分类号_____

U D C _____ 编 号 _____

成都理工大學

学 位 论 文

题名和副题名 OCR技术在简牍图像数字化中的应用

作 者 姓 名 刘 瑛

指导教师姓名及职称 王绪本 教授

申请学位级别 硕士 专业名称 信号与信息处理

论文提交日期 2007.11 论文答辩日期 2007.12

学位授予单位和日期 成 都 理 工 大 学（2007 年 12 月）

答辩委员会主席

评阅人

2007 年 11 月

分类号_____ 学校代码：10616

UDC _____ 学号：硕21-112

成都理工大学硕士学位论文

OCR技术在简牍图像数字化中的应用

刘 瑛

指导教师姓名及职称 王绪本 教授

申请学位级别 硕士 专业名称 信号与信息处理

论文提交日期 2007.11 论文答辩日期 2007.12

学位授予单位和日期 成 都 理 工 大 学（2007 年 12 月）

答辩委员会主席

评阅人

2007年 11 月

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__成都理工大学__或其他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。

学位论文作者签名：

2007年 12月 4日

学位论文版权使用授权书

本学位论文作者完全了解__成都理工大学__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__成都理工大学__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适用本授权书）

学位论文作者签名：

学位论文作者导师签名：

2007 年 12月 4日

OCR技术在简牍图像数字化中的应用

作者简介：刘瑛，女，1978年09月出生，2001年09月师从于成都理工大学王绪本教授，于2007年 12 月获硕士学位。

摘 要

简牍作为中国古代重要的一种书写材料，保存了丰富的历史文化信息，是我国宝贵的文化财富。简牍文献的数字化处理是简牍保护及其信息检索的重要研究内容，由于简牍图像具有干扰噪声大、部分文字不再使用的特点,目前市面上流行的文字识别软件很难适用于简牍图像的文字识别, 这给简牍文献的数字化处理工作带来很大困难。本文深入研究简牍图像特点，在简牍图像数字化处理的各阶段进行了大量实验，提出了一系列适用于简牍图像的文字处理算法：

（1）与一般文字图像相比，简牍图像的背景存在明显灰度差异，常用的二值化算法很难有效区分文字与简牍背景，针对该问题，本文提出八邻域灰度差值算法。该方法考虑到文字笔划与背景之间灰度差异要大于背景之间灰度差异的现象，先求出在文字图像内具有八邻域最大灰度差值的像素灰度值，然后根据该灰度值设计文字的二值化阈值。实验表明，八邻域灰度差值算法能有效地从简牍图像的复杂背景中提取出文字，并有效减少竹简边沿噪声的引入。

（2）本文提出了适合于简牍图像的文字切分算法。该文字切分算法先利用垂直投影切分出图像中的各列文字，再就每列文字进行水平投影切分出图像中的单个汉字，针对简牍的边沿条状噪声、节点噪声及腐蚀受损噪声等，综合运用文字合并、外延扩展、剔除竹简边沿、去除大噪声点等方法，实现了文字的正确切分。经实验验证该算法具有运算速度快、易于实现、不受竹简边沿噪声影响的特点。

（3）本文提出了一种快速有效的孔洞特征提取方法。该方法先利用孔洞填充算法填充文字外围空白区域，再依次填充文字内部各孔洞区域，直至文字图像中不存在空白区域。实验表明该方法能有效识别简牍文字中孔洞数量，必要时可以求出孔洞位置信息。

（4）本文改进了形态学文字细化算法。针对文字细化过程中因像素误删、导致文字连通性破坏和关键信息丢失的问题，增加两个保留模板;尤其针对两个像素宽度笔划出现断裂问题，采用数组记录保留像素点坐标，在细化迭代计算过程中查询该数组避免保留像素被删除。实验表明，改进的文字细化算法有效地保

持了原有文字的连通性。

（5）本文实现了中值滤波算法，以简牍图像为样本，对滤波参数进行整定，达到有效滤除简牍图像中椒盐噪声的目的；本文采用了双内插值算法对单字图像进行归一化处理，实验结果表明双内插值算法对图像缩放处理时可以保留大部份文字信息。

（6）本文初步研究几个常用的文字特征，选取在简牍图像中表现较为稳定的孔洞、特征点及水平垂直投影等特征，并实现上述文字特征的提取算法。

本文研究过程中注重理论结合实践，以应用为目的，充分协调项目的进度、资源及质量，选取下列平台作为研究的基础：1）以Internet、中文科技期刊数据库、中国优秀硕博士论文数据库及学校图书馆作为信息来源；2）以X86桌面PC作为通用硬件平台；3）以Windows XP作为开发、应用软件平台；4）开发环境采用Visual Studio .NET，软件框架采用MFC，图像库选用Paintlib，编程语言选用C++。通过选定上述研究平台，使研究成果可以快速转化为应用，同时有效减少辅助性质的工作量，使作者可以专注于技术难点、重点的研究。

关键词：简牍 文字识别 二值化 文字切分 细化

The application of OCR technology to bamboo scripts image digitization

Introduction of Author: Liuying, female, born in Sept. of 1978, was granted the master from chengdu university of technology whose tutor was professor Wangxuben.

Abstract

Bamboo scripts preserves the rich historical and cultural information. As a writing material in ancient China, Bamboo scripts is the valuable cultural wealth. Digital data processing of Bamboo scripts is an important method for Bamboo scripts protecting and retrieving, but the existing character recognition software can not be applied to the character recognition of bamboo scripts ,because bamboo scripts image has such characteristics as it has major sources of noise interference and some characters used by bamboo scripts have never been used any more. This paper designs one kind of bamboo scripts character recognition system according to bamboo scripts image' s characteristics and achieves the following specific functional modules:

Bamboo scripts preserves the rich historical and cultural information. As a writing material in ancient China, Bamboo scripts is the valuable cultural wealth. Digital data processing of Bamboo scripts is an important method for Bamboo scripts protecting and retrieving, but the existing character recognition software can not be applied to the character recognition of bamboo scripts because of bamboo scripts image' characteristics such as it has major sources of noise interference and some characters used by bamboo scripts have never been used any more, and this causes great difficulties to digital data processing of Bamboo scripts . This paper proposes a series of word processing algorithm suitable for bamboo scripts image after lucubration in the features of bamboo scripts image and a large number experiments done in all stages of digital image processing of Bamboo scripts.

(1) Compared with the general text images, there is obvious gray changes on the back ground of Bamboo scripts image, and this makes it very difficult to

distinguish word and background for existing binary processing algorithms. According the feature of bamboo scripts image, this paper designs 8 Gray Margin of Neighborhood. This method considers the gray differences between word and background is greater than between background, this method firstly calculates the gray value which has the largest margin of gray-scale pixel within 8 neighborhood, secondly designs the threshold value of gray, lastly sets pixel' gray value. Experiments show that 8 Gray Margin of Neighborhood can extract word from complex background and reduce edge noise effectively .

(2)This paper proposed word segmentation algorithm suitable for Bamboo scripts image. The first character segmentation algorithm cuts out words of each series making use of vertical projection map ,and then acquires the general location of words making use of level projection map .Several compensation measures such as text merger, outward expansion ,noise removing are used aimed at bamboo edge noise, nodes noise and corrosion noise.

(3) This paper presents a fast and efficient method used to seek holes feature. This method firstly fills gaps in word outlying regions using hole filling algorithm, and then fills holes within word until there is no blank region any more. Experiments show that this method can find out how many holes in one word effectively and these holes' location.

(4) Improving existing morphological thinning algorithm .This paper adds two reservation template aimed at such phenomena as text connectivity is destruct and key information loses because of pixel is deleted by mistake ;this paper makes use of array note pixel's location and inquiries array to avoid reserved pixel is deleted in allusion to the phenomenon stokes two pixel wide fracture. Experiments show that improved thinning algorithm maintain the original text connectivity commendably.

(5)This paper carries out median filtering algorithm, uses bamboo scripts image as sample, tunes filter parameter to filter salt and pepper noise; this paper unifies character to the same size using bilinear interpolation method ,and experiments show that most textual information is remained .

(6)This paper studies several common text feature preliminary, and selects

several stable features such as holes feature, feature points ,level projection feature and vertical projection feature. These features have been obtained.

This paper focuses on combining theory and practice, aims at application, coordinates relations between the process of the project, resource and quality, selects following platforms as the base of study:1) making full use of several resources such as Internet, VIP,CDMD and Chengdu University of Technology Library;2) using X86 desktop PC as general hardware platform;3)using Windows XP as software development and application platform;4) using Visual Studio .NET as software development tools, MFC as software framework, Paintlib as image manipulation library, and C++ as programming language. Through the setting of above platforms, research results can be translated into application relatively quickly, Auxiliary work reduces effectively ,and the author can concentrate on key issues.

Keywords: bamboo scripts, character recognition, binarization, character segmentation, thinning

目 录

6

第一章 引言

1.1 研究背景

中华民族创造了无比丰富的历史文化遗产，对于世界文化具有巨大贡献。古代典籍是中国历史文化遗产最为重要的物质载体，记载了古代思想、文化、生活、习俗及军事的方方面面，是了解古代思想文化的主要途径。中国古代留下丰富的典籍资料，据杨家骆先生1946年统计，仅西汉至清末的古籍就有181755部[1]，如果包括古代的甲骨、简牍，将是非常庞大的数目，仅1996年于湖南长沙发现的走马楼三国吴简，就有14万余枚。面对浩如烟海的古代典籍，相关信息的检索成为极其困难的事情，因此胡适之先生提出，必须系统地整理古籍。此后，学术界编纂了多种引得、通检、索引、汇编等工具书，部分完成了索引式整理的目标，但是人们也发现，中国古籍数量极大，经过系统整理的毕竟只是少数，方便的检索工具也太少，即使是已有索引的古籍，在用来解决具体问题时仍会感觉到种种不便。为此，人们需要有经过认真梳理的、反映全部前人研究成果的古籍文本供学术界使用，需要有便捷、高效、准确的查询工具为人文学术研究服务[2]。同时，由于部分古代典籍为孤本、善本，这些“国宝”，由于年代久远，加之受存放空间、设备等条件所限，有些古籍已出现纸张脆弱，字迹变色，破损，书页脱落，以及发霉、虫蚀等现象，出土的甲骨、简牍也由于长期埋存地下，出现了严重的腐蚀及霉烂的状况，过多的对此类典籍进行查阅，会严重影响典籍的保存寿命。由于上述原因，古代典籍的数字化成为考古学界和人文学术研究对信息工程提出的迫切要求。

古籍数字化是指运用计算机技术，对古籍文献进行加工和处理，建立书目数据库、全文数据库和综合检索系统，并通过光盘、网络等途径进行传播[3]。古籍数字化的瓶颈问题在于如何将大量的古代典籍输入计算机，如果采用人工的方式录入这些信息，将会是一个极其漫长而繁重的工作，即费时又昂贵，因为古代典籍使用的众多汉字属于古代的写法，或是繁体字，或是异体字，或是不再使用的汉字，对于录入人员来讲，即要求他们能够快速录入，又要求有足够的古文知识，是很不现实的，在《四库全书》电子化工程的预研阶段，从书中抽取了180万标题，采用传统的录入—打印—校对—修改的方法，仅1000万字的材料，仅录入就

用了30人月，更困难的是之后的校对过程，每一次校对，就要打印将近2万页，交出别人同原书对照，先后7次校对，历时一年，最终结果的错误率仍未低于千分之一，而所进行的内容只占到全部工作量的70分之1。因此，如何将古代典籍资料快速录入计算机成为古籍数字化研究工作的重要内容。

1.2 OCR 文字识别技术

OCR指光学字符识别技术，是自动识别技术研究和应用中的一个重要领域。光学字符识别技术的工作原理是通过扫描仪或数码相机等光学输入设备获取纸张、甲骨或简牍上的文字图片信息，利用各种模式识别算法分析文字形态特征，判断出文字的标准编码，将图像中的文字转换成文本格式，并按通用格式存储在文本文件或者数据库当中，还可以利用文字处理或者编辑软件做进一步加工。

通常，一个OCR系统由以下几个部分组成，如图1-1所示：

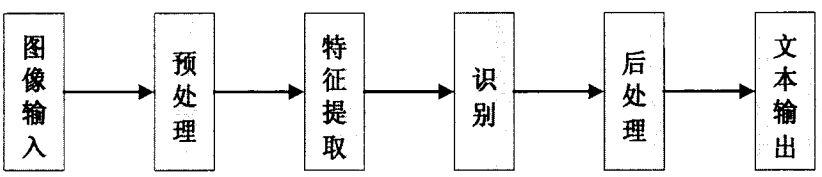


图1—1 OCR光学字符识别系统

图像预处理是OCR系统中需要解决问题最多的一个模块，是文字识别的重要一环，从输入图像获得的彩色图像，到分离出一个个的文字图像的过程，都属于图像的预处理。图像预处理模块可以把原始图像转换成特征提取所能接受的形式，消除一些与类别无关的因素（尺寸和位置的归一化），它包含了图像的灰度化、去噪、图像矫正等图像处理，以及图像分析、文字行与字分离的处理。由于一般在预处理后的图像上提取特征，如果这个步骤结果不理想，往往给后面的识别环节带来无法纠正的错误。

