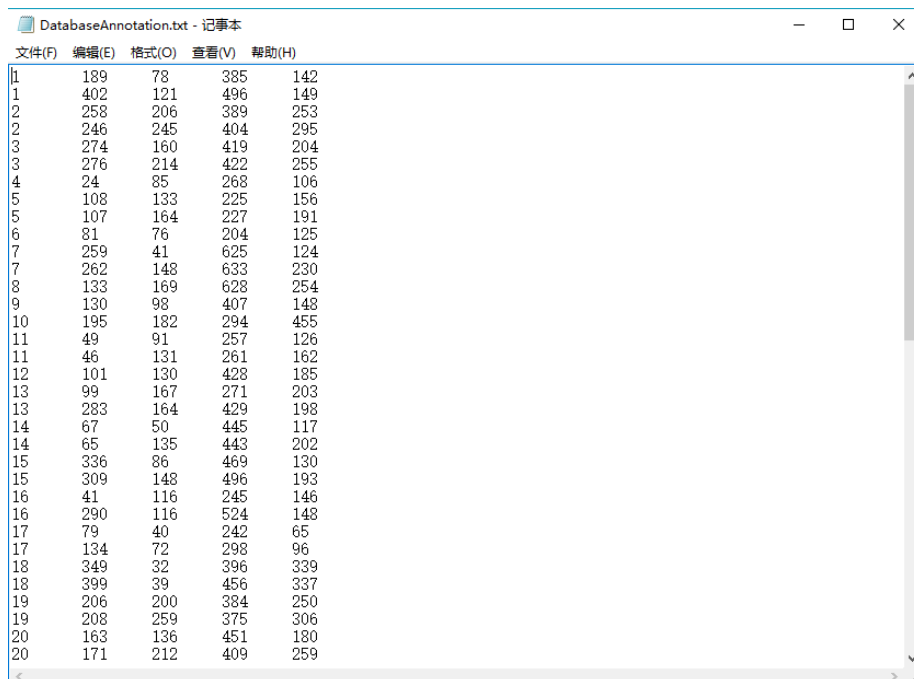


Image ID	X1	Y1	X2	Y2
11	189	78	385	142
1	402	121	496	149
2	258	206	389	253
2	246	245	404	295
3	274	160	419	204
3	276	214	422	255
4	24	85	268	106
5	108	133	225	156
5	107	164	227	191
6	81	76	204	125
7	259	41	625	124
7	262	148	633	230
8	133	169	628	254
9	130	98	407	148
10	195	182	294	455
11	49	91	257	126
11	46	131	261	162
12	101	130	428	185
13	99	167	271	203
13	283	164	429	198
14	67	50	445	117
14	65	135	443	202
15	336	86	469	130
15	309	148	496	193
16	41	116	245	146
16	290	116	524	148
17	79	40	242	65
17	134	72	298	96
18	349	32	396	339
18	399	39	456	337
19	206	200	384	250
19	208	259	375	306
20	163	136	451	180
20	171	212	409	259

图 2.7 实验图像数据集文本区域标注示例

§ 2.4 文本检测算法评价方法

对场景文本检测算法优劣的评估, 关键在于检测或识别到的文本是否正确和全面。为保证对检测算法评价的客观性和准确性, 综合考虑了场景文本和检测算法的特点, 近年来主要形成了以下几种常见的评价方法:

(1) 基于像素匹配的方法。该方法将图像中的像素点分为文本和非文本像素两大类。通过对比图像中全部的文本像素的数目和位置坐标, 来评价场景文本检测算法的好坏。

(2) 基于文字检测结果匹配的方法。用 OCR 软件验证从自然场景图像中检测出的文字, 与文本的标注结果进行匹配, 来评价算法性能的优劣。

(3) 基于矩形框匹配的方法。将从图像中检测到的每一个文本区域用其最小外接矩形来表征, 通过计算矩形块的重合情况, 来评价算法的性能。

基于矩形框匹配的方法是目下多数研究采用的方法，且该方法计算量较小，能较准确的评价算法的性能。ICDAR 文本定位竞赛的评价标准就是使用基于矩形框的方法。该评价标准使用准确率、召回率和 F 值对实验结果进行评价。准确率和召回率的定义分别如公式 (2.1)、(2.2)：

$$P' = \frac{\sum_{r_E \in E} m(r_E, T)}{|E|} \quad (2.1)$$

$$R' = \frac{\sum_{r_T \in T} m(r_T, E)}{|T|} \quad (2.2)$$

$$m_p(r_1, r_2) = \frac{a(r_1) \cap a(r_2)}{a(r_1) \cup a(r_2)} \quad (2.3)$$

$$m(r, R) = \max(m_{p(r, r')} \mid r' \in R) \quad (2.4)$$

其中， $a(r)$ 表示区域 r 的面积， E 指定位到的文本区域的集合， T 表示图像中原本存在的文本区域的集合， $|\cdot|$ 代表集合内的元素总数。 P' 表示准确率， P' 值越高，表明算法定位到正确的文本区域越多，误检率越低。 R' 表示召回率， R' 值越高，表明算法定位到正确的文本区域越多，漏检率越低。

如图 2.4 所示，用黑色标出的区域是 $a(r_1) \cap a(r_2)$ ，用虚线框出的区域是 $a(r_1) \cup a(r_2)$ 。 $m_p(r_1, r_2)$ 是指矩形框 r_1 和 r_2 的匹配面积，即 r_1, r_2 的重叠面积除以能框住二者的最小面积，也即图 2.8 中黑色区域的面积除以虚线框出的面积。 $m(r, R)$ 表示对待检测的图像中的矩形框，在标注集中都有一个可以与之匹配的面积最大的矩形框。

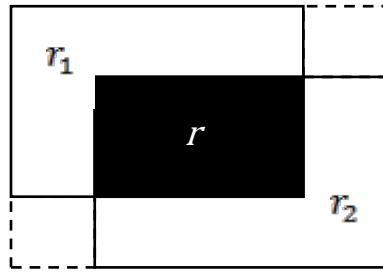


图 2.8 两个矩形框的匹配面积示意图

综合评价结果 F 值如公式 (2.5)：

$$F = \frac{1}{\frac{\alpha}{P'} + \frac{1-\alpha}{R'}} \quad (2.5)$$

在文本检测算法中, 如果其准确率较高, 其召回率不一定也高。只使用准确率或召回率都不能全面的反映算法的精度, 而 F 值兼顾了二者, 因此最终结果使用准确率、召回率和 F 值三者对实验结果进行综合评估比较可靠。其中, 一般取 $\alpha = 0.5$ 。

§ 2.5 本章小结

本章详细介绍了自然场景文本检测技术的相关知识, 主要对自然场景图像中的文本特点、国内外自然场景文本数据集、如何构建基于中文的自然场景图像数据集和文本检测算法的评价方法等几个方面分别展开详细的介绍, 为后续章节的研究奠定了基础。

第三章 自然场景图像中背景复杂的中文文本检测算法

§ 3.1 文本候选区域的提取

§ 3.1.1 MSER 算法原理

MSER 算法是当前最好的区域检测算法之一。由于在自然场景图像中的文字一般都具有相对较明显的对比度和统一的颜色，而且 MSER 算法的计算复杂度接近线性复杂度，很多研究都选择其作为场景文本检测中候选区域的提取算法。

MSER 算法是根据分水岭的概念提出的。分水岭 (water shed) 是地形学上的一个术语，指分隔相邻两个流域的山岭或高地。在图像处理时，可以把灰度图像当作一个起伏变化的地形图：地形图中的位置对应着图像中每个像素点的坐标，而地形图中的高度对应着每个像素点的灰度值大小。如图 3.1 所示，表示一个地形缓慢被水淹没的过程。最初，地形上完全没有水，然后开始向地形中不断注水，地形中低洼的地方开始缓慢出现积水。随着水流的不断注入，低洼地方的水面不断上升，直到到达一定高度，即分水岭处，水将流入其相邻的低洼地点。这样就逐渐形成了一个一个小水坑。随着水流的不断注入，相邻的小水坑逐渐合并成大水坑，相邻的大水坑又合并成池塘，最后整个地形全部被水淹没。在这个过程中，细微的水量变化会导致连通水域的水面面积的急剧变化。根据这个现象，出现了分水岭算法。该算法主要应用于图像分割技术，其主要关注的是区域合并时的水位，而此时的水面面积是不断变化的。MSER 算法在原理上跟分水岭算法是相同的，但二者的关注点却完全不同。MSER 算法关注的是稳定的区域，即该算法希望寻找的是当注入的水量极小时，水面面积的变化率最小的区域，此时水面的面积是近似稳定不变的。

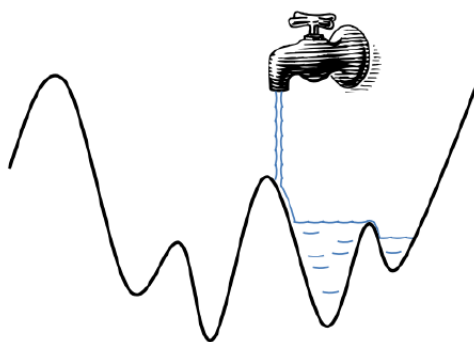


图 3.1 MSER 算法漫水过程示意图

MSER 算法是一种对灰度图像取阈值做二值化处理的算法。在运用该算法前，须

先将原图转为灰度图像。该算法使用图像 I 所有可能的值，包括了 0-255。把图像中小于阈值的灰度值定义为黑色，大于或者等于阈值的定义为白色。如图 3.2 为极值区域的形成过程，第一幅为原始图像，其他为随着阈值的变化得到的极值区域的结果图。当阈值从 0 开始发生变化时，像素点的灰度值都大于阈值，首先得到的是一幅白色的图，即图 3.2 中的第二幅图像。随着阈值的不断递增，一些局部最小值会渐渐大于给定的阈值，会慢慢变成黑色，并且其面积会随着给定阈值的增大而不断变大。当阈值等于某个值时，两个局部最小区域会合并成一个较大的区域。最后，当阈值增大到最大值 255 时，图像中所有像素点的灰度值都小于阈值，整幅图像变成黑色，即图 3.2 中的最后一幅图像。



图 3.2 极值区域的形成过程

在阈值从 0 到 255 不断变化的过程中，当阈值的变化量极小时，若区域面积没有出现比较明显的变化，则该区域为最大稳定极值区域。MSER 算法作为当前最好的区域检测算法之一，具有以下特点：

- (1) 仿射不变性：对图像每个像素点的灰度值具有仿射变换的不变性；
- (2) 稳定性：只有在一定阈值范围内保持不变的区域才会被 MSER 算法选中；
- (3) 多尺度检测：不用做平滑处理就可以实现图像多尺度的检测过程，即较大的和较小的结构都能够被该算法检测出来；
- (4) 时间复杂度较低：在 2008 年 Nister 等提出的改进算法中^[21]，MSER 算法的时间复杂度接近线性 $O(n)$ ， n 表示图像中的像素点总数。

§ 3.1.2 基于启发式规则的 MSER 算法改进

通过 MSER 算法得到的最大稳定极值区域是一些不规则的图形，不方便提取图像特征。一个文本候选区域的特征主要包括位置、长宽和质心等，通过对 MSERs 进行椭圆拟合^[56]，可以较易地得到这些特征。通过 MSER 算法提取到的最大稳定极值区域中既包括文本区域，也包括非文本区域，对椭圆拟合后的 MSERs 使用启发式规则可以将部分明显的非文本区域过滤掉，减轻后续算法的计算负担。TDSI 算法使用的基于 MSER 的启发式规则包括：

(1) 候选区域面积

候选区域中面积特别小的一般不是文本区域，需对其进行过滤。当将候选区域的面积阈值定为 20 时，结果最优，如公式 (3.1)：

$$ResultMSER1 = \{MSER_i | AreaMSER_i > 20\} \quad (3.1)$$

(2) 椭圆拟合后的长宽比

汉字笔画有的短粗、有的细长，如果拟合后的椭圆特别细，近似一条直线，说明该区域一定不是文本区域，需要将区域的长宽比大于某个阈值的区域过滤掉。当将阈值定为 5 时，结果最优，如公式 (3.2)：

$$ResultMSER2 = \left\{ ResultMSER1_i \mid \frac{LongAxis_i}{ShortAxis_i} \leq 5 \right\} \quad (3.2)$$

其中，长宽比是指拟合后的椭圆的长轴与短轴的比值。 $LongAxis_i$ 是拟合椭圆长轴的长度， $ShortAxis_i$ 是短轴的长度， i 表示 MSERs 的个数。

(3) 拟合椭圆与 MSERs 的面积比

拟合椭圆是对 MSERs 的拟合，其面积与 MSERs 存在一定差异。如果 MSERs 是非文本区域比如树叶，MSErs 的面积与拟合椭圆面积差异不大。相反，如果 MSERs 是文本区域，其面积与拟合椭圆面积差异较大。根据该规则，将拟合椭圆与 MSERs 的面积之比太小的区域去掉。当阈值取 1.35 时，结果最优，如公式 (3.3)：

$$ResultMSER3 = \{ResultMSER2_i \mid \frac{AreaEllipse_i}{AreaMSER_i} \geq 1.35\} \quad (3.3)$$

其中， $AreaEllipse_i$ 是拟合椭圆的面积， $AreaMSER_i$ 是 MSERs 的面积。

(4) 图像边界像素交集

场景图像中的文本区域一般不会出现在图像的边界位置，因此将含有图像边界像素的 MSERs 过滤掉，如公式 (3.4)：

$$ResultMSER4 = \{ResultMSER3_i \mid ResultMSER3_i \cap edge = \emptyset\} \quad (3.4)$$

其中， $edge$ 是图像的边界像素。

§ 3.2 文本候选区域的验证

§ 3.2.1 SWT 算法原理

Boris Epshtein 等^[57]依据局部区域的文本，尤其是相邻的文本常会具有相同或相似的笔画宽度值，提出了笔画宽度变换算法。最近 SWT 算法在场景文本检测算法中得到普遍使用，在此基础上的改进算法也屡见不鲜。

Zhang 等^[58]提出了一种基于角点和 SWT 算法的文本检测方法。该算法先在不同尺度的视频中做角点检测，生成文本候选区域。然后利用笔画宽度等特征对非文本区域进行过滤。由于自然场景图像中的干扰信息较多，且对仿射变换较敏感，导致 SWT 算法效果不好。Oh I S 等^[59]提出一种三维空间搜索最优方案的 SWT 算法。该算法借用了深度图像重建领域的搜索方法对 SWT 算法做改进。针对 SWT 算法对图像边缘较敏感的问题，Su F 等^[60]提出了一种基于种子变换的 SWT 算法，显著提高了原始 SWT 算法的鲁棒性。该算法首先搜索笔画种子，即一系列连续的、满足一定约束条件的相邻射线（双边缘像素），生成更多的笔画片段。然后利用笔画的宽度和方向信息来检测笔画宽度，并将部分由于错误和无用边缘导致丢失的笔画进行恢复。但该算法的计算效率不高。所以，TDSI 算法将在 SWT 算法的基础上添加启发式规则对其作出改进。

SWT 算法最大的优势在于，笔画宽度特征是文本的一个固有特征，无论是中文、英文还是其他语言，都具有笔画宽度这一特征。而文本的其他一些特征，比如连通区域、边缘特征、纹理特征等，除了文字，树叶、窗帘等也都包含这些特征。使用该特征能较准确的区分文本区域和非文本区域，是以该特征很适合用于文本候选区域的验证。

笔画是指图像中的连续区域，具有几乎不变的宽度和平行的边缘。字符的笔画宽度是指垂直于字符边缘的线段的长度。笔画宽度变换是一个局部图像算子，输出图像的尺寸与输入图像相同，但是输出图像的每个像素由原来表示灰度值变成了表示笔画宽度值^[61]。

SWT 算法的关键是如何计算每个像素点对应的笔画宽度值，Epshtein 等^[62]提供了计算笔画宽度值的具体过程，其算法描述如 3.1 所示：

算法 3.1:

算法输入：图像 I

算法输出：图像 I 对应的笔画宽度图像

- 1: 将原始图像转换为灰度图像；
 - 2: 初始化笔画宽度图像，其大小与输入图像一致，将其中每个像素点的灰度值初始化为无穷大；
 - 2: 使用 Canny 算子计算灰度图像的 Canny 边缘；
 - 3: 计算灰度图像的梯度方向；
 - 4: 提取灰度图像的边缘点，沿该点的梯度方向搜索方向与之相反的另一个边缘点；
 - 5: 如果找到符合条件的边缘点，则为该像素点的灰度值赋值为笔画宽度值，遍历整幅图像，最终得到原图所对应的笔画宽度图。
-

§ 3.2.2 基于启发式规则的 SWT 算法改进

当原始图像经过 MSER 算法得到文本候选区域后，使用 SWT 算法通过计算得到

相对应的笔画宽度图像，包括了文本区域和非文本区域，而 SWT 算法能较好的将二者做区分。通过基于 SWT 算法的启发式规则将部分非文本区域过滤掉，便于将剩余的文本区域聚集成文本行。使用到的启发式规则包括：

(1) 同一幅图像中汉字的笔画宽度值基本保持不变，即一个候选区域的笔画宽度值与整幅图像的平均笔画宽度值的差距较小。而标准差是用于表征一组给定的数据中某数据与其平均值间的差异程度。即当某个区域的笔画宽度值的标准差较小时，该区域为文本区域；而标准差较大时，则为非文本区域。把笔画宽度值的标准差大于 5.2 的区域认为是非文本区域，将其过滤掉，如公式 (3.5)：

$$ResultSWT1 = \left\{ ResultMSER4_i \mid \sqrt{\frac{1}{N} \sum_{j=1}^N (SWT_j - \mu)^2} < 5.2 \right\} \quad (3.5)$$

N 表示一幅图像中候选区域的个数， SWT_j 是一幅图像中第 j 个候选区域的笔画宽度值， μ 是一幅图像的笔画宽度值的算术平均值。

(2) 在同一幅图像中，一般相邻文本字号一致，其笔画宽度值相差不大。如果候选区域邻域像素的笔画宽度值与当前像素的笔画宽度值相差较大，说明该区域是非文本区域，需将之过滤掉。当邻域像素的笔画宽度值与当前像素的笔画宽度值之比小于 3 时，效果最佳，如公式 (3.6)：

$$ResultSWT2 = \{ ResultSWT1_i \mid \frac{NeiSW_i}{CurSW_i} < 3 \} \quad (3.6)$$

$NeiSW_i$ 是邻域像素的笔画宽度值， $CurSW_i$ 是当前像素的笔画宽度值。

(3) 将笔画宽度值限定在 (20,300) 之间，过滤掉笔画宽度值过大或过小的区域。如果笔画宽度值过小，一般是小的点或极细的线条，而不是字符区域，应该被过滤掉；而在拍摄的自然场景图像中，大多文字笔画宽度不会很大，需过滤掉笔画宽度值过大的区域，如公式 (3.7)：

$$ResultSWT3 = \{ ResultSWT2_i \mid 20 < SW_i < 300 \} \quad (3.7)$$

其中， SW_i 是笔画宽度值。

§ 3.3 基于汉字结构的算法改进

SWT 算法能较好的将文本区域和非文本区域区分开，但在将单个文字或字符组合成文本行时，英文文本和中文文本间存在着很大的不同。

在英文中大多数字母都是由一个完整的部分构成，只有“i”由两部分构成。但由于“i”上方的点很小，即使丢失也不影响最终结果。相对而言，汉字复杂，变化较多，包括上下结构、左右结构、全包围结构、半包围结构和品字形结构等，结构之间互不

相连。如果不对其进行处理,当图像中的文本行走向是水平方向,并且有汉字是上下结构时,就无法将文本聚合成文本行;反之亦然。须先将候选区域聚集成汉字,再将汉字聚合成文本行。

由于单个汉字各结构间的距离一定小于相邻汉字间的距离,根据该规则可以将候选区域组合成汉字。首先计算两个候选区域间的距离,从距离最小的两个开始,判断这两个候选区域是否满足以下规则:

- (1) 如果两个候选区域有重合部分,说明这两个区域是同一个汉字的两部分;
- (2) 如果两个候选区域的质心坐标近似重合,说明这两个区域是同一个汉字的两部分;
- (3) 如果两个候选区域间的距离小于等于两个候选区域的长宽平均值,可能是同一汉字的两部分;
- (4) 如果两个候选区域的像素值相差不超过 30,可能是一个汉字的两部分;
- (5) 如果两个候选区域的笔画宽度值相差不超过 100,可能是一个汉字的两部分。