图像的预处理流程如下：

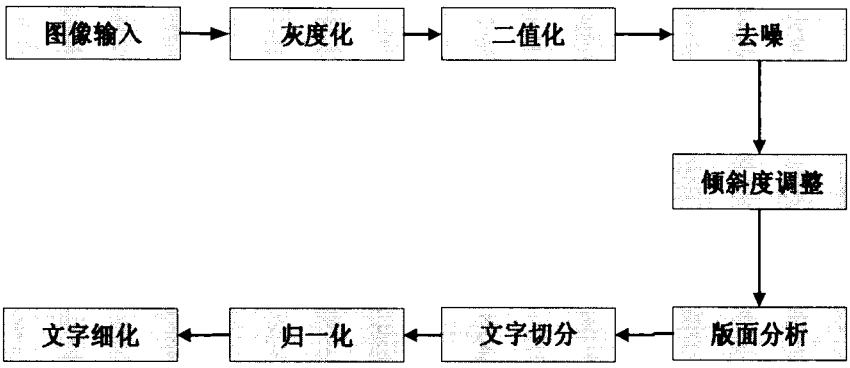


图1—2 OCR系统预处理模块

特征提取是文字识别的关键一步，用什么特征、怎么抽取，直接影响文字识别率，根据模式识别的统一熵理论，必须保证所提取的特征中含有足够的信息量才能获得较高的识别率。对于不同的识别问题，每种特征抽取方法所能满足要求的程度是不尽相同的。因此，对于一个特定的应用问题，哪些特征是最为有效的，一般都需要通过实验来评估。目前，在文字识别领域常用的特征有两类：

（1）统计特征

统计特征具有良好的抗噪声、抗干扰的性能。根据特征抽取区域的不同可粗略地分为全局统计特征和局部统计特征两大类：全局统计特征是将整个汉字点阵作为研究对象，从整体上抽取特征，主要包括全局变换特征，对汉字图像进行各种变换，利用变换系数作为特征，常用的变换有Wavelet变换，Fourier变换，DCT变换,，K-L变换等；全局笔画方向特征，这种特征反映了在整个汉字点阵中笔画的复杂度、方向及连接关系；背景特征，汉字图像的空白部分（即背景)和周围笔画的关系也含有一定的结构信息，提取背景点在各个方向的笔画密度作为背景特征；局部统计特征是将汉字点阵图像分割成不同区域或网格，在各个小区域内分别抽取统计特征，主要包括局部笔划方向线素特征和四角特征等[4]。

（2）结构特征

结构模式识别是早期汉字识别研究的主要方法。其主要出发点是汉字的组成结构。从汉字的构成上讲，汉字是由笔划(横竖撇捺等)、偏旁部首等结构单元所构成，由这些结构基元及其相互关系完全可以精确地对汉字加以描述，识别时，利用上述结构信息及其组成方法进行识别。用这种方法来描述汉字字形结构的主要优点在于对字体变化的适应性强，区分相似字能力强，但在实际应用中，使用结构特征构成的OCR系统文字识别率较低，抗干扰能力较差，这是因为在实际得

到的文本图像中存在着各种干扰，如倾斜、扭曲、断裂、粘连、纸张上的污点、对比度差等等，这些因素直接影响到结构基元的提取，假如结构基元不能准确地得到，自然不能准确的识别文字。此外结构模式识别的描述比较复杂，匹配过程的复杂度因而也较高[5]。

由于统计特征和结构特征各有优缺点，目前常用的文字识别系统是将统计特征和结构特征结合起来，利用统计特征进行文字的粗分类，运用结构特征对相似文字进行细分类。

分类识别就是利用事先建立好的特征字典库或识别规则，将待识别文字的特征与字典进行匹配，计算待识别字符跟字典中样本之间的距离，最后以相似度最大的类别作为识别结果输出。文字识别系统中采用的分类器有很多，如Bayes分类器、神经网络分类器等等。由于基于不同特征提取和匹配方法的各种分类器或多或少具有互补性，因此多个分类器的集成可以弥补单分类器的不足，提高总体性能[6]。

由于OCR的识别率无法达到百分之百，一些除错或辅助更正的功能，也成为OCR系统中必要的一个模块，字词后处理就是文字识别系统必不可少的一环。后处理就是利用词义、词频、语法规则或语料库等语言先验知识对识别结果进行校正的过程。

汉字的识别最早可以追溯到20世纪60年代。1966年，IBM公司的Casey和Nagy发表了第一篇关于印刷体汉字识别的论文，在这篇论文中他们利用简单的模板匹配法识别了1000个印刷体汉字。70年代以来，日本学者做了许多工作，其中有代表性的系统有1977年东芝综合研究所研制的可以识别2000汉字的单体印刷汉字识别系统；80年代初期，日本武藏野电气研究所研制的可以识别2300个多体汉字的印刷体汉字识别系统，代表了当时汉字识别的最高水平。此外，日本的三洋、松下、理光和富士等公司也有其研制的印刷汉字识别系统。这些系统在方法上，大都采用基于K-L数字变换的匹配方案，使用了大量专用硬件，其设备有的相当于小型机甚至大型机，价格极其昂贵，没有得到广泛应用。

我国对印刷汉字识别的研究始于70年代末、80年代初，大致可以分为三大阶段：

(1) 第一阶段从70年代末期到80年代末期，主要是算法和方案探索。

(Z) 第二阶段是90年代初期，中文OCR由实验室走向市场，初步实用。

(3) 第三阶段也就是目前，主要是汉字识别技术和系统性能的提高，包括汉

英双语混排识别率的提高和稳健性的增强。

同国外相比，我国的汉字识别研究起步较晚。但出于我国政府对汉字自动识别输入的研究从80年代开始给予了充分的重视和支持，经过科研人员十多年的努力，印刷体汉字识别技术的发展和应用，有了长足进步：从简单的单体识别发展到多种字体混排的多字体识别，从中文印刷材料的识别发展到中英混排印刷材料的双语识别。各个系统可以支持简、繁体汉字的识别，解决了多体多字号混排文本的识别问题，对于简单的版面可以进行有效的定量分析，同时汉字识别率达到了98%以上[7]。

1.3 目前技术存在的问题

虽然我国在汉字识别领域取得了长足的进展，但目前所使用的OCR文字识别系统很难适用于古代典籍的文字识别工作，主要原因如下：

（1）由于年代久远，许多古代书籍出现了字迹变色、破损、发霉、虫蚀等现象，而甲骨、简牍长期埋藏于地下，在出土时同样出现了字迹变色、发霉的现象，发霉、虫蚀的部分对于文字识别系统而言相当于大噪声点，如果这些大噪声点不能在预处理阶段删除，会被后续模块误认为是文字组成部分进行识别，影响文字识别的结果；当文字出现较严重的字迹变色时，由于输入图像是经过灰度化和二值化才进行文字识别，若字迹颜色变浅会被误识为背景噪声而被删除，导致二值化后的文字图像出现缺失笔画的现象，进而导致文字错误识别。现有OCR文字识别系统大多针对噪声较小的文字图像，很难应用于古代典籍这种干扰噪声较大的图像。

（2）中国古代使用繁体字进行书写，而现代中国使用的是简体字，部分人文工作者已不熟悉繁体字，如果古代典籍的OCR文字识别系统能够自动将繁体字转换为简体字，无疑会方便人文工作者开展研究工作。实际上，对于文字识别系统，在识别出文字图像中的繁体文字后，只需将繁体字的文字代码转换为对应的简体汉字的国标码，即可实现繁体汉字与简体汉字之间的文字转换。

（3）部分繁体字已不再使用，因此在繁体字所使用的Big5码和GB/T 12345-90辅助集中无法找到相应的文字编码，需要文字识别系统针对这类文字进行文字编码，显然，现有的文字识别系统无法实现该功能。

1.4 本文的主要工作

面对书写载体种类众多的古代典籍（甲骨、青铜器、石刻、陶器、砖瓦、兽皮、简牍、锦帛、纸张），本文作者选择简牍作为研究的主攻方向，原因如下：

（1）简牍，用竹子或木材做成的书写载体，其中单个的竹（木）片叫简，较宽的竹（木）片叫牍。简牍被作为书写材料始于殷商时期，在公元前770年至公元前221年的春秋战国时期成为当时的主要书写载体,直至公元4世纪的东晋，才由于纸张的广泛使用而退出历史，共1600余年，涉及内容非常广泛，有经书、子书、历谱、公文、律令、账簿、遗策等，是我国宝贵的文化遗产，对于简牍的研究也是学术界的重要内容。然而，目前已有的OCR文字识别系统，基本上都针对以纸张作为书写载体的古代典籍，针对简牍的文字识别系统研究较少。

（2）简牍数目极其庞大，需要有计算机辅助系统帮助考古学者进行研究。湖南长沙的马王堆汉简于1972年出土了610枚竹简，直至1981年才完成对竹简上文字的考释,我国科学工作者于1996年在湖南长沙走马楼发现14万余枚吴简，2002年又在湖南湘西龙山发现2万余枚秦简，我国已有简牍的数目达到了26万余枚，对这26万余枚简牍的考释将会是极其困难的工作，因此考古工作者迫切希望能够引入计算机信息处理技术帮助进行考古研究。

（3）长沙简牍博物馆与我校合作，提出了国家十五科技攻关项目，研究简牍图像的信息处理技术，该项目完成了建立简牍数据库的工作，但对于简牍图像的自动识别还没有涉及，因此，长沙简牍博物馆希望我校继续进行简牍文字的自动识别工作。

根据OCR文字识别理论和简牍图像的特点，设计简牍图像文字识别系统如下：

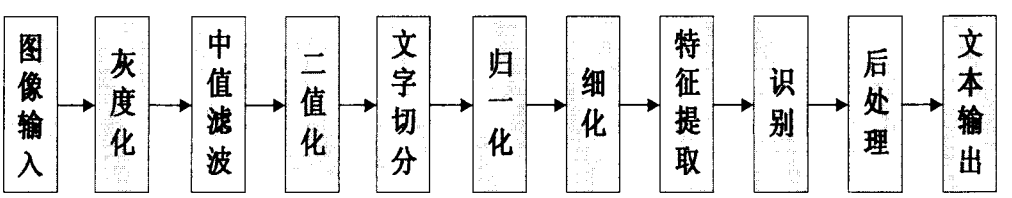


图1—3 简牍图像文字识别系统

本文的第二章介绍了简牍OCR系统的灰度化实现方法。

第三章首先介绍了中值滤波算法的原理，然后针对简牍图像特点，通过大量实验设计了简牍图像中值滤波算法的参数。

第四章在仔细分析已有图像二值化方法的基础上，根据文字图像特点，改进已有图像二值化算法，并在编程实验的基础上验证了改进算法的有效性。

第五章根据简牍图像特点，设计一种分步文字切分算法，先利用灰度直方图获得文字的大致位置，利用后续补偿措施，获得精确的文字位置信息，同时去除竹简边沿噪声。

第六章针对已切分出的单个文字，运用比较成熟的文字归一化算法，将原本大小不一的简牍文字，统一成相同的大小40×40。

第七章在研究已有的文字细化算法的基础上，改进文字细化算法，并对改进后的文字细化算法做实验验证文字细化效果。

第八章针对文字特点，提出几种文字特征提取算法，并设计了相应的软件实现算法。

最后对本文所做工作做出总结，并提出后续的研究方案。

在软件设计过程中，本文采用的软件开发工具为Visual C++（VC++），原因在于：（1）C++语言具有灵活度性好、效率高、运行速度快、可移植性强的特点，VC是微软制作的产品，具有强大的MFC库；（2）实际可以交付给考古人员使用的软件程序是由编程语言所生成的可执行文件，只有在进行算法实验时使用编程语言进行算法的验证，才能确切知道实际使用过程中算法所需要的计算时间及内存容量的大小，对软件及算法进行合理的设计。

第二章 灰度化处理

2.1 作用

自然界中的所有颜色都可以由红（R），绿(G)，蓝(B)组合而成。有的颜色含有红色成分多一些，如深红；有的含有红色成分少一些，如淡红。针对含有红色成分的多少，可以分成0到255共256个等级，0级表示不含红色成分，255级表示含有100%的红色成分。同样，绿色和蓝色也被分成256级。这种分级的概念被称作量化。这样，根据红、绿、蓝各种不同的组合我们就能表示出$256\times256\times256$，约1千6百万种颜色。数码相机所拍摄的图像存储格式一般为24位真彩色BMP格式或JPEG格式的位图，24位真彩色BMP格式用3个字节（即24bit）表示一个像素点的彩色值，其中R、G、B分量各占一个字节，可表示$256\times256\times256$种颜色，JPEG图像存储格式是一种比较成熟的有损压缩格式，文件扩展名为jpg或jpeg,在JPEG文件中，RGB色彩格式需要先根据公式$Y=0.299R+0.587G+0.114B$、公式$U=-0.147R-0.169G+0.5B$和公式$V=0.5R-0.4187G-0.0813B$转换为YUV的格式，其中Y分量表示亮度信息，UV分量表示色差信息，然后进行DCT变换、量化和编码，因此在JPEG格式文件中是不能直接获得图像的R、G、B数值的，而目前大多数的图像处理算法是针对颜色的R、G、B分量进行设计的，因此在进行数字图像处理时通常先将JPEG格式文件转换为BMP格式文件。

彩色图像虽然能够比较好的保存图像信息，但在图像处理过程中需要对每个像素点的R、G、B三个分量进行运算处理，运算量大，运算时间长，而简牍图像文字识别系统的目的在于识别出图像中的文字，文字或简牍的颜色并不是主要的考虑对象，因此，在设计系统时选择先将简牍图像进行灰度化，以减小后续算法的运算量。