若满足,则将两个候选区域合并成一个汉字。然后根据距离从小到大依次进行组合,直到没有符合条件的候选区域为止,这样就将候选区域组合成一个个汉字。

文本行中的汉字一般都在同一条直线上,这些汉字质心的纵坐标(或横坐标)大小相差不大,每个汉字的最高点的纵坐标(最左侧点的横坐标)、最低点纵坐标(最右侧点的横坐标)都大致相同。根据这些特性,将汉字聚合成文本行。

§ 3.4 中文文本检测算法及描述

目前的自然场景文本检测算法对图像背景复杂时的处理效果不佳,误检率较高,且大多数算法都是针对英文环境进行的研究,鲜少有专门研究基于中文环境的场景文本检测算法,本章针对图像背景复杂时算法对中文文本的检测效果差的问题提出了改进算法 TDSI。使用 MSER 算法得到的最大稳定极值区域中,包括了文本区域和非文本区域。MSER 算法侧重于获取文本候选区域,如果仅使用该算法,得到的区域中包含大量非文本区域,将导致后期过滤非文本区域的工作量剧增。使用 SWT 算法得到相应的笔画宽度图像,根据笔画宽度值将文本区域和非文本区域区分开。SWT 算法侧重于对候选区域进行甄别,如果仅使用该算法,干扰信息较多,容易出现误判。TDSI 算法将 MSER 和 SWT 二者的优势相结合,既使用 MSER 去掉大量干扰信息,又使用 SWT 算法根据候选区域的笔画宽度值区分文本区域和非文本区域。通过论文提出的改进的 MSER 算法和改进的 SWT 算法过滤掉大量非文本区域。最后根据汉字的结构特征将文本区域聚集成单个汉字,再将其聚合成文本行。TDSI 算法流程如下:

(1) 通过 **MSER** 算法得到最大稳定极值区域即文本候选区域，使用启发式规则过滤掉部分明显的非文本区域；

(2) 通过 **SWT** 算法得到笔画宽度图像，运用相应的启发式规则将非文本区域过滤掉，得到文本区域；

(3) 根据汉字的结构特征将文本区域聚集成中文单字；

(4) 把汉字聚集成文本行，使用矩形框对其进行渲染。

算法流程图如图 3.3 所示：

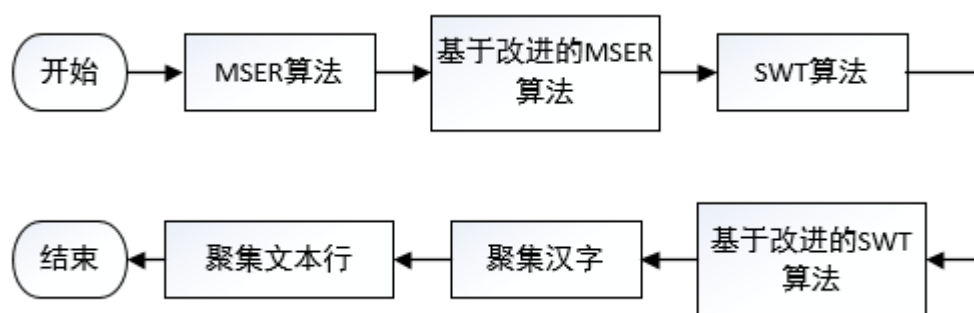


图 3.3 TDSI 算法流程图

§ 3.5 实验结果与分析

§ 3.5.1 实验结果

实验使用 Busta^[13]算法、Chen^[26]算法和本文提出的 TDSI 算法做对比，Busta 算法是近两年提出的自然场景图像中英文文本检测算法，Chen 算法使用了 MSER 和 SWT 进行英文文本的检测。实验过程中，TDSI 算法忽略了所有字符数量少于 3 个和包含了非法字符诸如数字、标点符号的文本区域。实验结果如图 3.4 所示。图 3.4(a)是从自建图像数据集中任意抽取的两幅原图，图 3.4(b)是由 Chen 算法获得的实验结果图，图 3.4(c)是由 Busta 算法得到的实验结果图，图 3.4(d)是由 TDSI 算法得到的实验结果图。图中蓝色矩形框框出的部分即为算法检测到的文本区域。在图 3.4(b)中把第一幅图像中的商品错误识别成文本区域，第二幅图像只检测出部分文本区域。在图 3.4(c)中第一幅图像正确检测出部分文本区域，把一部分商品错误识别成文本区域，第二幅图像中把背景中的人错误识别成文本区域。在图 3.4(d)中文本区域定位基本正确，说明针对背景复杂的自然场景图像中的中文文本，TDSI 算法比 Chen 算法、Busta 算法有明显优势。



(a) 原图



(b) Chen 算法实验结果图



(c) Busta 算法实验结果图



(d) TDSI 算法实验结果图

图 3.4 TDSI 算法与 Chen 算法、Busta 算法实验结果对比图：

a: 原图；b: Chen 算法实验结果图；c: Busta 算法实验结果图；d: TDSI 算法实验结果图

§ 3.5.2 结果分析

准确率、召回率和 F 值取自建图像数据集中所有图像检测结果的平均值。实验结果采取 ICDAR 竞赛的评价标准^[55]。检测结果对比如表 3.1、表 3.2 和图 3.6 所示。

表 3.1 TDSI 算法与部分改进算法结果对比

方法	准确率	召回率	F 值
TDSI 算法	0.713	0.896	0.794
TDSI-MSER	0.547	0.910	0.683
TDSI-SWT	0.605	0.764	0.675
TDSI-Ch	0.453	0.981	0.62

表 3.1 中，TDSI-MSER 算法是指在 TDSI 算法基础上去掉对 MSER 算法的改进，TDSI-SWT 算法是指在 TDSI 算法基础上去掉对 SWT 算法的改进，TDSI-Ch 算法是指在 TDSI 算法基础上去掉针对汉字结构的改进。由表 3.1 可知，TDSI-MSER 算法、TDSI-SWT 算法和 TDSI-Ch 算法的准确率较低，召回率较高，即误检率较高。TDSI-MSER 算法和 TDSI-SWT 算法只对 MSER 和 SWT 中的一个算法进行改进，最终得到的文本区域中有部分非文本区域，所以 TDSI 算法比 TDSI-MSER 算法和 TDSI-SWT 算法检测效果好。TDSI-Ch 算法没有考虑汉字的结构特征，部分文本区域无法聚集成文本行，因此 TDSI-Ch 算法误检率较高，说明汉字结构特征对提高中文文本检测结果是有帮助的。

表 3.2 检测结果对比

方法	准确率	召回率	F 值
TDSI 算法	0.713	0.896	0.794
Busta 算法	0.529	0.873	0.659
Chen 算法	0.580	0.667	0.620

由表 3.2 和图 3.5 可知，TDSI 算法的准确率、召回率和 F 值均最高。Busta 算法的准确率最低，召回率和 F 值较高。Chen 算法的召回率和 F 值最低，准确率较高。

据文献[13]介绍，Busta 算法提取的候选区域比 MSER 算法多，而 Chen 算法使用 MSER 提取候选区域，因此 Busta 算法的召回率比 Chen 算法高。又因为 Chen 算法使用笔画宽度值过滤非文本区域，而 Busta 算法没有使用任何方法过滤非文本区域，所以 Busta 算法的准确率没有 Chen 算法高，即 Busta 算法的误检率较高。

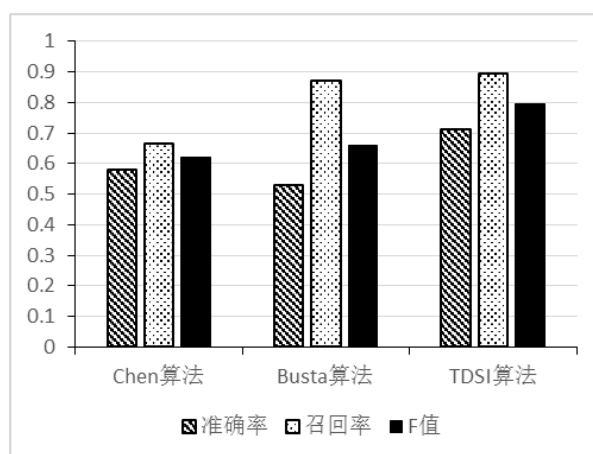


图 3.5 检测结果对比

TDSI 算法使用改进的 MSER 算法，比 Busta 算法提取的候选区域少，且提取的候选区域大多是文本区域，因此 TDSI 算法的召回率比 Busta 算法高。另外，TDSI 算法使用改进的 SWT 算法过滤非文本区域，而 Busta 算法没有过滤非文本区域，因此 TDSI 算法的准确率比 Busta 算法高。虽然 Chen 算法同时使用 MSER 和 SWT 算法，但 Chen 算法只对 MSER 算法进行改进。而 TDSI 算法分别对 MSER 算法和 SWT 算法都做了改进，且根据汉字的结构特征对算法进行改进，因此 TDSI 算法的结果比 Chen 算法好。综上所述，与 Chen 算法和 Busta 算法相比，TDSI 算法的检测结果较好，误检率较低。

§ 3.6 本章小结

本章详细介绍了 MSER 算法和 SWT 算法，并提出了一种基于自然场景图像的中文文本检测算法 TDSI。该算法对 MSER 和 SWT 算法做出了改进，并利用了汉字的结构特征聚集中文字单字。实验结果表明，对于背景复杂的自然场景图像，TDSI 算法对中文文本的检测效果较好，能较准确地检测出文本区域。

第四章 基于颜色矩的自然场景中文文本检测算法

§ 4.1 颜色空间

颜色空间也被称为彩色模型，主要用于在特定规范下用寻常可接受的方式表征颜色。颜色空间作为一种对色彩的描述方法，种类较多。下文将对几种较常见的颜色空间做简要介绍。

§ 4.1.1 RGB 颜色空间

RGB 颜色空间是按照物体的发光原理定义的，R、G、B 三个分量与光的三原色逐一对应。它是图像处理过程中最简单的颜色空间。通过数码相机或智能手机拍摄的图像，一般都被分为 R、G、B 三色加以保存。RGB 颜色空间如图 4.1 所示。