将彩色图像转化成为灰度图像的过程称为图像的灰度化处理。

灰度图像是R、G、B三个分量取值相同的一种特殊的彩色图像，其一个像素点的变化范围为256种，在数字图像处理中一般将各种格式的图像转变成灰度图

像以减小后续图像处理算法的计算量，而灰度图像仍然反映了整幅图像的整体和局部的色度和亮度等级的分布和特征[8]。

2.2 方法

设 $R=G=B=Y$，其中Y被称为灰度值，它应位于某个范围之内：

$$Y_{\min} \le Y \le Y_{\max} \qquad (2\text{--}1)$$

理论上要求Y仅为正的，且为有限值，区间 $[Y_{\min}, Y_{\max}]$ 称为灰度级，一般常用灰度级为[0，255]，这里 $Y_{\min}=0$ 为黑，$Y_{\max}=255$ 白，所有中间值是从黑到白的各种灰色调[9]。

彩色图像可以由公式（2-2）变为灰度图像：

$$Y = 0.299R + 0.587G + 0.114B \qquad (2\text{-}2)$$

2.3 本文采用方法

由于灰度化处理已有很成熟的算法并且有许多应用软件面世，因此在系统设计过程中本文利用已有软件工具包paintlib实现简牍图像的灰度化处理。

paintlib是一个可移植的用于图像加载、保存和处理的C++类库，支持BMP, GIF, IFF, JPEG, PCX, PGM, PICT, PNG, PSD, SGI, TGA, TIFF 和 WMF等格式文件的读取，并且以BMP, JPEG, PNG 和 TIFF格式进行存储。Paintlib于1995年由Zadow作为开源软件项目提出，经过10年世界各地图像处理爱好者的努力，paintlib在功能上日臻完善；同时，由于paintlib为开源软件，公布全部代码，在实际使用过程中可以根据实际需要进行裁减或更改，所以本文选择paintlib作为简牍文字识别系统的支持类库，实现简牍图像的读入和灰度化处理。实验中采用的版本为paintlib-2.6.2。

2.4 灰度化效果

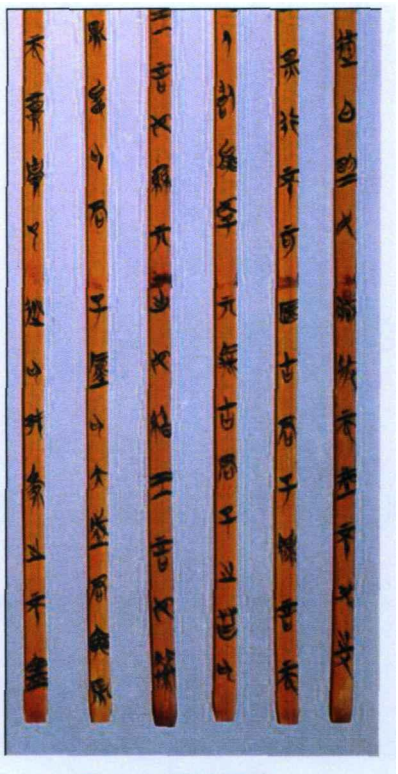


图 2-1 简牍图像

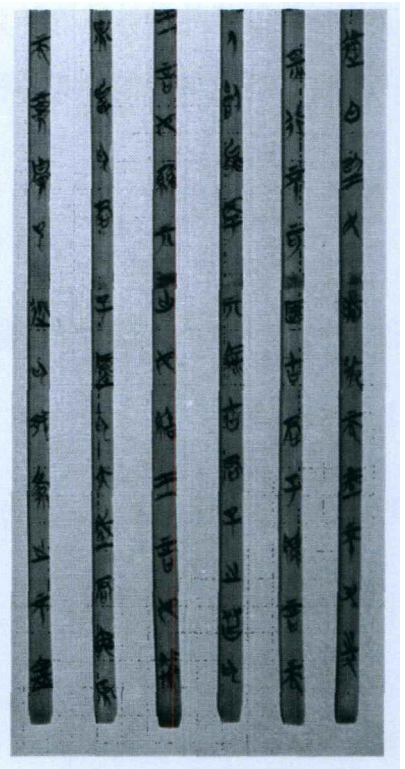


图 2-2 灰度化图像

图2-1为输入系统的彩色简牍图像，图2-2为经过灰度化处理后的图像，由对比可知，灰度化后的图像比较好的保存了原图像中色度和亮度的分布信息。

第三章 中值滤波

3.1 作用

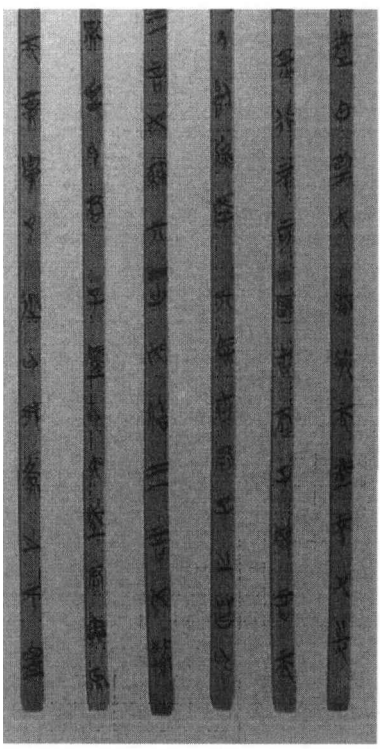


图 3-1 灰度化图像

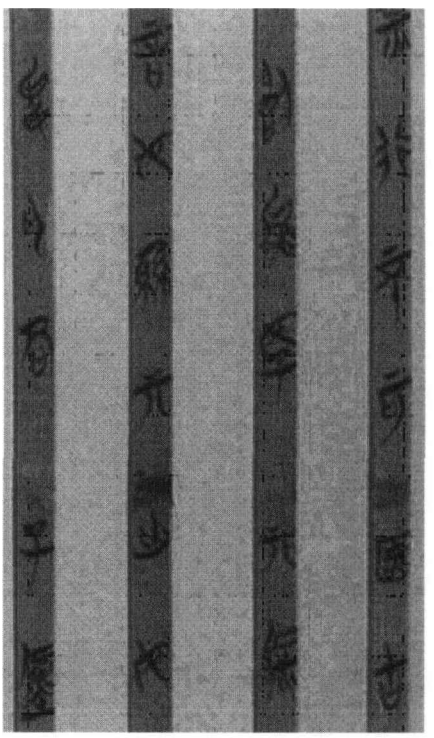


图 3-2 部分放大后的灰度图像

仔细研究灰度化后的简牍图像，会发现在图像上存在一些很小的灰色点，将灰度图像中灰点的位置与原彩色图像对比后发现，灰色点的出现原因在于：由于简牍实物长期埋藏于地下，竹简（木牍）出现了腐蚀现象，这些灰色点就是在简牍实物文物保护处理过程中无法消除的霉点。

这些小的灰色点，一部分可以在后续的二值化处理过程中被除去，但也有相当一部分由于颜色较深，无法在二值化处理过程中消除，这些小的灰色点会影响文字识别的效果，因此有必要将这些灰色点去除。

由于这些灰色点出现具有不规则性，且大部分是以单个点的形式出现，或以较小的数目（2个或3个）成群出现，因此可以将这些灰色点看成椒盐噪声，而中值滤波算法在处理椒盐噪声时具有去噪效果好、软件实现容易的优点。

3.2 原理

3.2.1 中值滤波算法

中值滤波是一种非线性平滑法，它对一个滑动窗口内的像素灰度值排序，并用其中值代替窗口中心像素的灰度值。对脉冲干扰及椒盐噪声具有良好的抑制作用，并且在抑制随机噪声的同时能有效保护文字边沿边缘，防止出现模糊现象。二维中值滤波的窗口形状有多种，如线状、方形、十字形、圆形、菱形等。不同形状的窗口产生不同的滤波效果，一般情况下，多选择方形窗口。

P0	P1	P2
P3	P4	P5
P6	P7	P8

图 3-3 3X3 滑窗

图3-3为3×3的方形滑窗，在滑窗滑动时，需要确定灰度值的像素点位于滑窗的中间位置，即P_4。为了求得P_4点的灰度值，需要查看P_0~$P_8$9个像素的灰度，按照从小到大的顺序排列，这时的中间值应该是排序后9个像素中第5个的灰度值。

4	5	5
3	200	6
5	4	5

图 3-4 示例一

如图3-4所示的灰度图像数据，为了要求O所围的像素点的灰度值，将9个像素点的灰度值从小到大排列，即3、4、4、5、5、5、5、6、200，其中第5个灰度值为5，因此将P_4点的像素灰度值设为5，P_4原来的灰度值为200，接近于白色，

而其周围8邻域像素点的灰度值均小于7，接近黑色，因此可以判定P_4点应为噪声点，而中值滤波后P_4点的灰度被置为5，接近黑色，噪声点被消除了。

实际上，由于椒盐噪声的灰度值与周围像素的灰度值相差很远，在按照大小排列时会集中排列在数列的左端或右端，因此选取数列的中间值作为P_4的灰度，可以去除椒盐噪声。

3.2.2 加权中值滤波算法

加权中值滤波算法在1991年提出，此滤波算法是由中值滤波算法改良而来，不但可以去除噪声，还可以较好地保留图像细节。中央加权中值滤波算法的处理步骤跟中值滤波算法很相似，同样先设定长宽一致的滑窗，对滑窗内中央点P_4复制 w次，然后排序输出中间值，取w等于0时，加权中值滤波算法就变成了中值滤波算法[10]。

对于图3-4所示的灰度图像数据，设w=2，则按大小进行排列后的数据为3、4、4、5、5、5、5、6、200、200、200，这时位于数列中间位置的灰度值为5，与中值滤波算法得出效果相同。

100	112	115
130	120	110
100	134	132

图 3-5 示例二

图3-5所示灰度数据，按照中值滤波算法，求得P_4处像素的灰度值为115，而分析P_4周围8邻域像素的灰度值可以发现，该点并不是噪声点，只是与周围像素间存在灰度差而已，经过中值滤波后，该点的像素灰度由120变为115，虽然在人眼看来，灰度值差5区分并不明显，但如果在进行二值化处理过程中，二值化阈值设在115与120之间时，原本P_4点应被置为白点，经过中值滤波后却被置为了黑点。若采用加权中值滤波算法，并设w=2，由小到大排列像素的灰度，得到的灰

度数列为100、100、110、112、115、120、120、120、130、134、135，数列中间位置的灰度为120，为P_4点原本的像素灰度值。

加权中值滤波算法效果的好坏，在很大程度上取决于w的选择，当w太大时，P_4点处的灰度值会以很大概率出现在灰度数列的中间位置，会降低椒盐噪声点被重置灰度的概率，导致中值滤波的效果不好；当w太小时，P_4点处的灰度值出现在灰度数列中间位置的概率会减小，该点灰度重置的概率增加，滤除椒盐噪声的概率增加，滤波效果变好，但同时也会导致图像文字边沿变模糊，图像细节信息丢失过多，因此w的选择是中值滤波算法设计的关键。

3.3 实验效果分析

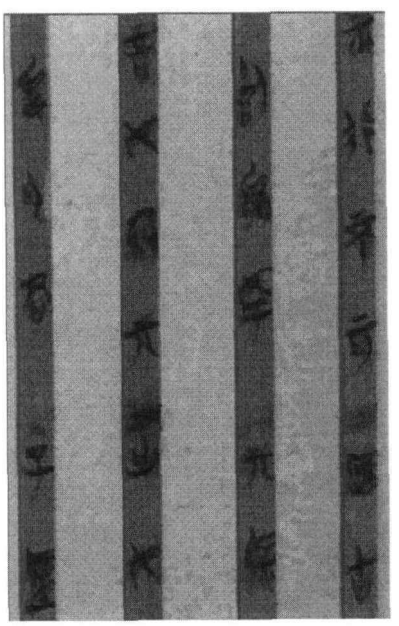


图 3-5 w=2 中值滤波效果

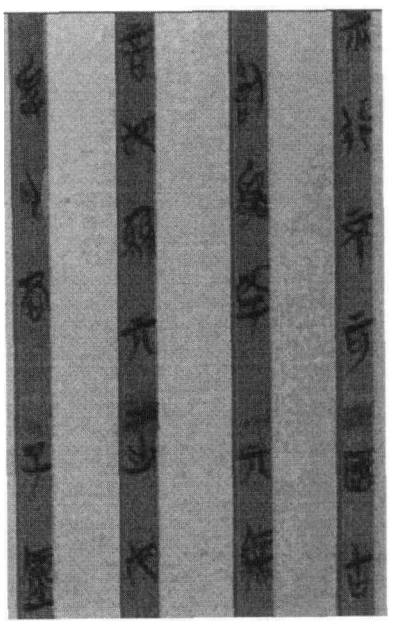


图 3-6 w=4 中值滤波效果

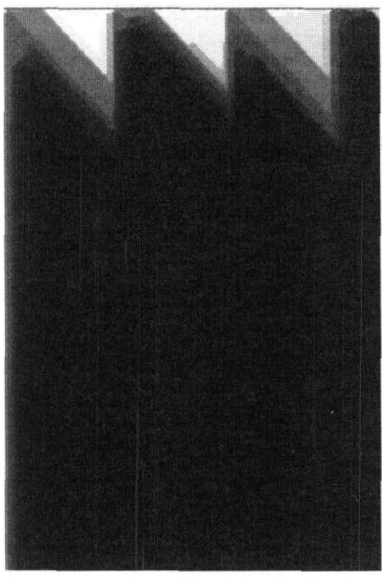


图 3-7 w=6 中值滤波效果

图3-5~图3-7是分别w取2、4、6时进行加权中值滤波后所得结果，由图3-5知，当w取2值时，椒盐噪声基本被滤除，但同时文字笔画变得模糊；图3-6是当w取4值时，椒盐噪声基本被滤除，文字的笔划边沿也比较好的保存下来，模糊较小；图3-7是当w取6时，中值滤波后的图像已不能显示出文字图像。至于w=1,3,5,7情况，由于在w取奇数值时，灰度数列的中间位置有两个，因此在实践过程中较少采用。

由实验效果可知，当w=4时，加权中值滤波不仅能够较好的滤除椒盐噪声，而且文字的边沿也较好的保存了下来，因此，在实际设计简牍文字识别系统时，使w值为4。

第四章 二值化

4.1 作用

图像二值化是指图像上所有点的灰度值只有二种可能，不是“0”，就是“255”，即把灰度值超过某一阈值的像素赋以最大灰度值255,其余像素则赋予最小灰度值0[11]。这样有利于在对图像做进一步处理时，图像的集合性质只与像素值为0或255的点的位置有关，不再涉及像素的多级灰度值，使处理变得简单，而且数据的处理和压缩量小。

实际上，对于简牍图像而言，二值化处理还有第二个作用：去除简牍背景，将简牍上面的文字分离出来。对于简牍文字识别系统而言，其目的在于提取出来简牍上面的文字并加以识别，至于竹简（或木牍）的情况，不在识别系统的考虑范围之内，而且竹简（或木牍）的存在，还会使后续的文字切分模块变得异常复杂。

4.2 常用方法

二值化处理的关键在于阈值T的选择，一旦确定了阈值，就可以将多灰度级图像g(x,y)在每一点的值分别与T进行比较。若g(x，y)≥T则将其灰度值置为1，反之则将该点的灰度值置为0，从而形成新的二值图像[12]。

选择阈值的方法有很多种：全局阈值法、局部阈值法、动态阈值法等。全局阈值法根据图像的直方图或灰度空间分布确定一个阈值，以此实现灰度图像到二值图像的转化，典型的全局阈值方法包括Ostu方法、大津法、灰度直方图法，全局阈值算法简单，对于目标和背景明显分离、直方图分布呈双峰的图像效果良好，但对于由于光照不均匀、噪声干扰较大等原因使直方图分布不呈双峰的图像，二值化效果明显变差。局部阈值算法通过定义考察点的邻域，并由邻域计算模板，实现考察点灰度与邻域点的比较，非均匀光照条件等情况虽然影响整体图像的灰度分布却不影响局部的图像性质，从而使局部阈值算法比全局阈值算法有更广泛的应用，BemSen算法是一种典型的局部阈值算法，但局部阈值算法也存在以下

问题和缺点：如实现速度比全局阈值算法慢；不能保证字符笔划的连通性；容易出现伪影(ghost)现象(背景区域受到噪声干扰出现笔划)等。是一种自适应得二值化算法，它利用了像素自身及其邻域灰度变化特征，充分考虑了每个像素邻域的特征，能够更好的突出背景和目标的边界，使得相距很近的两条线不会产生粘连现象，动态阈值法时间开销大，另外也会在图像的某些部位上产生整体阈值法所没有的失真[13]。

4.2.1 大津法

大津法是日本的大津展之在1980 年提出的。该方法的基本思想是：设阈值将图像分割成两组，一组灰度对应目标，另一组灰度对应背景，使这两组灰度值的类内方差最小，两组的类间方差最大。方法具体实现过程如下：

设一幅图像的灰度值为1 ~ m级，在k 处将其分成两组 $C_0=\{1\sim k\}$，和 $C_1=\{k+1\sim m\}$，分别计算各组产生的概率 ω_0 和 ω_1，各组的组内平均值 μ_0 和 μ_1 及整体图像的灰度平均值 μ，两组间的方差用式（4-1）求出

$$\begin{aligned}\sigma^2(k)&=\omega_0(\mu_0-\mu)^2+\omega_1(\mu_1-\mu)^2\\&=\omega_0\omega_1(\mu_1-\mu_0)^2=[\mu\omega(k)-\mu(k)]^2/\omega(k)[1-\omega(k)]\end{aligned}\qquad(4\text{-}1)$$

从1～m之间改变k ，求上式为最大值时的k ，即求 $\max\sigma^2(k)$ 时的 k^* 值，此时的 k^* 值便是所求的阈值。 $\sigma^2(k)$ 叫做阈值选择函数。 大津法不管图像的直方图有无明显的双峰，都能得到较满意的结果[14]。

4.2.2 灰度直方图法

灰度直方图法针对具有双峰直方图的图像而言，具有较好的效果。由于图像由前景和背景组成，在灰度直方图上，前后二景都形成高峰，在双峰之间的最低谷处就是图像的阈值所在。

由于噪声的影响，直方图上的峰和谷都不会是完美的，“山峰”常常是由若干个窄的峰组成。但是，其中的最高点所在的灰度一般可以代表物体内部或背景点的典型灰度值。把阈值设在相对于两峰的某个固定位置，如中间位置上，一般来说，这个结果将比直接寻找最少出现的灰度值即估计直方图的谷的位置更为可

靠[15]。

文献[16]针对物体和背景在图像中所占的面积相差悬殊的情况，提出一种改进型的灰度直方图二值化方法。这种改进方法的运算方法如下：

(1)计算像素灰度平均值(avg)、标准偏差(sma)。

(2)以像素平均值为分界点，分别求出左、右部分的最大值的位置。

(3)若两峰值位置相距较近(在标准偏差范围内)，说明该直方图的双峰中有一个峰很低，因此需另寻低峰的位置，否则至第(7)步)。

(4)求出像素灰度中值点位置(midpos)。

(5)如果midpos>avg，表明小峰在大峰左边(较低灰度级)；否则，表明小峰在大峰右边(较高灰度级)。相应调整分界点位置。

(6)重新求出大、小峰值的位置。

(7)以两峰值位置的中点作为所要求取的阈值。

4.2.3 BemSen 算法

考虑以(x,y)为中心的$(2w+1)\times(2w+1)$模板，$f(x,y)$表示(x,y)处的灰度值，则Bemsen算法可描述为：

(1)计算每点阈值

$$T(x,y)=0.5\times(\max_{-w\le k,l\le w} f(x+k,y+l)+\min_{-w\le k,l\le w} f(x+k,y+l)) \tag{4-2}$$

(2)逐点二值化

$$b(x,y)=\begin{cases}0 & f(x,y)<T(x,y)\\ 1 & f(x,y)\ge T(x,y)\end{cases} \tag{4-3}$$

式中$b(x,y)$为进行二值化后置给点(x,y)的灰度值[17]。

文献[18]针对BemSen算法进行了改进，提出在运用BemSen算法进行全局二值化后，计算文字笔划的宽度，对属于文字笔画范围内的像素点进行局部二值化处理，其余点则进行全局二值化。

4.3 本文采用方法

上述常用方法虽然在各自的应用领域取得了较好的二值化效果，但在进行文

字图像二值化的过程中没有充分考虑文字图像特点，二值化效果较差；改进的BemSen算法虽然针对文字图像的二值化处理进行调整，但在二值化处理过程中要对一幅图像进行两次处理，运算时间较长，而且需要二值化模块在处理时同时保存原始的灰度图像和第一次全局二值化后的图像，对计算机的内存容量要求较高。

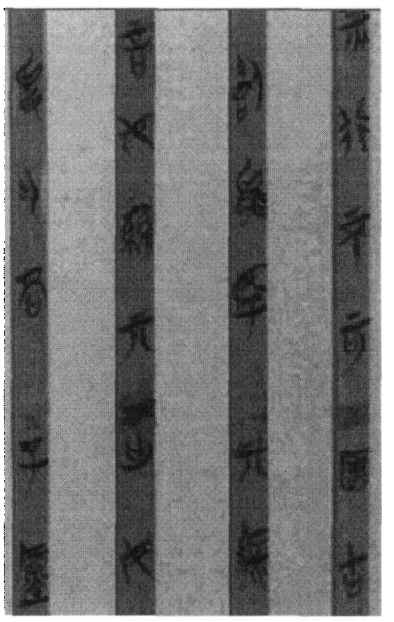


图 4-1 中值滤波后的简牍图像

仔细研究中值滤波后的简牍图像会发现，在文字笔画的左右两侧有明显的灰度变化，而且由左至右灰度值是先减小再增加，由此可将观察到的现象归结如下：

（1）文字笔画必处于灰度递减与递增的区域之间。

（2）文字笔画所在处的灰度必为先递减再递增。[19]

因此，在文献[19]中，作者提出使用灰度的差分值来求二值化处理的阈值。

在简牍图像中，能否使用灰度的差分值来求二值化处理的阈值呢？在简牍图像中，不仅在竹简（或木牍）与文字之间存在灰度差别，在竹简（或木牍）和背景之间也存在灰度差别，如果单纯利用具有最大灰度差分处的灰度值作为二值化处理的阈值，很有可能会误将竹简（或木牍）处的灰度值最为二值化处理的阈值，这样二值化后的图像必然是简牍处全为黑色，无法提取出文字，因此有必要对原有的灰度差分方法进行改进。

虽然在竹简（或木牍）和背景之间存在灰度差别，但其灰度差和竹简（或木牍）与文字之间的灰度差存在区别：由于竹简（或木牍）为纵向排列，因此竹简

（或木牍）与背景之间的灰度差别基本为由左至右灰度值逐渐减小再增加，但在纵向上灰度差别极小；文字与竹简（或木牍）之间，由于文字位于竹简（或木牍）之上，笔画与竹简之间的灰度差别不仅体现在由左至右方向上，同时在笔画边界的像素点的上、下、左、右、左上、左下、右上、右下8邻域上都会有所反映，因此，从像素点的8个邻域方向上计算灰度差，具有最大的灰度差值的像素灰度就应该是文字的边沿灰度，进而找出二值化处理的阈值，因此可以将这种方法称之为8邻域灰度差值法。

在设计识别系统的二值化处理模块时，计算每个像素点与周围8邻域的灰度差之和与该点灰度的乘积：

$$S_{(i,j)}=(\sum_{x=i-1}^{x=i+1}\sum_{y=j-1}^{y=j+1}\left|k_{(i,j)}-k_{(x,y)}\right|)\times k_{(i,j)} \tag{4-4}$$

式中，$k_{(i,j)}$ 为点 (i,j) 处的灰度值，设

$$S_p=\sum_{k_{(i,j)}=p}S_{(i,j)} \tag{4-5}$$

其中 $0\le p\le 255$，设 $S_{p(\max)}$ 处的灰阶值具有最大的灰度差异。

设阈值为 $K=\alpha\times S_{p(\max)}$，当点（i,j）处的灰阶值大于或等于K时，该点为背景点，即 $k_{(i,j)}=255$，否则该点为黑点，即 $k_{(i,j)}=0$。

4.4 实验效果分析

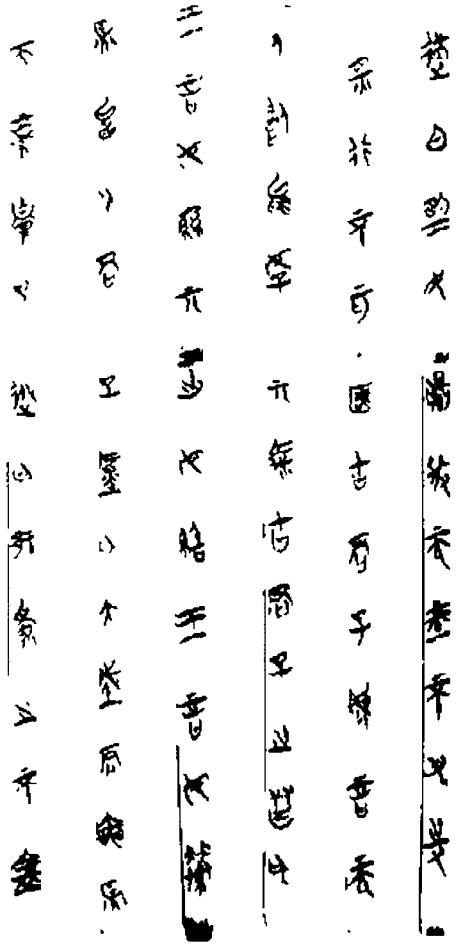


图 4-2 $\alpha = 0.29$

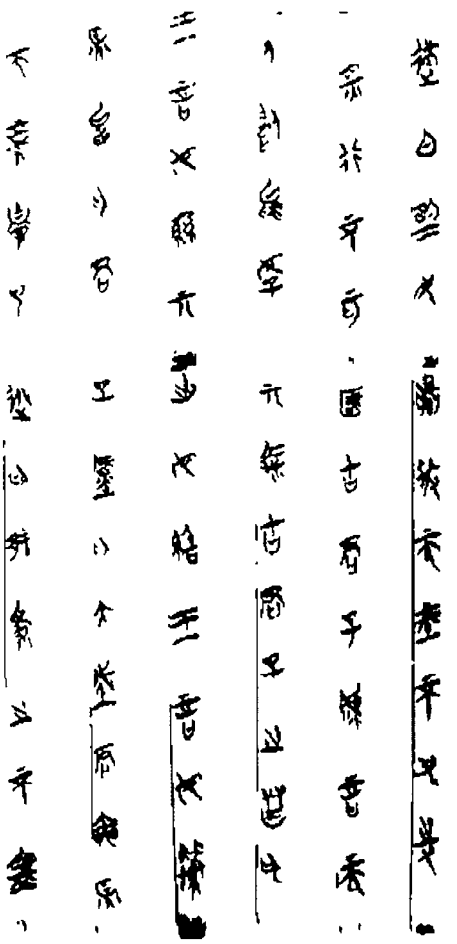


图 4-3 $\alpha = 0.35$

图 4-4 $\alpha = 0.45$

图 4-5 $\alpha = 0.55$

从实验结果看，二值化处理效果的好坏在很大程度上取决于α的取值，当α取值较大时，二值化阈值过大，导致简牍背景也被误认为是文字，置成了黑色，使后续的文字切分模块无法准确切分出文字；当α取值过小时，二值化阈值太小，会使文字笔划出现断笔、甚至出现笔划缺失的现象，导致过高的文字误识率；当$\alpha = 0.35$时，文字即没有出现断笔的情况，同时大部分的竹简（木牍）被滤除掉，效果最好。