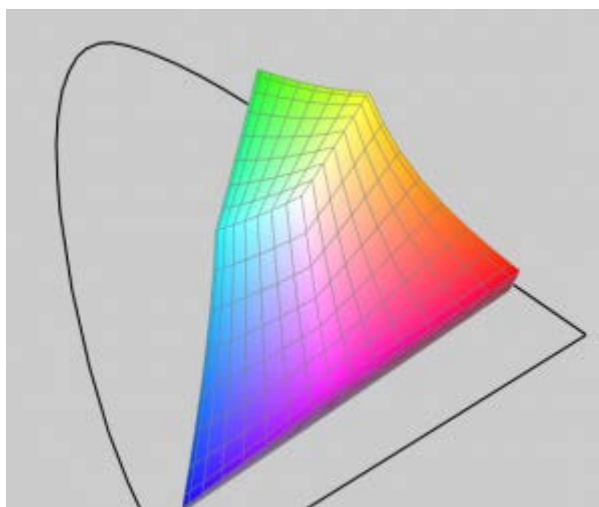


图 4.1 RGB 颜色空间

RGB 颜色空间一个重要的不足是无法直接从 R、G、B 三个分量中看出所表征的颜色的亮度、饱和度等信息，是以 RGB 颜色空间不适合人眼对颜色的认识规律。并且 RGB 颜色空间是最不均匀的，即通过计算两个不同像素点的颜色值的差值无法表征两种颜色间的感知差别。此外，在自然环境下获取的图像易受光照、障碍物、阴影等因素的影响，即对亮度较敏感。亮度出现改变时，R、G、B 三个分量也会随之呈现改变。RGB 颜色空间相对比较适合在显示仪器上显示图像，但并不适合用于图像处理。

§ 4.1.2 CMYK 颜色空间

CMYK 颜色空间即印刷四色模式，是依据光的反射原理定义的，主要用在印刷行业中，彩色印刷机和彩色打印机均采用 CMYK 颜色空间。C、M、Y、K 四个分量分别代表青、品红、黄、黑四种颜料，取值范围均为[0%，100%]。CMYK 颜色空间主要用于描述在印刷过程中需要使用哪种油墨，需要使用的油墨量所占的比例。理论上，通过印刷或打印出来的颜色都能通过 C、M、Y 这三种颜色根据完全不同的比例混合而成。

§ 4.1.3 YUV、YCrCb 颜色空间

YUV、YCrCb 颜色空间多用于电视系统中。YUV 颜色空间是欧洲电视系统采用的颜色模型，该颜色空间将颜色值分为一个亮度值和一个色差值来进行传输。YCrCb 颜色空间与 YUV 很相似，YCrCb 只是 YUV 的压缩和偏移后得到的一种版本。因为人眼对光照的敏感水平远大于对其他颜色的敏感水平，于是让一幅图像中相邻的像素用同一个颜色值表现。

§ 4.1.4 CIE L*a*b*、CIE L*u*v*颜色空间

1976 年，CIE（Commission Internationale de L'Eclairage，国际照明委员会）制定了两种不同的颜色空间。这两种颜色空间分别是 CIE L*a*b*和 CIE L*u*v*，它们对颜色的感知比其他颜色空间更加均匀，为人们评估两种完全不同的颜色之间的近似程度提供了一种有效的方法。该方法可以通过色差 ΔE 来表现两种不同颜色的差别。这两种颜色空间多用于计算机色调的调整和色彩的校正方面。

§ 4.1.5 HSV（HSB）颜色空间

HSV 颜色空间是一种基于视觉感知的颜色模型。如图 4.2 为 HSV 颜色空间。HSV 颜色空间具有以下两个特点：图像的亮度与图像的颜色信息无关；图像的色调和饱和度与人眼感知颜色的途径有着很大的关联。以上的这些特性使 HSV 颜色空间适用于那些符合人眼的视觉特征来认知颜色的处理方式。HSV 颜色空间与人眼感知色彩的视觉特性的三要素相对应，而 HSV 的三个通道间都是相互独立、互不影响的。使用 HSV 颜色空间能够完全独立地感知不同颜色的不同变化，而它的色调更能影响人眼对颜色的判断。HSV 颜色空间也被称为 HSB（B 指 Brightness）颜色空间在表示颜色

时，相对于 RGB 等显得更自然一些。

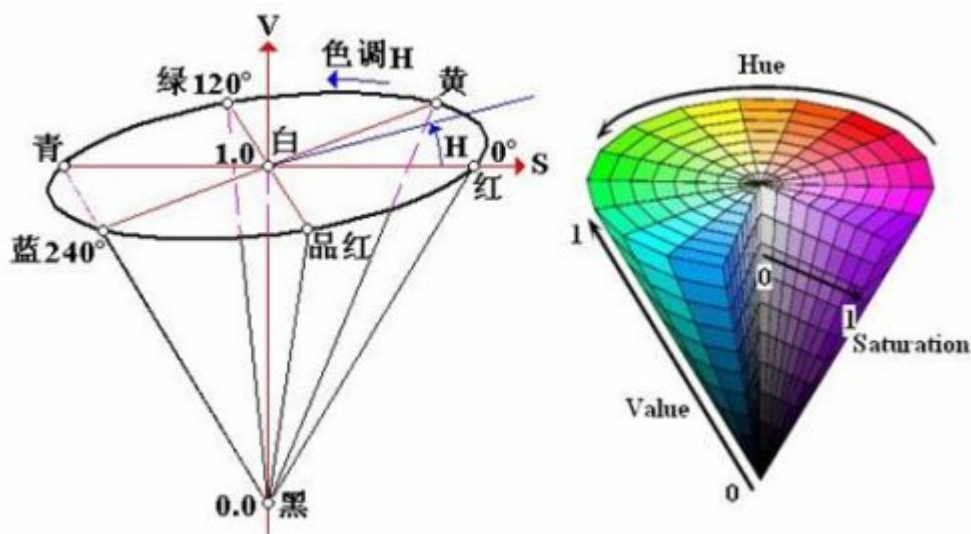


图 4.2 HSV 颜色空间

大多数自然场景图像都是通过智能手机和数码相机采集到的，这些设备拍摄的图像都是采用 RGB 颜色空间对图像进行存储。但 RGB 颜色空间并不适用于图像处理。论文算法将先把图像从 RGB 转为 HSV 颜色空间，然后再对其进行后续的处理。

§ 4.2 颜色特征提取方法

图像中包含了许多个像素点，此中包括了丰富的特征，比如颜色、纹理、边缘和空间关系等特征。这些特征通常可以通过算法或人机交互的方法提取到^[63]。本节将针对图像中的颜色特征做简洁明了的说明。

颜色是图像处理过程中使用最频繁的特征之一。每幅图像都包含的最基本的特征就是颜色。颜色特征是全局特征，与图像的大小、方向、视角等信息完全无关，具有较好的鲁棒性。通常的图像处理算法会先对图像进行灰度化，这样就会把图像的颜色信息全部丢失。如果遇到灰度图的处理效果不佳时，就有必要使用图像的颜色特征对图像做一定的其他处理。另外，颜色特征比其他特征诸如纹理特征、边缘特征的计算复杂度低，能够较大幅度提升图像处理的效率。颜色特征的表现方式多种多样，论文将对一些常见的颜色特征作简要介绍。

§ 4.2.1 颜色直方图

在颜色特征的表现方式中，使用最普遍的便是颜色直方图。颜色直方图指的是在整幅图像中不同的颜色值呈现的概率。颜色直方图作为表征颜色的一种方式，在特征

提取和相似度计算时相对较简练。颜色直方图还包括随图像的尺度、旋转等变化不敏感的优秀特征。

假设一幅图像用 $(f_{xy})_{M \times N}$ 表征，则图像的颜色直方图 h_c 如公式 4.1 所示：

$$h_c = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(f_{ij} - c), \forall c \in C \quad (4.1)$$

f_{xy} 代表的是像素点 (x, y) 处的颜色值， $M \times N$ 指整张图像的尺寸，图像中所包含的颜色集则被记为 C 。

颜色直方图具有很多的优点，但其并不能较好的显示颜色值的空间位置，即不能很直观的表征图像中的事物。颜色直方图还存在特征维数较高的问题，相应的其计算复杂度也会比较大。

§ 4.2.2 颜色熵

John 等^[64]主要按照的是颜色直方图的特点和信息论中信息熵的概念，第一次提出了使用颜色的信息熵来表征颜色特性的方法。颜色熵，就是图像颜色的信息熵。使用颜色熵来表征颜色特征，比起使用图像的颜色直方图来说，将多维数据降到一维数据，能够较明显的提升算法的计算效率。

假设图像的颜色直方图可以用 (h_1, h_2, \dots, h_n) 来表征。把颜色直方图当作是不同的颜色值在同一幅完整的图像中出现的概率。则按照信息熵的理论，图像的颜色熵可以用公式 4.2 来表示：

$$E = - \sum_{i=1}^n h_i \log_2(h_i) \quad (4.2)$$

使用颜色熵能够大幅度的减少颜色直方图的维数，但它的分辨能力较低。大多算法往往会将颜色熵与其他一些颜色特征联合使用，这样效果会更好一些。

§ 4.2.3 颜色聚合向量

针对颜色直方图不能表征颜色的空间分布信息的不足，Pass 等^[65]提出了一种新的描述颜色特性的方式：颜色聚合向量（color coherence vector）。该方法是颜色直方图的另一种改进，它的关键之处是将原本属于颜色直方图中每个区域的像素点分成两部分：若该区域内的一些像素所占的连续区域的面积大于某个给定的阈值，则该区域内的像素为聚合像素，不然即非聚合像素。图像的颜色聚合向量可以表示为：

$$\langle (\alpha_1, \beta_1), (\alpha_2, \beta_2), \dots, (\alpha_n, \beta_n) \rangle \quad (4.3)$$

公式 4.3 中， α_i 和 β_i 分别表征第 i 个区域的聚合像素和非聚合像素的个数，而 $\langle \alpha_1 + \beta_1, \alpha_2 + \beta_2, \dots, \alpha_n + \beta_n \rangle$ 是指该图像相对应的颜色直方图。

§ 4.2.4 颜色相关图

Huang^[66]首次提出了颜色相关图（color correlogram）的概念。它是图像中的颜色分布的另一种表现形式。颜色相关图是使用图像中不同像素间的颜色关系来表征图像的颜色特征的一种方式。该特征不光描述了某种颜色的像素数占据整幅图像的比值，而且还表征了不同颜色间的空间相关性。