从图4-3可知，二值化后的图像无法完全去除简牍背景，这是由于部分简牍背景的灰度值小于文字笔划的灰度，如果为了去掉简牍背景而降低二值化阈值，会导致文字笔划出现断笔或缺失，所以在二值化处理模块中无法完全去掉这些背景噪声（由于简牍背景对于文字识别来讲是无用信息，同时背景的存在还会导致文字切分和特征提取的困难，所以可以将其看成噪声）。

虽然在二值化处理模块无法完全去掉简牍背景噪声，但由二值化后图像可以看出，文字已经基本被提取出来，实现了二值化处理的目标。

在进行二值化处理时，之所以没有采用具有最大灰度差的灰度数值作为二值化处理阈值，是由于通常情况下，文字的边沿位置灰度数值会大于文字内部灰度

的数值，而如果直接利用文字笔划边沿灰度值作为二值化的阈值，由于其数值较大，会引入较多的背景噪声，因此在实际处理过程中，以边沿灰度作为基础将其数值适当降低作为二值化处理的阈值。

8邻域灰度差值法利用了文字图像的特点（在文字笔划的边沿位置，其8邻域方向上存在灰度差），寻找文字笔划的边沿像素点灰度值，进而找出二值化处理的阈值灰度。由于该方法充分利用了文字图像的特点，因此具有较好的二值化效果，但同时也存在着明显的缺点—α的取值。从前面的实验效果可知，二值化效果在很大方面取决于α的取值，对于不同的文字图像，当其在光照效果不同的情况下，竹简颜色深浅程度不同的情况下，需要选取不同的α值才能取得最好的二值化效果，而目前的8邻域灰度差值法α值还无法根据图像的实际情况进行调整，在对不同的输入图像情况下需要手工进行修改，使用起来不是很方便。

因此， 8邻域灰度差值法的下一步修改目标为将8邻域灰度差值法与动态阈值法的优点结合起来，尝试通过估算简牍背景灰度与文字边沿灰度之间的对比关系，动态调整α的大小，使之能够自动适应不同光照及不同背景情况下的文字图像二值化处理。

第五章 文字切分

5.1 作用

文字切分是整个识别系统中极为重要的一个环节，因为正确的识别往往依赖于正确的切分，如果不能正确的从文字图像中切分出单个文字，即使有很好的识别算法，也无法获得高性能识别系统。由于在文字切分之后的单个汉字会被归一化为统一的大小，并使文字字符充满整个空间，故行字切分只需要找出每个文字字符的外接矩形，但这个外接矩形不必是此字符的最小的外接矩形，只要在矩形中完整地包括此字符的笔画，同时没有其它字符的笔画就可以了[20]。

5.2 常用方法

文字切分主要有三种切分策略和这三种策略混合使用产生的多种方法。这三种切分策略为：标准切分法，基于识别的切分法，整体切分法[21]。标准切分法主要根据图像本身所具有的属性进行处理，对于图像质量好、字与字之间有固定间距的文档处理效果好，如垂直投影法、连通域法。基于识别的方法提供多个切分假设，然后对切分结构进行选择得到最优的切分结果，这种方法效果好，但是相对比较复杂、耗时，在实际中应用得较少，如HMM法。整体切分法则是把一个单词作为一个整体来进行识别，这种方法虽然避免了单词内部字符切分的问题，但它依赖于现有的定义好的词典，这大大限制了该方法的应用范围，在汉字识别中则没有使用[22]。

5.2.1 垂直投影法

垂直投影法在文字切分过程中被广泛使用。垂直投影$V(x)$是通过把在x坐标处所有的黑点相加获得。如果在某一列上存在有文字，则其$V(x)$的值必不为0，而在理想情况，当文字之间不存在粘连的情况时，文字之间的背景区域下$V(x)$的

值等于0.通过求$V(x)$的值即可求出文字的大致位置。

当文字图像存在有多行文字时，可以先对文字做水平投影，求出每行文字所在的位置，再通过垂直投影求出没个文字的具体位置。

垂直投影法适用于文字之间存在有明显间距的情况，当文字间存在粘连时切分效果则无法产生很好的效果。

5.2.2 连通域法

连通域单元就是每个连通域的最小矩形区域 。连通域算法的基本思想是基于连通域的组合(这里指的仅仅是像素点阵的连通域-也就是说如果一个汉字由几个分离的部件构成，那么它将对应多个连通域，其中由每一个部件求得一个连通域)[23]。就汉字的结构来说，组成一个汉字的各个连通域单元之间的位置关系可分为上下、左右和包含三类 。根据两个连通域单元之间的距离和重叠量等位置参数，结合汉字本身的结构特点可合理的合并连通域单元[24]。

连通域算法的基本思想是基于连通域的组合，所以算法的第一步是求字符点阵连通域。通常可以采用区域扩展法来求解连通域。定义连通域单元i的左上角坐标为(L_i,T_i)，右下角坐标为(R_i,B_i)，高度为H_i，宽度为W_i，单元i和单元j的水平间隙为$G_{(i,j)}$，单元i和单元j合并后的宽度为$W_{(i,j)}$，则有：

$$G_{(i,j)} = \max(L_i,L_j) - \min(R_i,R_j) \tag{5-1}$$

$$W_{(i,j)} = \max(R_i,R_j) - \min(L_i,L_j) \tag{5-2}$$

合并步骤如下：

（1）合并相邻有重叠的连通域单元。对于第i个连通域单元，如果满足$G_{(i,j)} < 0$，则合并单元i和单j。

（2）估计字符高度H。鉴于同一人书写的简牍文字中，字与字之间的高度差异不会太大，用中值滤波的方法可得到字符高度的估计值H。方法如下：

对于连通域单元1,2,…… ,n，把各连通域单元的高度H_1，H_2，……，H_n按大小顺序排列为$H_{k1} \le H_{k2} \le \cdots \le H_{kn}$，则有

$$H = Med(H_{k1}, H_{k2}, \cdots, H_{kn}) \tag{5-3}$$

（3）根据字符高度H合并过分的字符。对于满足如下条件的单元i：

$$W_i < \alpha H \times \min(W_{i-1,i}, W_{i,i+1}) < \beta H \tag{5-4}$$

当$W_{i-1,i} \le W_{i,i+1}$时，合并单元$i-1$与单元i，当$W_{i-1,i} > W_{i,i+1}$时，合并单元i和单元$i+1$。α和β为常数。

（4）根据字符高度H确定含有粘连字符的连通域单元。对于单元i，如果满足$W_i > \lambda H$，则认为该连通域单元中含有需要切分的粘连字符，其中λ为常数。

5.3 本文采用方法

5.3.1 简牍图像特点

通过对大量的简牍图像进行分析，归结出简牍图像具有如下特点：

（1）一幅简牍图片中有多枚简牍，且简牍数目不定。

（2）简牍纵向排列，且简牍之间存在明显的间隔区域。

（3）多数简牍上只有一列文字，且文字之间间隔明显。

（4）由于简牍实物长期埋藏于地下，出土时简牍上出现明显的腐蚀现象，腐蚀部分在二值化后图像中成为大噪声点。此外，竹节为大噪声点的另一主要来源。

（5）简牍文字大小不一，不能采用统一的标准进行切分。

（6）由于光照及简牍变色等原因，简牍边沿的颜色往往深于简牍内部颜色，在二值化过程中不能全部去除简牍边沿，导致在二数值化后的简牍图像中出现呈现带状分布的大片噪声区域。经分析，带状噪声区域位于文字左右两边，且多跨越在多个文字之间。

5.3.2 解决方案

针对简牍图像特点，设计文字切分算法流程如下：

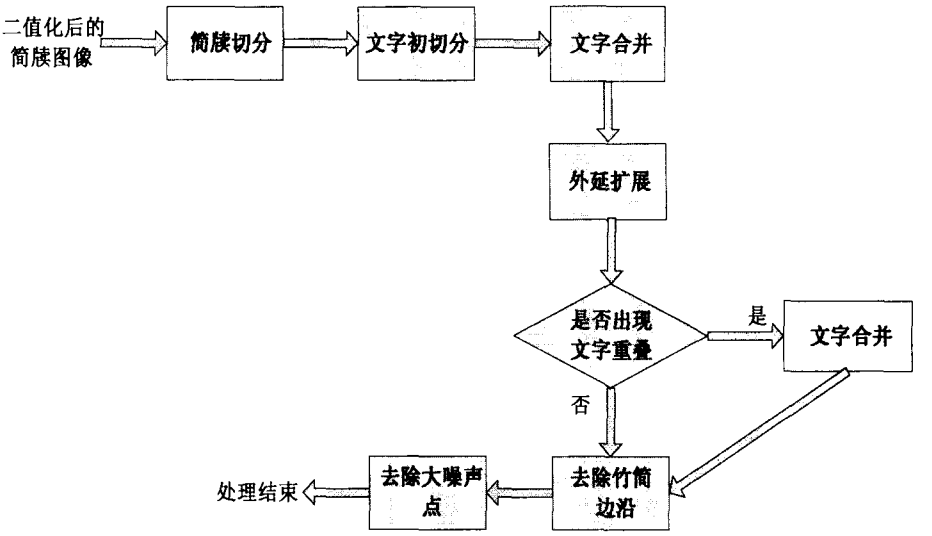


图5-1　　文字切分流程

5.3.2.1 单个简牍切分

图5-2 二值化简牍图像

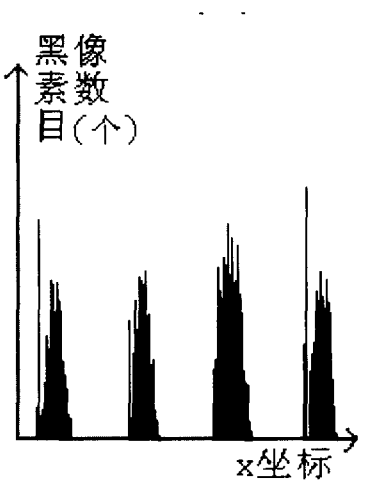


图5-3　垂直投影图

图-2是一幅二值化后的简牍图像，图5-3是对图5-2做的垂直投影，其中 x 轴为对应简牍图像的行坐标，纵轴为对应该坐标对应的一列具有的黑像素的个数。由图可知，简牍之间的区域黑像素数目趋于0，简牍所在处与空白区域之间差别明显且没有过渡地带。利用这个特点，设计简牍切分算法如下：

（1）设 C_x 为图像列性质标志变量，当 $C_x=1$ 时，表示该列存在文字，当 $C_x=0$ 时，表示该列为图像背景，且设 $C_{-1}=C_{NW}=0$（NW 为该图像宽度）。

（2）从$X=0$处开始计算每一列的垂直投影N_x，若$N_x \geq 4$且$C_{x-1}=0$，则该列存在文字且x为单枚简牍的起始x坐标x_b，若$N_x \geq 4$且$C_{x+1}=0$，则x为简牍的终止x坐标x_e。

（3）计算简牍的平均宽度，若某一枚简牍的宽度小于平均宽度的$1/2$，将其于距离该枚简牍最近的简牍合并。

5.3.2.2 文字切分与文字合并

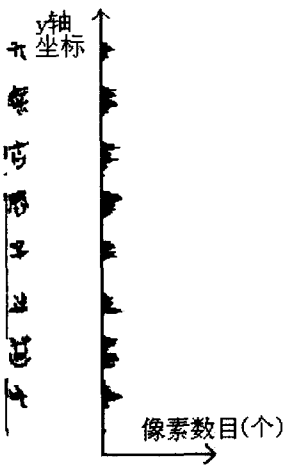


图 5-3 水平投影图

图5-3是单枚竹简的水平方向投影图，图中纵坐标是对应简牍图像的y坐标，横轴是对应该坐标的一行具有的黑像素的个数。从图中分析可知，文字所在处黑像素的数目远大于文字间的黑像素的数目，文字间的区域黑像素数目不一定为0，由此设计文字初切分算法如下：

（1）设L_y为图像行性质标志变量，当$L_y=1$时，表示该行存在文字，当$L_y=0$时，表示该行为图像背景，设$L_{-1}=L_{NH}=0$（NH为该图像高度）。

（2）从$y=0$处开始统计每一行图像的黑像素数目N_y，若$N_y \geq N_{th}$（黑像素数目阈值）且$L_{y-1}=0$，则该行存在文字且y为文字的起始y坐标y_b，若$N_y \geq N_{th}$且$L_{y+1}=0$，则y为文字的终止y坐标y_e。N_{th}的选择是该算法的关键，若N_{th}太小，会导致简牍边沿被误认为是文字而产生误切分，若N_{th}太大，又会

产生切割过度的现象，使文字丢失部分笔画。经实验证明，当$N_{th}=2$时软件可取得较好的切分效果。

（3）计算文字的平均高度，若某一文字的高度小于平均高度的$1/2$，将其与距离该文字最近的文字合并。

（4）在文字合并时，取y坐标较小的文字的y_b作为合并后文字的y坐标起始值，取y坐标较大的文字的y_e作为合并后文字的y坐标的终止值。

（5）统计该枚简牍的文字总数，记为N_{word}。

（6）重复步骤（1）～（5），直到每枚简牍的所有文字都切分完毕。

5.3.2.3 外延扩展

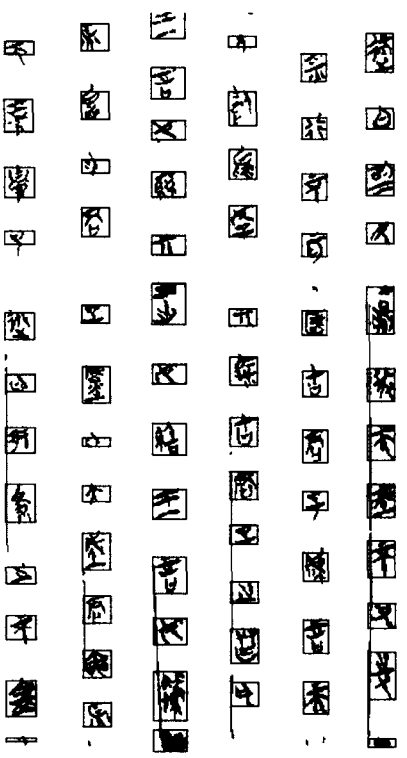


图 5-4 文字初切分合并效果

图5-4是文字初切分及文字合并后的效果图，由图可以看出，部分位于文字上下两端的笔画被误切割掉。经分析，这是由于笔画宽度小于N_{th}个像素导致被误认为是文字间背景噪声而被切割掉的，此时需要对文字进行外延扩展，找回被切割掉的笔画。设计文字外延扩展算法如下：

（1）设K_{Hn}(K_{Ln})为文字n上（下）边沿扩展标志变量，并设其初始值为0。设y_b为文字n的起始y坐标，y_e为终止y坐标。设$y=y_b-1$。

（2）统计第y行黑像素数目N。为避免简牍边沿被误认为是缺失笔画，在进行像素统计时，以坐标x_b+4为统计的起始坐标，以坐标x_e-3为终止坐标。

（3）当N大于或等于1时，判定有笔画被误切除,$y=y-1$，同时K_{Hn}加1。并重复（2）。

（4）当N小于1时，判定已无笔画缺失，令$y_b=y_b-K_{Hn}$。

（5）令$y=y_e+1$。

（6）统计第y行黑像素数目N。为避免简牍边沿被误认为是缺失笔画，在进行像素统计时，以坐标x_b+4为统计的起始坐标，以坐标x_e-3为终止坐标。

（7）当N大于或等于1时，判定有笔画被误切除,$y=y+1$，同时K_{Ln}加1，并重复（6）。

（8）当N小于1时，判定无无笔画缺失，。令$y_e=y_e-K_{Ln}$。

（9）重复步骤（1）～（8），直到每枚简牍的所有文字都确认完毕。

5.3.2.4 文字重叠判断及文字合并

在软件进行文字外延扩展之后，部分文字之间可能出现重叠。文字重叠现象的出现意味着原有的一个文字被错误切分为两个字，此时需要将错误切分的两个字还原为一个文字。因此在文字外延扩展模块之后需要对外延处理后的图像进行文字重叠判断，若出现重叠现象，进行文字合并处理。

此部分算法如下：

（1）设文字n是简牍中的第n个文字（$1\le n\le(N_{word}-1)$），比较终止y坐标y_{en}与下一个文字的初始y坐标的$y_{b(n+1)}$之间的关系，若$y_{en}>y_{b(n+1)}$，则该文字n与下一文字$n+1$没有重叠，不需进行文字合并;若$y_{en}\le y_{b(n+1)}$，该文字n与下一文字$n+1$发生重叠，需进行文字合并步骤。

（2）若文字n与文字$n+1$发生重叠，将文字n和文字$n+1$合并为一个字，

新字的起始y坐标为文字n的起始y坐标y_{bn}，新字的终止y坐标为文字$n+1$的终止y坐标$y_{e(n+1)}$。

（3）简牍总字数减1，即$N_{word}=N_{word}-1$。

（4）重复步骤（1）～（3），直到每枚简牍的所有文字都确认完毕。

5.3.2.5 去除竹简边沿

针对简牍边沿的特点（位于多个文字之间，位于文字的左右两边），可以设想如果我们搜索文字之间的空白区域，就可获知是否存在简牍边沿以及简牍边沿的位置，利用这些信息就可以去除文字内部的边沿噪声了。设计软件实现算法如下：

（1）取一枚简牍，设文字n是该枚简牍的第n个字（$1\le n\le N_{word}$），设变量k为计数变量，初始值为0。设$P_{(x,y)}$为点(x,y)处图像的灰度值，当$P_{(x,y)}=0$时，该像素点为黑点，当$P_{(x,y)}=255$时，该像素点为白点。设坐标$x=x_{bn}$。

（2）在坐标x处纵向统计文字n与文字$n+1$之间空隙的黑像素的数目N_y，即$N_y=\sum_{y=y_{en}+1}^{y_{b(n+1)}-1}\frac{(255-P_{(x,y)})}{255}$。令$x=x+1$。

（3）若$N_y\ge B_1\times(y_{b(n+1)}-y_{en})$（其中$B_1$为比例系数），简牍边沿同时穿过文字$n$与文字$n+1$,需对文字$n$与文字$n+1$进行去除边沿处理：将文字$n$与文字$n+1$的坐标$x$处的所有像素灰度值设置为255，即设置为白像素。

（4）若$N_y\ge B_2\times(y_{b(n+1)}-y_{en})$（其中$B_2$为比例系数），简牍边沿只穿过一个文字。若$P_{(x,y_{en}+1)}=0$，简牍边沿穿过文字$n$，将文字$n$的坐标$x$处的所有像素灰度值设置为255；若$P_{(x,y_{b(n+1)}-1)}=0$,简牍边沿穿过文字$n+1$,将文字$n+1$的坐标$x$处的所有像素灰度值设置为255。

（5）若文字为单枚简牍的最后一字，设置$y_{b(n+1)}=HEIGHT-1$(其中$HEIGHT$为图像高度)。

（6）若文字为单枚简牍的首字，只需将$y_{en}=0$，$y_{b(n+1)}$设定为首字的起始y

坐标即可。

（7）重复执行（2）～（6）至$x = x_{en}$结束。

（8）重复执行（1）～（7），直至所有简牍的所有文字都进行完判断为止。

（9）经实验证明，当$B_1 = 0.9$，$B_2 = 0.4$时，软件去除边沿的效果最好。

5.3.2.6 去除大噪声点

通过分析，大噪声点所在处黑像素占像素总数的比例高于文字图像中黑像素占像素总数的比例，利用这个特点，设计软件算法如下：

（1）调整文字坐标。由于在进行去除文字边沿模块之后，部分文字的起始x坐标和终止x坐标处不存在黑像素，在此基础上计算黑像素密度不能如实反映其真实的像素比例，同时对文字坐标的调整还可以更加精确的给出文字的实际起始（x_b，y_b)和终止坐标（x_e，y_e）。

（2）计算每个文字区域内黑像素的数目，即计算$N_b = \sum_{i\in(x_b,x_e),j\in(y_b,y_e)} P_{(i,j)} = 0$，计算$q = \frac{N_b}{(x_e - x_b + 1)\times(y_e - y_b + 1)}$（$q$为文字区域内黑像素密度）。

（3）若$q \geq Q_{th}$（Q_{th}为黑像素密度阈值），黑像素占总像素比例过高，判定该文字为大噪声点，从文字列表中清除；否则，保留该文字信息。经实验，当$Q_{th} = 0.65$时，软件即可以去除大部分的大噪声点，同时又能避免文字被误判断删除。

5.4 实验结果分析

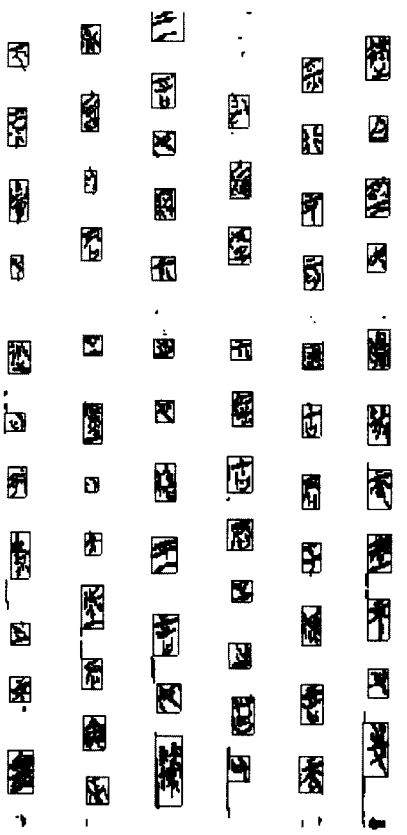


图 5-5 文字切分效果图

从文字切分的实验效果图来看，本文所采用的文字切分算法有效的将文字从图像中提取了出来，基本去除了大噪声点，并且部分除掉了干扰文字识别的简牍边沿，基本达到了文字提取的目的。

但部分简牍的边沿仍保留在切分后的文字内部，原因如下：

（1）由于简牍的边沿噪声并非均匀分布在简牍上面，同一边沿噪声的宽度会出现变化，导致位于文字中间的噪声宽度与文字内部的噪声宽度不相等，当文字中间的边沿噪声宽度小于文字内部的边沿噪声宽度时，就会出现边沿噪声没有滤除干净的现象。

（2）由于部分简牍的边沿已经与文字相连，如果在设计过程中，将与边沿噪声相连的点全部去除掉，会误删除文字的笔画。

而且，本文所设计的文字切分算法，只适用于单个简牍上面只有一列文字的情况，当单个简牍上面有多列文字时，文字切分的效果并不理想。

因此，在以后的研究中，必须针对算法的这两个缺陷进行改进：

（1）对于无法完全去掉简牍边沿的情况，由于在二值化图像中是很难准确

区分边沿和文字笔画的，可以考虑在彩色简牍图像中利用简牍边沿与文字笔画之间的颜色差异进行区分。

（2）对于单个简牍上面具有多列文字的情况，可以考虑在设计文字识别系统时对文字切分模块加入一个选择分支，即在输入简牍图像时直接告诉系统输入的图像为单列文字情况或多列文字情况：当输入为单列情况时，利用本文所述方法进行文字切分；若输入的简牍上面有多列文字时，可先用本文提供的方法切分出文字词组，再结合其它文字切分算法进行精确切分。

第六章 归一化处理

6.1 作用

归一化处理是预处理中非常重要的一环，无论手写体字符还是印刷体字符，由于原始图像在大个、形状等方面存在着很大的差异，必须进行归一化处理，使其具有相同的大小和形状，以利于特征提取和识别的进行[25]。统计识别中，基于各种特征的相关匹配要求特征向量和模板向量具有相同的推数，各个分量一一对应，否则距离或相似度的计算便难以进行；结构识别中，虽然更注重字符的结构和笔划之间相互的关系，但笔划长度也常常作为一个很重要的属性被加以利用，若字符大小不做归一化处理，这一属性就失去意义[26]。

6.2 常用方法

归一化包括位置归一化和大小归一化，大小归一化又有两种方法，即线性归一化和非线性归一化。

6.2.1 位置归一化

汉字点阵的位置归一化方法主要有两种，一是重心归一化，二是外框归一化。重心归一化方法是计算出汉字的重心后将重心移到汉字点阵的规定位置，如中心位置上，即重心归一化后汉字的重心位于点阵中心。外框归一化是将汉字的外框移到点阵规定位置上。因为重心计算是全局性的，因此抗干扰能力强；各边框搜索是局部性的，易受干扰影响。而大多数字笔划分布左、右、上、下比较均匀，汉字的重心和汉字字形的中心相差不多，重心归一化不会造成字形失真，但对个别汉字如乎、于、丁等字，上下分布不匀，重心归一化使字形向下移动，以致字形下端超出点阵范围而造成失真[27]。

6.2.2 大小归一化

常用的大小归一化方法有两种，一种是以文字的外接边框大小为依据，将文字按比例线性放大或缩小，成为规定尺寸的文字，即线性归一化；另一种是根据文字在不同方向上黑像素的分布进行归一化，即非线性归一化。

（1）线性归一化

线性归一化实现的方法是对像素坐标进行到线性变换，将不同尺寸的图像转换成固定大小的图像。

设某一点(x,y)经过归一化后其位置变为(X,Y)，则两者之间存在有如下关系：

$$X = ax \tag{6-1}$$

$$Y = by \tag{6-2}$$

a,b分别是x方向、y方向的放大率。a，b比1大时放大，比1小时缩小。对于所有的像素点(x,y)进行计算，把输入图像上的点(x,y)的灰度值代入输出图像上的点(X,Y)处，就可以把图像放大或缩小了[9]。

（2）非线性归一化

将图像中的黑像素分别向水平轴和垂直轴两个方向进行投影，得到两个方向的点密度如式(6-3)，(6-4)所示：

$$H(i) = \sum_{j=1}^{J} f(i,j) + \alpha_H \tag{6-3}$$

$$V(j) = \sum_{i=1}^{I} f(i,j) + \alpha_V \tag{6-4}$$

其中α_H和α_V为常数，其作用是为了控制转换的强度，其值越大，则变换强度越小，可以根据实际情况进行选取。变换后的新坐标位置(X,Y)可以由式（6-5），（6-6）得到[25]：

$$X = \sum_{k=1}^{i} H(k) \times \frac{M}{\sum_{k=1}^{I} H(k)} \tag{6-5}$$

$$Y=\sum_{l=1}^{j}V(l)\times\frac{N}{\sum_{l=1}^{j}V(l)} \tag{6-6}$$

非线性归一化方法要优于线性归一化方法，但运算速度较慢。

6.3 本文采用方法

本文在实践过程中采用线性归一化方法。

对于输出图像上的所有像素(X,Y)，运用公式（6-1）、（6-2）的逆公式（6-7）、（6-8）求解其对应在输入图像上的像素，写入这个像素的灰度值。

$$x=X/a \tag{6-7}$$

$$y=Y/b \tag{6-8}$$

由于公式（6-7）和公式（6-8）进行的是实数运算，而a和b可能带有小数，所以计算出的x和y可能包含小数位。然而，输入图像的像素地址必须是整数，所以对于地址计算，有必要采取某种形式对地址进行整数化。