§ 4.2.5 颜色矩

颜色矩由 Stricker 等^[67]初次提出，是另外一种相对比较简洁有效的颜色特征的表达方式。它所依赖的原理主要是图像中的任意颜色的分布都能够通过使用图像的矩来表现。

矩是指用来表征随机变量的一种数字特征。假设 X 为随机变量， c 为一个常数， k 为正整数，那么 $E[(x-c)^k]$ 就代表了 X 关于 c 点的 k 阶矩。其中，当 $c=0$ 时， $a_k = E(X^k)$ 为 X 的 k 阶原点矩。当 $c=E(X)$ 时， $\mu_k = E[(X-EX)^k]$ 为 X 的 k 阶中心矩。一阶原点矩即期望，一阶中心矩 $\mu_1=0$ ，二阶矩 μ_2 便是 X 的方差 $Var(X)$ ，三阶矩 μ_3 用来衡量分布是否有偏。

对一幅图像来说，若是能把像素的坐标当做一个二维的随机变量 (X,Y) ，则一幅图像就能用二维密度函数来表现，即能够用图像的矩来描述这幅图像的特征。

颜色矩即图像中的颜色的矩，用以表征图像中的颜色特征。颜色矩主要包含颜色一阶矩（均值）、二阶矩（方差）和三阶矩（斜度）。用颜色矩表征图像时，不消对颜色特征做空间量化，而且特征维数比较少，运算量也比较小。在实际应用的过程中，常常使用颜色矩来表现图像的颜色特征。

颜色一阶矩是通过一阶原点矩计算得到的，即颜色的均值，如公式 4.4 所示，表示图像的整体明暗程度。颜色一阶矩 μ_i 的值越大，图像的亮度越大。

$$\mu_i = \frac{1}{N} \sum_{j=1}^N p_{i,j} \quad (4.4)$$

i 表示图像的颜色通道数。如果图像为灰度图像，则 $i=1$ ，如果为彩色图像，则 $i=3$ 。 $p_{i,j}$ 指彩色图像中的第 i 个颜色通道中灰度值为 j 的像素点在整幅图像中出现的概率， N 指整幅图像中全部像素的总量。

颜色二阶矩是二阶颜色中心距的平方根，即颜色的标准差，如公式 4.5 所示，主要用来表示图像中颜色的分布情况。颜色二阶矩 σ_i 的值越大，则图像的颜色分布范围越广。

$$\sigma_i = \left(\frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}} \quad (4.5)$$

颜色三阶矩是三阶颜色中心距的立方根，即颜色的斜度，如公式 4.6 所示，表示图中颜色分布的对称性。当 $s_i=0$ 时，图像的颜色分布是近似对称的。当 $s_i \neq 0$ 时，图像的颜色分布是不对称的，即：当 $s_i > 0$ 时，称之为图像的颜色分布是左偏的或者负偏的；当 $s_i < 0$ 时，称之为图像的颜色分布是右偏的或者正偏的。

$$s_i = \left(\frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}} \quad (4.6)$$

因为图像中颜色的分布信息大部分都存在于低阶矩中，所以仅使用颜色一阶矩、二阶矩和三阶矩就能够相对较精准的表现不同颜色间的分布信息。图像中的每个像素都具有红绿蓝三个不同的颜色通道，是以图像的颜色矩可用 9 个分量来表现。与颜色直方图相对比，颜色矩无须进行颜色空间量化，而且颜色矩的特征向量维数不高，计算量较小。颜色矩还拥有平移不变性、旋转不变性和缩放不变性等特性，可以用颜色矩来表征颜色特征。论文算法将使用颜色矩来表现颜色特征。

§ 4.3 颜色距离计算

在对彩色图像进行处理时，通常会涉及到对颜色间差异的度量。在实际研究过程中，一般使用色差度量不同颜色间的差异。计算色差的方法多种多样，常见的有基于 CIEL*a*b* 颜色空间的色差计算、欧式距离和基于 HSV 颜色空间的欧式距离等，下文将大致介绍这些常见的计算方法。

§ 4.3.1 基于 CIEL*a*b* 颜色空间的色差计算

CIEL*a*b* 颜色空间是基于人眼对颜色的感知特性的，可以相对较好的表征人眼所能观察到的全部颜色值。其中， L^* 指图像的明度， a^* 指红绿颜色色差， b^* 指蓝黄颜色色差。图像中两个不同颜色间的色差 ΔE^* 计算公式如公式 4.7 所示。

$$\Delta E^* = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2} \quad (4.7)$$

其中， ΔE^* 表现图像中两个不同的颜色间的色差， ΔL^* 、 Δa^* 、 Δb^* 分别代表两个不同颜色在不同分量的差值。在计算基于 CIEL*a*b* 颜色空间的色差时，大多数算法都会先将图像从 RGB 转换到 CIEL*a*b* 颜色空间。但是这个转换过程的计算量较大，该方法较少被用到。

§ 4.3.2 欧式距离

颜色距离指两个不同的颜色值间的差异。颜色距离越大，两个颜色的差异就越大。在计算图像的颜色距离时，一般采取欧式距离。欧式距离是一种常见的距离定义，用来表现两个点间的真实距离。在图像处理过程中，两个不同颜色间的欧式距离公式如公式 4.8 所示。

$$\|C_1 - C_2\| = \sqrt{(C_{1,R} - C_{2,R})^2 + (C_{1,G} - C_{2,G})^2 + (C_{1,B} - C_{2,B})^2} \quad (4.8)$$

其中， C_1 和 C_2 分别指图像中两个完全不同的颜色 1 和颜色 2， $C_{1,R}$ 表示颜色 1 的 R 通道颜色值。

§ 4.3.3 基于 HSV 颜色空间的欧式距离

因为 RGB 颜色空间是线性并且相互正交的，而人眼的感知系统却不是线性的，其并不能很好的表征人眼对颜色的感知情况，其相对应的颜色距离也不能真实的反映两个不同的颜色是否相似。而 HSV 颜色空间能够相对较好的反映人眼的视觉特征。计算两个不同的颜色间的颜色距离时，通常会先将图像从 RGB 转换成 HSV 颜色空间，再对其做运算。

HSV 颜色空间是一契合人眼对色彩感知的颜色空间，能够把图像的亮度和颜色较准确的区分开来。但是 HSV 颜色空间有自身的不足之处：当 $s=0$ 时， h 没有定义；当 s 取极小值时， h 不够稳定。为了考虑这两种情况，使用公式 4.9 来表征 HSV 颜色的差异：

$$dist(h's'v', hsv) = \left\| \begin{pmatrix} v \times s \times \cos(2\pi \times h), v \times s \times \sin(2\pi \times h), v \\ -(v' \times s' \times \cos(2\pi \times h'), v' \times s' \times \sin(2\pi \times h'), v') \end{pmatrix} \right\| \quad (4.9)$$

§ 4.4 基于颜色矩的文本检测算法

目前很多学者在对场景文本检测技术进行研究时，一般先对图像进行灰度变换，将图像中的颜色特征去掉，然后对得到的灰度图像进行后续处理。因为一般通过数码相机或智能手机得到的图像都是基于 RGB 颜色空间的，这并不符合人眼的视觉特征。虽然 HSV 颜色空间相对较合适人眼观察事物的视觉规律，但由于从 RGB 转换到 HSV 颜色空间的过程相对较复杂，如果在数码相机或手机上拍摄完图像后直接进行 RGB 到 HSV 颜色空间的转换，无法实时将拍摄的图像展现给用户，导致用户体验较差。在一般的图像处理过程中，颜色特征起到的作用较小，多数算法将颜色特征忽略掉。

然而，在场景文本检测过程中，图像的颜色特征却起着很重要的作用。因为图像中的文字一般都具有相同或相似的颜色，而且有与图像背景完全不同的颜色值。相对于计算图像的纹理特征或边缘特征，颜色特征计算简单，计算量较小，且可以较直观的将图像中的背景部分和文本部分区分开来。综上所述，论文提出一种改进算法 TDSI-C，在 TDSI 算法的基础上添加颜色特征来提高自然场景图像中文本区域的检测精度。

MSER 算法要求被处理的必须是灰度图像，忽略了很重要的颜色特征。SWT 算法也是对灰度图像进行处理，也未考虑颜色特征。但在自然场景图像中，颜色是一个非常重要的特征。图像中的文本通常具有相同或相近的颜色值，一是为了美观，二是为了能够与背景有较明显的区别，让人们较容易的在图像中观察到文本，起到标识和警告的作用。为了与图像背景作区分，同时也是为了让人们能清晰明了的从图像中观察到文本内容，图像中文本的颜色一定跟图像中的背景颜色区别很大。算法将在原有的基础上添加颜色特征，来区分文本区域和非文本区域。

RGB 颜色空间能较好的表征图像的颜色，但却与人眼的视觉特征并不契合，而 HSV 颜色空间相对而言更符合人眼观察事物的视觉规律。TDSI-C 算法在计算文本候选区域的颜色矩之前，将图像从 RGB 转换到 HSV 颜色空间。

设 RGB 颜色空间中的颜色值分别为 R, G, B ，其中 $R, G, B \in [0,1]$ 。则颜色空间的转换公式如公式 4.10, 4.11, 4.12 所示：

$$V = \frac{1}{\sqrt{3}}(R + G + B) \quad (4.10)$$

$$S = 1 - \frac{\sqrt{3}}{V} \min(R, G, B) \quad (4.11)$$

$$H = \begin{cases} \theta, G \geq B \\ 2\pi - \theta, G < B \end{cases} \quad (4.12)$$

$$\text{其中, } \theta = \arccos \left[\frac{\frac{1}{2}[(R - G) + (R - B)]}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right]。$$

由于 HSV 颜色空间的色调、饱和度和明度与人眼观察事物的视觉规律较吻合，而且与其他颜色空间相对比，HSV 颜色空间能较好的表征人眼对颜色的认知特性。论文算法将采用 HSV 颜色空间来表示图像。

图像中用于表征颜色特征的方法多种多样，其中的颜色矩特征不需要对图像中的颜色特征进行空间量化。而且颜色矩的特征维数较少，计算量也较小，算法将使用颜色矩来表征颜色特征。