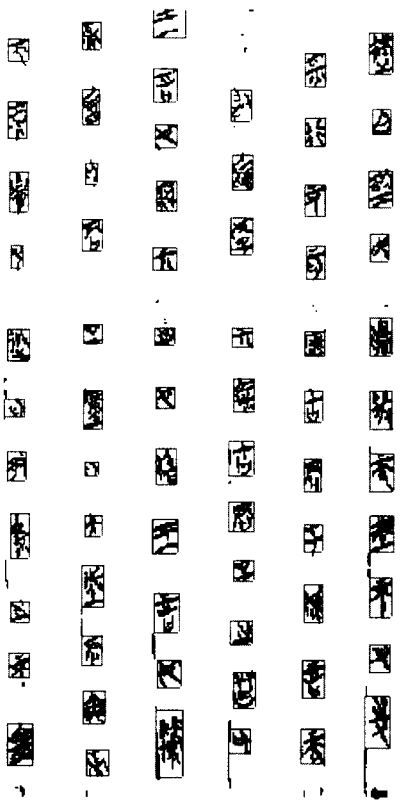
在设计系统时，本文采用了双线性内插的方法。这种方法是当所求的地址不是整数时，求得(x,y)到相邻4个像素点的距离之比，用这个比率和4个相邻像素的灰度值进行灰度内插。

其计算公式如下：

$$\begin{aligned}d(x,y)&=(1-q)\times(1-p)\times d([x],[y])+(1-q)\times p\times d([x]+1,[y])\\&+q\times(1-p)\times d([x],[y]+1)+p\times q\times d([x]+1,[y]+1)\end{aligned} \tag{6-9}$$

其中，$d(x,y)$为点(x,y)处的灰度值，$[x]$和$[y]$分别是不超过x和y的整数值，$p=x-[x]$，$q=y-[y]$。

6.4 实验结果分析



6-1 二值化后简牍图像

图 6-2 归一化效果图

在实验过程中，本文将简牍文字大小统一为 40×40。对比二值化后简牍图像与归一化后的简牍图像，可以发现归一化后的简牍文字较好的保存了原文字的结构信息。由于在设计文字识别系统时，设计图像显示页面为固定大小，在将文字大小做了归一化后，部分文字笔划超出了显示范围，所以在归一化的效果图上会发现有部分的文字笔画出现了缺失，主要体现在简牍图像的顶部，部分文字的笔画丢失。

第七章 细化

7.1 作用

对于字符识别来说，原始图像中的信息存在着一定程度的冗余，这种冗余表现在笔划的宽度上。通常扫描进来的原始图像，其笔划宽度一般都在5个像素左右，信息的冗余是显而易见的。因为要描述字符的结构，单像素宽度的笔划就足够了。

细化又称为骨架化，即在不影响原图像拓扑连接关系的条件下，尽可能用最少的迭代次数，快速而准确的将宽度大于一个像素的图形线条转变为一个像素宽的线条的处理过程，也就是抽取图像的骨架。

细化的必要性不仅仅只是压缩冗余信息的需要，同时还是对模式进行结构分析的需要。对于字符来说，细化后的图像更符合人类的识别习惯，便于我们进行结构分析并设计出更为直观的识别算法。同时，细化后的图像有利于某些重要特征的提取，如端点、交叉点、拐点等特征点及各个部件间的连接关系，一般都是基于骨架进行提取的。另外，输入所产生的噪声往往集中在字符笔划的边缘上，通过细化便可以消除这些噪声点，并保持原始图像重要的几何拓扑特。

细化的目的是在保持原有模式重要特征的前提下.压缩冗余信息，清除局部的噪声，以利于特征提取的进行。要达到上述目标，并做到细化时不产生新的畸变，同时又要使算法尽可能简单、快速，是一件非常困难的事情。正因为如此细化算法的研究一直是一个非常活跃的领域[28]。

7.2 常用方法

一个好的细化算法应该达到如下要求：

（1）骨架图像必须保持原图像的连通性。

（2）骨架图像应尽可能是原图像的中心线。

（3）细化结果要尽可能细，争取得到一个像素宽的线条图像。

（4）应使用尽可能少的迭代次数。

（5）保留曲线的端点。

（6）保持原有字符的拓扑，几何特征，不应产生严重的畸变。

细化过程要不断重复扫描图像，在图像处理中是耗时较多的操作。因此，研究如何用最少的迭代次数得到完全细化的图像，缩短细化时间是很有意义的。已有的细化算法按迭代方式不同分为串行算法和并行算法。在串行细化算法中，当前迭代的结果不仅取决于前一次的迭代图像，而且与当前处理情况有关；而在并行方式中，当前迭代仅仅由前一次的迭代情况决定。串行细化算法的处理结果依赖于对像素处理的先后顺序，因而像素点的消除或保留不可预测；而并行细化算法对图像进行细化时利用相同的条件同时检测所有像素点，其结果具有各向同性，因此从算法原理上并行方法优于串行方法[29]。

由于并行细化算法具有快速而准确的特性，因此它一直是人们研究的热点，并且相应的提出了许多并行细化算法，如OPTA细化算法、Rosenfeld细化算法、Zhang&Suen细化算法以及ZR细化算法等等。

7.2.1 HSCP 细化算法

P1	P2	P3
P8	P0	P4
P7	P6	P5

图 7-1　8 连通示意图

HSCP 算法可以简述如下：

步骤1：针对所有的边缘点 P_0，若满足如下条件，则判断其为可删除的：

（1）其8连通域中的边缘点数 $E(P_0)$ 为 $2 \le E(P_0) \le 6$。

（2）P_0 的8连通域中包含且只包含一个4连通边缘点。

步骤2：遍历所有的可删除点，若满足如下条件之一，则保留；否则删除：

（1）P_2、P_6 为边缘点，而 P_4 为可删除的。

（2）P_4、P_8 为边缘点，而 P_6 为可删除的。

（3）P_4、P_5、P_6均为可删除的。

7.2.2 Hilditch 细化算法

Hilditch 经典细化算法是一种比较有效的二值图像细化算法，其主导思想是每次删除图像上目标的轮廓像素，直到图像上不存在可删除的轮廓像素为止。二值图中目标像素值为1，背景像素为0。图像中任一点P_0的8 邻域如图7-2 所示。在此算法中，如果目标像素满足下列条件即为可删除像素[30]：

P1	P2	P3
P8	P0	P4
P7	P6	P5

图 7-2 8 连通示意图

（1）$P_2 + P_4 + P_6 + P_8 \le 3$。

（2）$N_{C'} = 1$。

（3）$P_k (1 \le k \le 8)$中至少有一个目标像素为1。

（4）$P_2 = 1$或$N_c 2 = 1$。$N_c 2$为假设$P_2 = 0$时，P_0的联结数。

（5）$P_8 = 1$或$N_c 8 = 1$。$N_c 8$为假设$P_8 = 0$时，P_0的联结数。

其中，联结数为8邻域中互相分离的联结成分的个数。

7.2.3 基于 Delaunay 三角网的骨架提取算法

Delaunay三角剖分是Voronoi的对偶图。Delaunay三角剖分具有两个非常重要的性质[31]：

（1）外接圆性质:顶点集V所形成的Delaunay三角剖分中，每个三角形的外接圆均不包含顶点集V中的其它任意点。

（2）最小角最大性质：在由顶点集V所能形成的三角剖分中，Delaunay三角剖分所形成的三角形的最小角度是最大的。

以上两个性质决定了Delaunay三角剖分具有极大的应用价值，同时它也是二维平面三角剖分中唯一的、最好的。

Delaunay三角剖分根据要剖分对象的不同分为离散点集的Delaunay三角剖分和域的Delaunay三角剖分。

（1）离散点集的Delaunay三角剖分把离散点集中的点连接成一些符合最小角最大的三角形。除了一些特殊的情况外，离散点集的Delaunay三角剖分是唯一的。Delaunay三角剖分被证明是好的三角剖分。在广泛应用于文字识别、地球物理、科学计算可视化等很多方面。

（2）域的Delaunay三角剖分把平面的多边形区域剖分成一些形状饱满、满足Delaunay标准的三角形。

基于Delaunay三角网的骨架提取算法，首先计算图像边界的逼近多边形，然后基于边界多边形的顶点进行Delaunay三角剖分，并对生成的三角形进行调整，从而得到该边界多边形的比较理想的三角剖分结果，接着根据每一个三角形的类型来计算局部骨架，最后这些局部骨架的集合就构成了整个带状图像的骨架。算法流程如下：

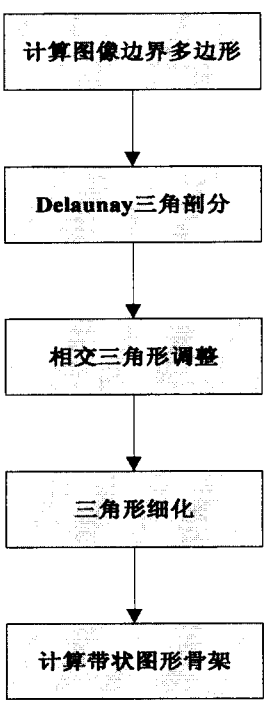


图 7-3　算法流程

7.3 本文采用方法

本文在设计细化算法时，选择形态学方法进行细化处理。即当像素点灰度为0（即当前像素为黑像素），且与周围8邻域的像素满足模板所列出情况时，将该点像素删除，否则保留。

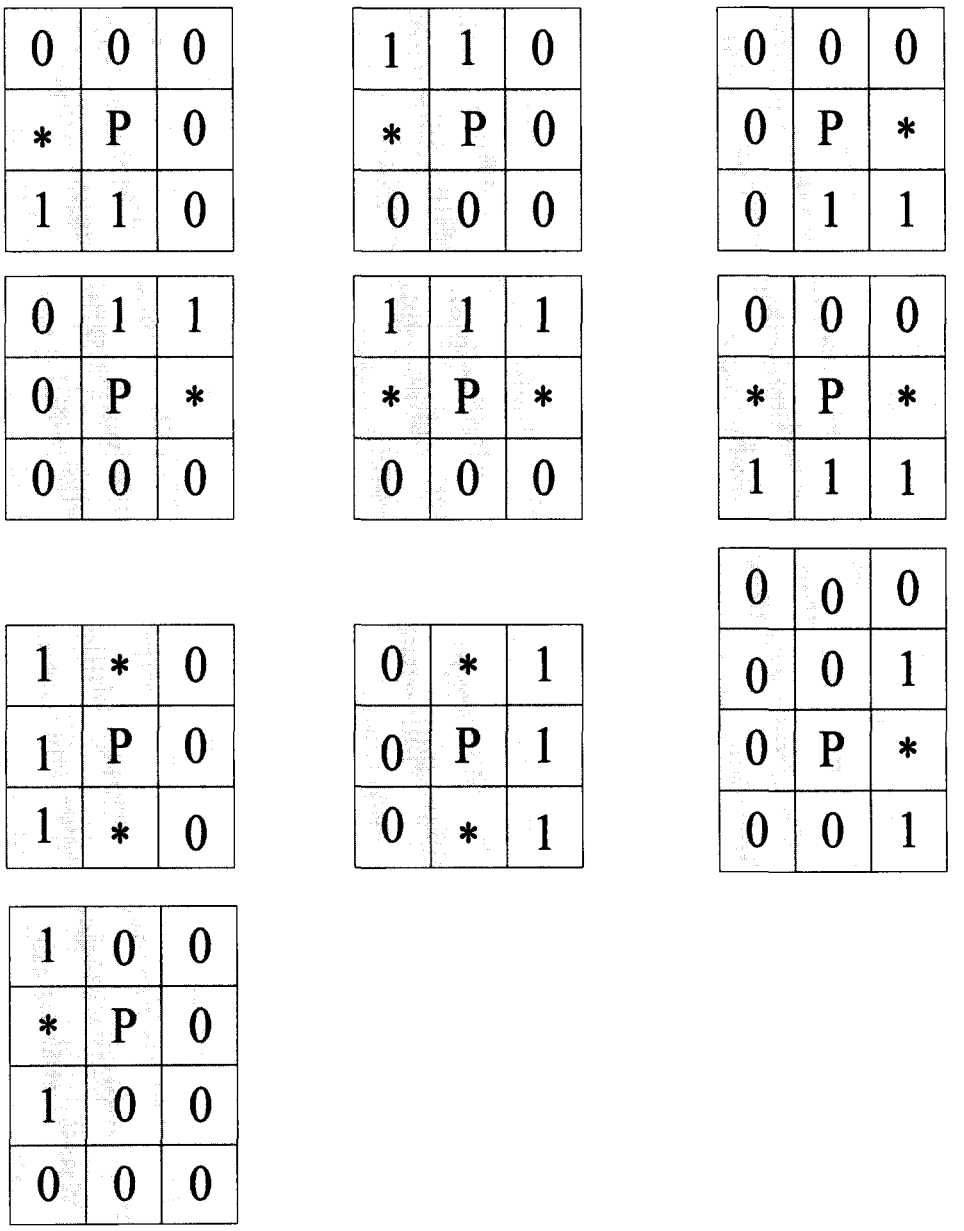


图 7-4　系统所选用模板

模板中，数值为0的部分表示像素为黑色，数值为1表示像素为白色，数值为*表示该点像素可以为黑色也可以为白色。

在实现软件设计的过程中，设计算法为并行处理算法，即一次性判断图像中的像素是否应该被删除后，再进行删除处理，因此在实现过程中为了标志一个像

素是否应该被删除，会将黑像素点的灰度置成灰色（像素灰度为100），因此在进行模板匹配时，将判断条件$P_k = 0$改为$P_k != 255$，即判断该点不是白色像素点。