由于每个像素都由 H、S、V 三个分量构成，因此将文本候选区域的前三阶颜色矩构成了九维向量。使用颜色矩表现颜色特征的九维向量如公式 4.13 所示：

$$F_{color} = [\mu_H, \sigma_H, s_H, \mu_S, \sigma_S, s_S, \mu_V, \sigma_V, s_V] \quad (4.13)$$

得到文本候选区域的颜色矩后，需计算不同文本候选区域颜色矩间的相似度。可以通过计算颜色矩间的距离判断文本候选区域的颜色矩的相似度。算法使用基于 HSV 颜色空间的欧式距离来描述不同文本候选区域颜色矩间的相似度。这是因为 TDSI-C 算法已将颜色空间从 RGB 转换为 HSV，而基于 HSV 颜色空间的欧式距离既能较准确的计算颜色矩间的相似度，又兼顾了 HSV 颜色空间能表征人眼对颜色特征的感知情况。其算法的时间复杂度为 $O(n^2)$ 。相似度越高，文本候选区域的颜色越相似，是文本区域的概率越大。如果使用基于 CIEL*a*b* 颜色空间的色差来度量颜色矩间的相似度，计算量较大。因此 TDSI-C 算法将使用基于 HSV 颜色空间的欧式距离计算不同文本候选区域间颜色矩的相似度。

计算不同文本候选区域的颜色矩相似度的算法描述如算法 4.1。

算法 4.1:

算法输入：一幅图像的所有文本候选区域

算法输出：一幅图像中的所有正确的文本区域

- 1: 将文本候选区域的颜色空间从 RGB 颜色空间转换为 HSV;
- 2: 计算每个候选区域的九维颜色矩;
- 3: 计算每个文本候选区域与其他文本候选区域的颜色矩间的相似度;
- 4: 若是颜色矩相似度大于阈值，则该文本候选区域为文本区域，将之保留；若是小于阈值，则该文本候选区域为非文本区域，将之去掉。

算法中，颜色一阶矩、二阶矩和三阶矩相似度阈值分别为 90、1000、2000。阈值的参数确定过程参见实验部分。

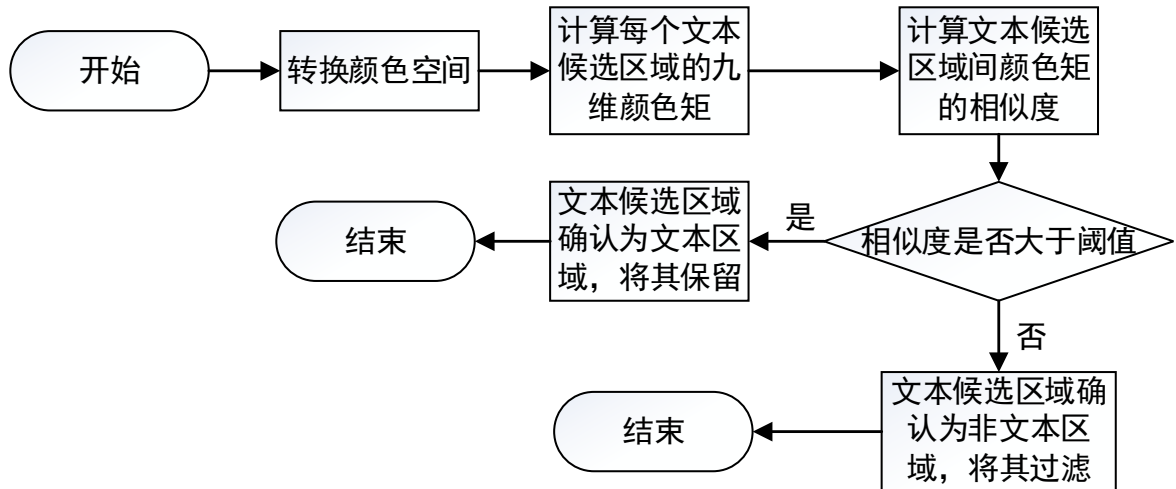


图 4.3 TDSI-C 算法流程图

TDSI-C 算法思想：首先使用改进的 MSER 算法提取文本候选区域；其次，使用

SWT 算法获取文本候选区域的笔画宽度图像，运用启发式规则将非文本区域过滤掉；然后，计算一幅图像中每个文本候选区域的颜色矩特征值，并计算不同文本候选区域的颜色矩间的相似度，如果相似度大于阈值，则文本候选区域确认为文本区域，将其保留，如果相似度小于阈值，则文本候选区域确认为非文本区域，将其过滤掉；最后根据汉字的结构特征将文本区域聚集成文本行。具体算法流程如图 4.3 所示。

§ 4.5 实验结果与分析



图 4.4 部分实验结果图示例

TDSI-C 算法使用 ICDAR 评价标准对算法的性能进行评估,实验数据使用自建实验图像数据集。部分实验结果如图 4.4 所示,图中用蓝色矩形框框出来的部分表示 TDSI-C 算法检测到的图像中的文本区域。从实验结果可知,TDSI-C 算法的检测结果较好,能将实验图像数据集中大多数自然场景图像中的中文文本检测出来,且检测准确率较高。

颜色矩相似度的阈值对算法结果的影响如图 4.5、4.6、4.7 所示。图中的横轴表示相似度阈值,纵轴表示准确率 P、召回率 R 和 F 值。

图 4.5 是指在 TDSI 算法的基础上只添加颜色一阶矩时,随着颜色一阶矩的相似度阈值的变化得到的算法结果的变化规律。当颜色一阶矩的相似度阈值为 90 时,P 值为 0.785,R 值为 0.873,F 值为 0.827,算法结果最佳。实验结果表明,在 TDSI 算法的基础上添加颜色一阶矩对算法结果有明显的改进。

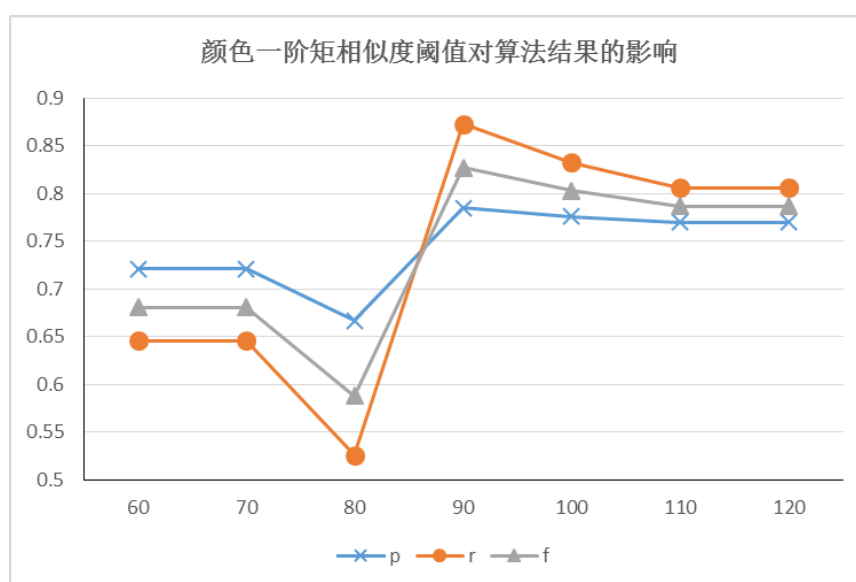


图 4.5 颜色一阶矩相似度阈值对算法结果的影响

图 4.6 是指在 TDSI 算法的基础上添加颜色一阶矩和颜色二阶矩,其中颜色一阶矩的相似度阈值取 90 时,随着颜色二阶矩的相似度阈值的变化得到的算法结果的变化规律。当颜色二阶矩的相似度阈值为 1000 时,P 值为 0.788,R 值为 0.886,F 值为 0.834,此时算法的检测结果最佳。根据实验结果,颜色二阶矩对算法结果有改进,但不太明显。

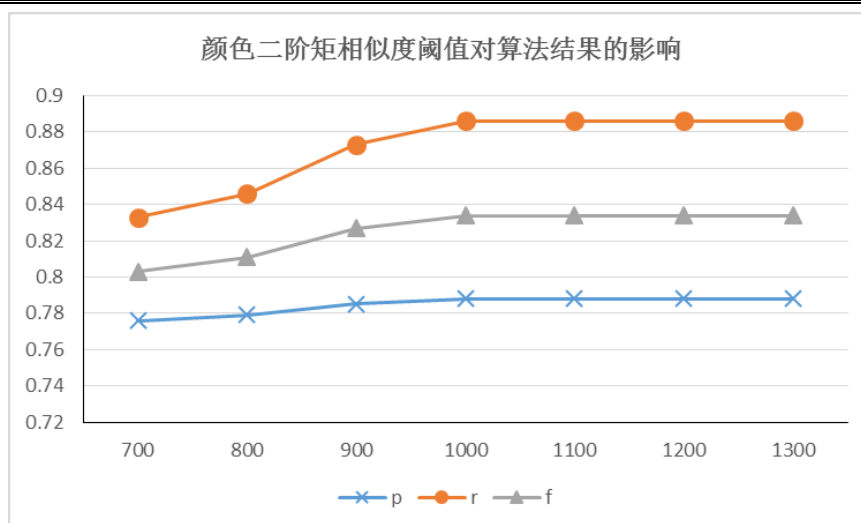


图 4.6 颜色二阶矩相似度阈值对算法结果的影响

图 4.7 是指在 TDSI 算法的基础上添加了颜色一阶矩、二阶矩和三阶矩三个特征，其中，颜色一阶矩的相似度阈值取 90，颜色二阶矩的相似度阈值取 1000 时，随着颜色三阶矩的相似度阈值的变化得到的算法结果的变化规律。当颜色三阶矩的相似度阈值为 2000 时，P 值、R 值、F 值分别为：0.799、0.939、0.864，此时算法的结果最优。当颜色一阶矩、二阶矩和三阶矩的相似度阈值分别为 90、1000、2000 时，TDSI-C 算法的检测结果最佳。实验结果表明，颜色三阶矩对算法结果有较好的改进。

综合图 4.5、4.6、4.7 可知，在 TDSI 算法的基础上添加颜色矩特征，对场景文本检测算法的结果有较明显的改进。这是因为自然场景图像中相邻的文本一般具有相同或相近的颜色值，图像中颜色矩特征能较好的表征图像中颜色的亮度、分布范围和偏色的情况。图像中的文本区域和非文本区域，颜色不同，颜色的亮度、颜色值的分布范围、颜色的偏色情况也不相同，是以颜色矩特征能较好的将非文本区域过滤掉，添加颜色矩特征能提高场景文本算法的检测结果。

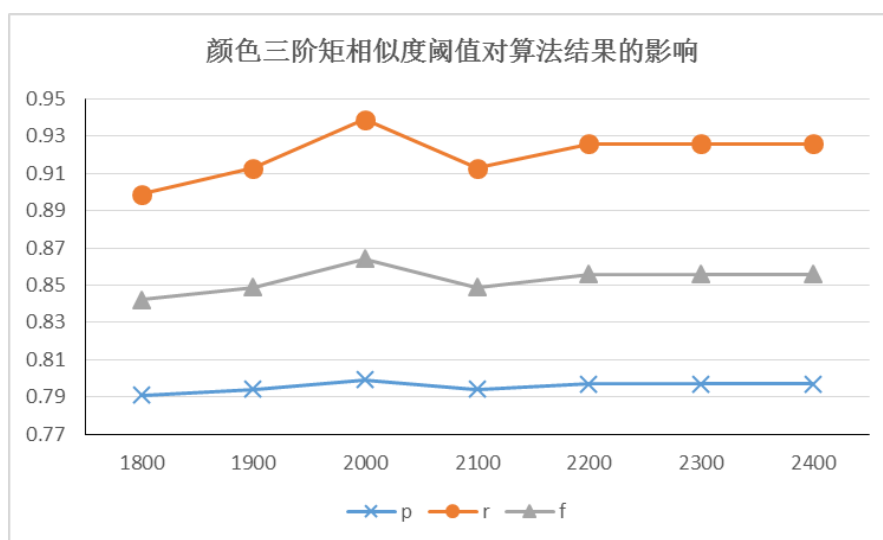


图 4.7 颜色三阶矩相似度阈值对算法结果的影响

为了比较使用颜色矩的文本检测效果,使用 TDSI 算法、Busta^[13]算法、Chen^[26]算法与 TDSI-C 算法作对比,结果如表 4.1。由表中的数据可知,TDSI-C 算法相较其他三个算法检测效果好,其准确率 P、召回率 R、F 值都是最高的。Chen、Busta、TDSI 三个算法在检测图像中的文本时都先将原图转换为灰度图像再做后续处理,都忽略了颜色特征,TDSI-C 算法在 TDSI 的基础上添加了颜色特征即颜色矩。从实验结果可以看出,颜色特征对文本检测起到了很重要的作用,说明颜色矩能将文本区域与图像背景较好地区分开。

表 4.1 实验结果对比

方法	准确率	召回率	F 值
Chen 算法	0.58	0.667	0.62
Busta 算法	0.529	0.873	0.659
TDSI 算法	0.713	0.896	0.794
TDSI-C 算法	0.799	0.939	0.864

§ 4.6 本章小结

本章首先就颜色空间、常见的颜色特征提取方法、颜色距离计算等方面做了详细的介绍。然后介绍了改进的基于颜色矩的自然场景图像中的中文文本检测算法 TDSI-C。最后做了对比实验并对实验结果进行了分析。实验结果表明,与其他不使用颜色特征的算法相比,加入颜色特征的 TDSI-C 算法的检测结果较好。

第五章 总结与展望

§ 5.1 研究总结

随着信息技术的高速发展,图像数据的获取变得越来越便捷,检测自然场景图像中的文本信息成为图像处理领域的一项研究热点。论文从自然场景文本检测技术的检测方法入手,主要对自然场景文本检测做了研究,重点研究了图像背景复杂时如何检测其中的中文文本和颜色特征对自然场景文本检测算法结果的影响。

论文完成的主要工作包括:

(1) 构建纯中文文本的实验图像数据集。当前公开的国内外自然场景图像数据集大部分是基于英文环境的,少数是基于中英文混合环境。基于此,论文建立了一个完全基于中文文本的自然场景图像数据集。利用高分辨率的智能手机和数码相机拍摄了学校内和学校周围的自然场景图像,并添加部分从网上找到的较合适的图像,一共403幅左右,构建了纯中文文本的实验图像数据集以备实验所用。

(2) 提出了一种基于自然场景图像的中文文本检测算法 TDSI。首先利用 MSER 算法对原图进行二值化处理,使用 MSER 算法提取最大稳定极值区域作为文本候选区域。利用基于 MSER 算法的启发式规则过滤掉部分明显的非文本区域。然后计算图像的笔画宽度值得到图像对应的笔画宽度图像,根据文本区域具有与图像背景部分明显不同的笔画宽度值和一些启发式规则,对文本候选区域进行筛选。根据汉字的结构特征,即候选区域的质心、重合区域等特征将文本区域先聚集成汉字,再将之聚集成文本行。TDSI 算法降低了文本检测的误检率,准确率为 0.713,召回率为 0.896, F 值为 0.794。

(3) 提出基于颜色矩的自然场景文本检测算法 TDSI-C。针对现有的算法大多忽略图像的颜色特征的问题,提出了 TDSI-C 算法。首先使用 MSER 算法得到文本候选区域,然后使用 SVM 算法对文本候选区域做验证,用到了包括图像的几何特征、笔画宽度值、颜色矩在内的一些特征。最后,根据汉字的结构特征,将得到的字符区域聚集成中文单字,并聚集成文本行。实验结果表明, TDSI-C 算法检测结果较好,准确率为 0.799,召回率为 0.939, F 值为 0.864。

§ 5.2 工作展望

在自然场景文本检测中，技术发展现状与实际的应用需求还存在着很大的差距，说明这方面还存在很多值得探究的难题。论文主要就部分问题做了针对性研究，但还有很多难题有待解决，包括但不限于：

（1）如果拍摄图像时，出现光照不均匀现象，可能导致图像中部分文本区域由于存在高光导致无法将之检测出来，目前还没有合适的方法解决该问题。

（2）构建的实验图像数据集规模较小，只包括了约 403 幅图像，下一步将拍摄更多符合实验条件的图像加入到数据集中，以提升算法的普适性。

（3）当图像出现抖动模糊的情况时，文本检测的效果较差。在后续研究中，应针对图像的抖动模糊现象做改进。

参考文献

- [1] 张健. 复杂图像文本提取关键技术与应用研究[D]. 南开大学, 2014.
- [2] 王润民. 自然场景文字检测方法研究[D]. 华中科技大学, 2015.
- [3] 姜维. 基于视觉显著性与颜色的复杂场景文字提取方法的研究[D]. 西安电子科技大学, 2014.
- [4] Tang Y, Wu X. Scene Text Detection via Edge Cue and Multi-features [C]// International Conference on Frontiers in Handwriting Recognition. IEEE Computer Society, 2016: 156-161.
- [5] Meng Q, Song Y, Zhang Y, et al. Text detection in natural scene with edge analysis[C]// IEEE International Conference on Image Processing. IEEE, 2013:4151-4155.
- [6] Yu C, Song Y, Meng Q, et al. Text detection and recognition in natural scene with edge analysis [J]. Computer Vision Iet, 2015, 9(4):603-613.
- [7] Yu C, Song Y, Zhang Y. Scene text localization using edge analysis and feature pool [J]. Neurocomputing, 2015, 175(PA):652-661.
- [8] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar J. Leite, Jorge Stolfi. T-HOG: An effective gradient-based descriptor for single line text regions. Pattern Recognition[J], 2013, 46(3):1078-1090.
- [9] Huang W, Qiao Y, Tang X. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees[C]// Computer Vision—eccv. 2014:497-511.
- [10] Galih Hendra Wibowo, Riyanto Sigit, Aliridho Barakbah. Feature Extraction of Character Image using Shape Energy [C]// International Electronics Sysposium, 2016: 471 – 475.
- [11] Jonathan Fabrizio, Myriam Robert-Seidowsky, Séverine Dubuisson, Stefania Calarasanu, Raphaël Boissel. TextCatcher: a method to detect curved and challenging text in natural scenes[J]. International Journal on Document Analysis and Recognition (IJDAR), 2016, 19(2):99-117.
- [12] 尹芳. 场景文本识别关键技术研究[D]. 哈尔滨理工大学, 2012.
- [13] Busta M, Neumann L, Matas J. FASText: Efficient Unconstrained Scene Text Detector [C]// IEEE International Conference on Computer Vision. IEEE, 2015:1206-1214.
- [14] 颜建强. 图像视频复杂场景中文字检测识别方法研究[D]. 西安电子科技大学, 2014.
- [15] 孙雷. 自然场景图像中的文字检测[D]. 中国科学技术大学, 2015.
- [16] Ma, Jianqi, Shao, Weiyan, Ye, Hao, et al. Arbitrary-Oriented Scene Text Detection via Rotation Proposals [J]. 2017, arXiv: 1703. 01086.
- [17] Liao M, Shi B, Bai X, et al. TextBoxes: A Fast Text Detector with a Single Deep Neural Network [J]. 2016, arXiv: 1611. 06779v1.
- [18] 张树业. 深度模型及其在视觉文字分析中的应用[D]. 华南理工大学, 2016.

-
- [19] Basavanna M, Shivakumara P, Srivatsa S K, et al. Adaptive histogram analysis for scene text binarization and recognition[J]. Malaysian Journal of Computer Science, 2016.
- [20] J. Matas, O. Chum, M. Urban and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions [J]. BMVC, 2002, pp. 384-393.
- [21] Nister, D, Stewenius, H. Linear Time Maximally Stable Extremal Regions[C]. 10th European Conference on Computer Vision (ECCV 2008), Marseille, 2008, 5303: 183-196.
- [22] Iqbal K, Yin X C, Yin X, et al. Classifier comparison for MSER-based text classification in scene images [J]. 2013:1-6.
- [23] WenHung Liao, YiChieh Wu. An integrated approach for multilingual scene text detection [J]. International Journal of Computer Information Systems and Industrial Management Applications, 2016, 8(1):33-41.
- [24] Neumann L, Matas J. Efficient Scene text localization and recognition with local character refinement [C]// International Conference on Document Analysis and Recognition. IEEE, 2015: 746- 750.
- [25] Liu J, Su H, Yi Y, et al. Robust text detection via multi-degree of sharpening and blurring[J]. Signal Processing, 2015, 124(C):259-265.
- [26] Chen, H., Tsai, S.S., Schroth, G., Chen, D.M., Grzeszczuk, R., Girod, B. Robust text detection in natural images with edge-enhanced maximally stable extremal regions [C]. 18th IEEE International Conference on Image Processing (ICIP), Brussels, BELGIUM, 2011.
- [27] Wang Q, Lu Y, Sun S. Text detection in nature scene images using two-stage nontext filtering[C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 2015: 106-110.
- [28] Rong X, Yi C, Yang X, et al. Scene text recognition in multiple frames based on text tracking[C]// IEEE International Conference on Multimedia and Expo. IEEE, 2014:1-6.
- [29] Wang Y, Shi C, Xiao B, et al. MRF based text binarization in complex images using stroke feature [C] // International Conference on Document Analysis and Recognition. IEEE, 2015: 821- 825.
- [30] Liu X, Lu T. Natural Scene character recognition using Markov Random Field[C]// International Conference on Document Analysis and Recognition. 2015:396-400.
- [31] Wen-Hung Liao, Yi-Chieh Wu. An integrated approach for multilingual scene text detection [J]. International Journal of Computer Information Systems and Industrial Management Applications, 2016(8), 33-41.
- [32] Chen Y, Huang L. Chinese Traffic Panels Detection and Recognition from Street Level Images [C] // MATEC Web of Conferences, 2016:06001.
- [33] Jiakai Gao, Lei Li, Lei Yang. A new text location method in natural scene images based on color

- reduction and AdaBoost [C]// Systems and Informatics (ICSAI), 2016.
- [34] Turki H, Halima M B, Alimi A M. Scene Text Detection Images With Pyramid Image and MSER Enhanced[C]// International Conference on Intelligent Systems Design and Applications. 2015.
- [35] Jufeng Liu, Linlin Huang, Boya Niu. Road sign text detection from natural scenes [J]. Information Science, Electronics and Electrical Engineering (ISEEE), 2014: 1547-1551.
- [36] Zhang Yang, Wang Chunheng, Xiao Baihua, et al. A New Method for Text Verification based on Random Forests[C]. International Conference on Frontiers in Handwriting Recognition, Bari, 2012: 109-113.
- [37] Huang Rong, Shivakumara P, Uchida S. Scene Character Detection by an Edge-Ray Filter[C]. IEEE International Conference on Document Analysis and Recognition, Washington DC, 2013: 462-466.
- [38] Xu, Hailiang, Xue, Like, Su, Feng. Scene Text Detection Based on Robust Stroke Width Transform and Deep Belief Network [J]. 12th Asian Conference on Computer Vision (ACCV), 2014, 9004: 195-209.
- [39] Zhang, Yong, Lai, Jianhuang, Yuen, Pong C. Text string detection for loosely constructed characters with arbitrary orientations [J]. Neuro computing, 2015, 168: 970-978.
- [40] Shi B, Bai X, Yao C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, PP(99):1-1.
- [41] Michael Opitz, Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig. End-to-End Text Recognition Using Local Ternary Patterns, MSER and Deep Convolutional Nets [J]. Document Analysis Systems (DAS), 2014: 186-190.
- [42] Ankush Gupta, Andrea Vedaldi, Andrew Zisserman. Synthetic Data for Text Localisation in Natural Images [J]. Computer Vision and Pattern Recognition (CVPR), 2016: 2315-2324.
- [43] Veit A, Matera T, Neumann L, et al. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images [J]. 2016, arXiv:1601.07140v2.
- [44] Gupta A, Vedaldi A, Zisserman A. Synthetic Data for Text Localisation in Natural Images [J]. 2016, arXiv:1604.06646v1.
- [45] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition [J]. Eprint Arxiv, 2014.
- [46] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading Text in the Wild with Convolutional Neural Networks[J]. International Journal of Computer Vision, 2016, 116(1):1-20.
- [47] Mishra A, Alahari K, Jawahar C V. Scene Text Recognition using Higher Order Language Priors[C]// 2012.

- [48] Yao C. Detecting texts of arbitrary orientations in natural images[C]// Computer Vision and Pattern Recognition. IEEE, 2012:1083-1090.
- [49] Wang K, Babenko B, Belongie S. End-to-end scene text recognition[C]// International Conference on Computer Vision. IEEE Computer Society, 2011:1457-1464.
- [50] Jung J H, Lee S H, Cho M S, et al. Touch TT: Scene Text Extractor Using Touchscreen Interface[J]. Etri Journal, 2011, 33(1):78-88.
- [51] Campos T E D, Babu B R, Varma M. Character Recognition in Natural Images[C]// Visapp 2009 - Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Lisboa, Portugal, February. DBLP, 2009:273-280.
- [52] Yi, Tian Y L. Text string detection from natural scenes by structure-based partition and grouping [J]. IEEE Transactions on Image Processing (TIP), 2011, 20(9):2594-2605.
- [53] Yin X C, Pei W Y, Zhang J, et al. Multi-Orientation Scene Text Detection with Adaptive Clustering [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1930.
- [54] 刘晓佩. 自然场景文本信息提取关键技术研究[D].西安电子科技大学,2014.
- [55] ICDAR 2013 Robust Reading Competition] [Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition[J]. 2013, 2(2-3):1484-1493.
- [56] 王永明,王贵锦.图像局部不变性特征与描述[M].国防工业出版社,2010:112-115.
- [57] Boris Epshtein, etc. Detecting Text in Natural Scenes with Stroke Width Transform [J]. CVPR, 2010: 2963-2970.
- [58] Zhang B. Y., Liu J. F., Tang X. L. Multi-scale Video Text Detection based on Corner and Stroke Width Verification [J]. Visual Communications and Image Processing, Kuching, 2013:17-20.
- [59] Oh I S, Lee J S. Smooth Stroke Width Transform for Text Detection [M]// Artificial Intelligence: Methodology, Systems, and Applications. 2016.
- [60] Su F, Xu H. Robust seed-based stroke width transform for text detection in natural images[C]// International Conference on Document Analysis and Recognition. 2015: 916- 920.
- [61] Boris Epshtein, etc. Detecting Text in Natural Scenes with Stroke Width Transform [J]. CVPR, 2010: 2963-2970.
- [62] Nister, D, Stewenius, H. Linear Time Maximally Stable Extremal Regions[C]. 10th European Conference on Computer Vision (ECCV 2008), Marseille, 2008, 5303: 183-196.
- [63] 孙君顶等.图像特征提取与检索技术[M].北京:电子工业出版社,2015.
- [64] John Z M. An information theoretic approach to content based image retrieval [D]. Baton Rouge: Louisiana State University and Agricultural and Mechanical College, 2000.
- [65] Pass G, Zabih R, Miller J. Comparing images using color coherence vectors[C]. Proceedings of the fourth ACM international conference on multimedia, 1997: 65-73.

- [66] Huang J. Color-spatial image indexing and applications [D]. New York Cornell University, 1998.
- [67] Stricker M, Orengo M. Similarity of color images[C]. Proceedings of SPIE Storage and Retrieval for Image and Video Database, 1995, 2420:381-392.

致谢

时光荏苒，岁月如梭，三年的研究生生活一闪而逝。还记得三年前复试的那天，天空下着小雨，带着一丝不安和期待，我来到了山水甲天下的桂林。跟北方完全不同的感觉，潮湿的空气，碧绿的草木，就像来到了原始森林。在见到导师缪裕青的那一瞬间，就被她的气质折服。也许是看出了我的忐忑和不安，在面试的过程中，缪老师一直微笑着，使我渐渐放松下来。在之后的三年里，让我一直庆幸的就是当初选择来桂林面试，遇到了缪老师。

在读研期间，跟着缪老师学到了很多。在此非常感谢缪老师，她严谨的治学态度、渊博的理论知识、一丝不苟的工作作风，深深激励着我，让我不断前行。本论文从选题、开题、成稿到结题，缪老师都倾注了无数的心血。未来我将以缪老师为榜样，时刻反省自己，不断提升自己。

感谢计算机与信息安全学院及研究生院的各位领导、老师和同学的关心和帮助。感谢研究生院 14 级 9 班的全体同学，感谢你们三年来的陪伴。感谢实验室 3404 的邵其武师兄、谢益均师兄、高韩师兄、同门欧威健、邹巍师弟、汪俊宏师弟、缪永进师弟、王洪磊师弟、凌海彬师弟，还有刘哥刘同来和师姐张万桢，他们在我的研究生生涯中给予了我很多的帮助和支持，尤其是在论文书写过程中，给予我信心。

感谢室友姜潇薇、王辉、贺秋歌、朋友项敏、王荣等，与我一起分享学习和生活中的喜怒哀乐。

感谢我的父亲刘建龙和母亲白建媛，感谢您们的养育之恩，在我失意时给我鼓励，在我得意时教会我胜不骄败不馁，感谢您们在我的人生道路上付出的所有心血。同时也感谢我的妹妹刘永清给我的支持和鼓励。

最后，谨以此文献给所有关心、支持我的亲人、老师、同学和朋友们！

攻读硕士学位期间的主要研究成果

发表论文

- [1] 缪裕青, 刘水清, 张万桢, 欧威健, 蔡国永. 自然场景图像中的中文文本检测算法研究[J]. 计算机工程与设计, 2018. 3.
- [2] 缪裕青, 欧威健, 刘同来, 刘水清, 文益民. 基于情感极性与 SMOTE 过抽样的虚假评论识别方法[J]. 计算机应用研究, 2018. 9.

申请软著

- [1] 在线虚假商品评论特征提取系统 V1.0 软件著作权. 登记号: 2017SR090219, 授权日: 2017 年 3 月 24 日.
- [2] 多平台用户信息融合研究系统 V1.0 软件著作权. 登记号: 2016SR225149. 授权日: 2016 年 8 月 18 日.
- [3] 瓶灌装饮用水数据采集系统 V1.0 软件著作权. 登记号: 2016SR225144. 授权日: 2016 年 8 月 18 日.

基金项目与获奖情况

- [1] 国家自然科学基金. 基金号: 61363029.
- [2] 广西自然科学基金. 基金号: 2014GXNSFAA118395.
- [3] 桂林电子科技大学研究生教育创新计划. 项目号: 2016YJCX72.
- [4] 第五届“中国软件杯”大学生软件设计大赛 全国总决赛三等奖.