7.4 实验结果分析

图 7-5 细化后图像

由细化后的简牍图像可以发现，细化后的文字图像虽然满足了细化条件，即骨架图像保持了原图像的连通性，保留曲线的端点，细化结果宽度基本为1个像素，但是在细化后图像中出现了严重的笔划分叉现象，如果将这种细化后的简牍图像送入特征提取模块进行特征提取，会导致特征提取模块错误提取文字笔划，进而导致错误识别或拒绝识别。

产生笔划分叉的原因在于文字笔划本身边沿并不是均匀分布的，图7-6所示

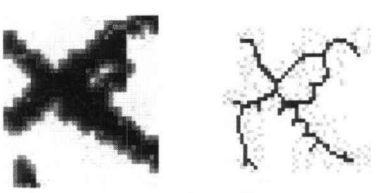


图 7-6 文字图像

为放大后的简牍文字图像，由图像可以看出，归一化后的文字边沿为锯齿形状（原始的文字边沿本身为锯齿状，经过归一化处理后锯齿状边沿被放大），这些锯齿状的文字边沿在进行细化处理过程后就成为了文字笔划的分叉点，会对部分文字特征造成影响。

7.5 改进算法

在文献[32]中提出了一种有效的文字细化算法，该算法采用了7个保留模板和50个删除模板，分别如下：

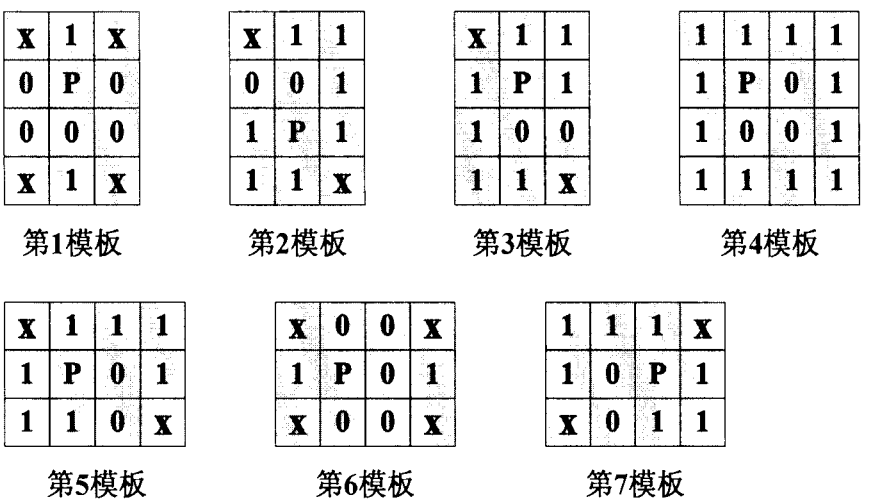
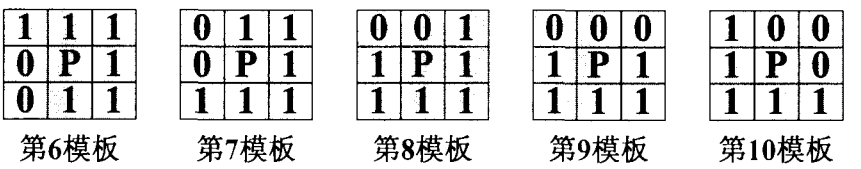
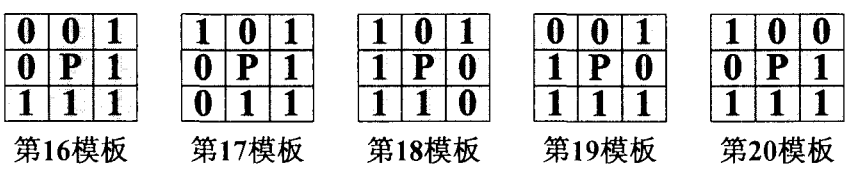
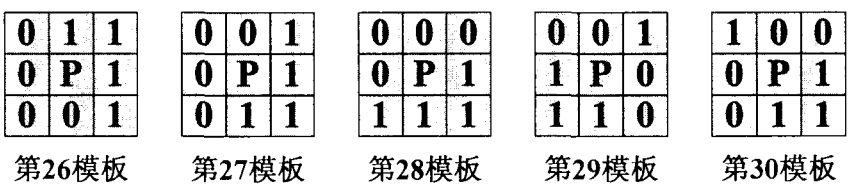
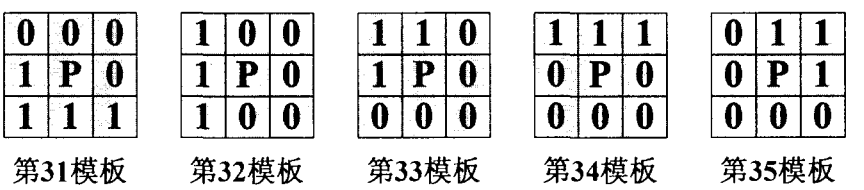


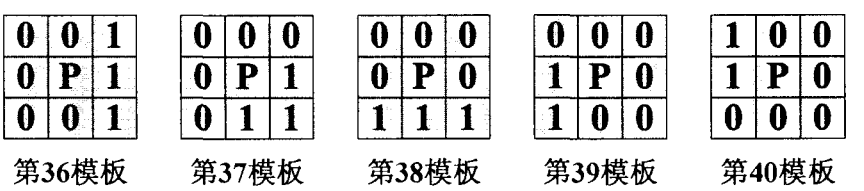
图7-8 保留模板

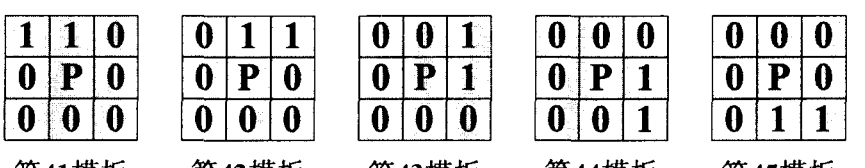












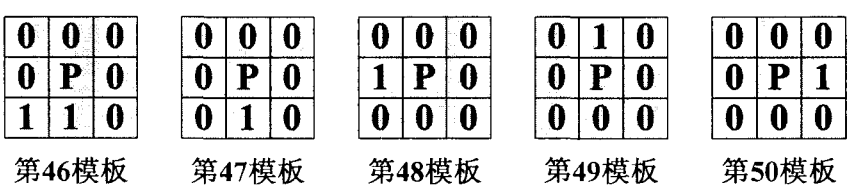


图 7-9　删除模板

图中标志为1的点为白像素点，标志为0的点为黑像素点，标志为P的点为当前像素点，标志为X点表示不关心该点的像素灰度。

运用该算法对文字图像进行细化处理，得到文字细化效果如下：

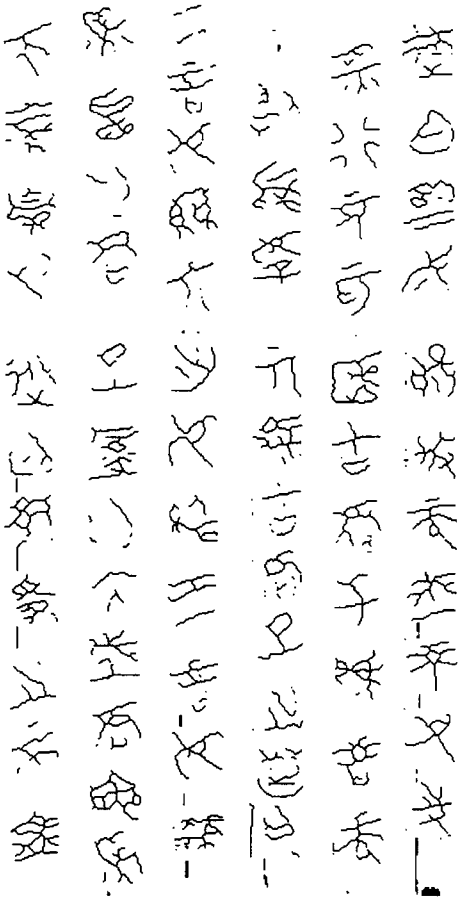


图7-10　文字细化效果图

由文字细化后效果图可以看出，细化后的文字图像笔划宽度为1个像素，同时文字笔划基本没有出现分叉现象，但是文字笔划出现了明显的断笔，如第3列第5个及第6个文字：

图7-11　文字断笔示意图

图7-12　断笔文字归一化后图像

由归一化后的文字图像明显看出，发生断笔的文字其归一化后的图像并没有断开，因此文字笔划的断开完全是由于细化算法对文字进行了过度细化造成的。

为了研究文字出现断笔的原因，在实验过程中，对文字的细化过程进行深入研究：

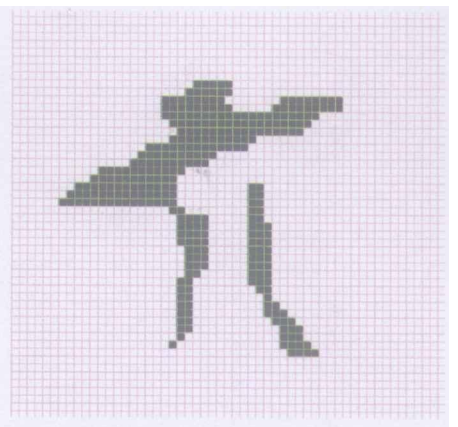


图7-13　进行一次细化算法后图像

图7-13为对文字进行一次细化处理后所获得的图像，由该图像知在进行一次细化后文字没有出现断笔现象。

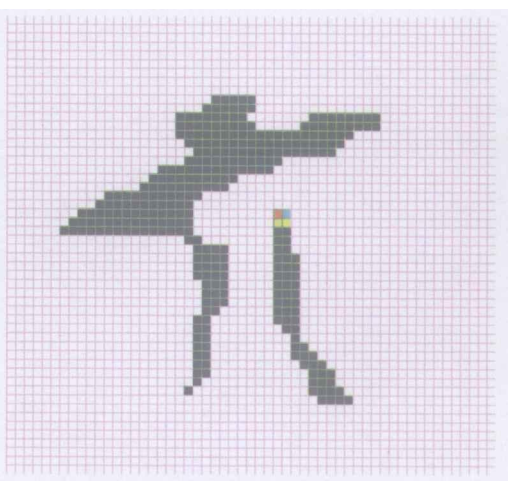


图7-14　细化示意图

但仔细分析图7-13，并且将其画成图7-14的示意图，红色像素点的位置满足删除模板12，浅蓝色像素点所在位置满足删除模板14，同时这两个像素点所在位置都不满足保留模板，因此会被删除。当这两个像素点被删除后，两个黄色像素点会满足删除模板12、14，因此也会被删除掉……依此类推，文字的大部分竖笔

划会被删除，必然会导致出现缺失笔划和文字断笔的情况。

这种情况出现的原因在于当笔划宽度为2个像素时，由于其两个像素会同时满足删除模板，因此会全部被删除。在设计简牍图像的文字处理系统时，针对该算法的缺点，本文设计了一种补偿算法，其实现步骤如下：

（1）当文字像素点满足删除模板10或12时，查看当前像素点右侧第二个像素点的灰度值，若灰度为0（即为黑像素点），设置当前像素点的灰度为白色；若灰度为255（即为白像素点），保留当前黑像素点，同时将 P_4 处像素点的坐标存入标志数组中。

（2）当文字像素点满足删除模板1或4时，将当前像素点 P_4 位置处的像素位置与标志数组中的坐标进行比较：若与标志数组中的某个坐标相同，则保留当前像素点；若在标志数组中没有坐标与 P_4 相同，将当前像素点的灰度置为255。

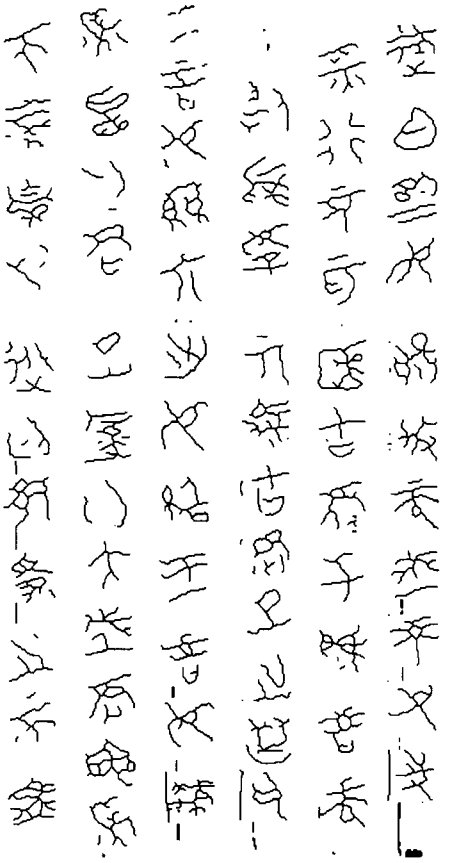


图7-15　细化效果图

图7-15为改进后的细化算法实现的文字细化效果图，由细化后的效果图可以看出：（1）文字笔划基本被细化成一个像素的宽度；（2）文字笔划出现的分叉很少；（3）文字笔划基本没有出现断笔；（4）细化后的文字图像基本保持了原文字

的拓扑特征。因此，该文字细化算法实现了细化目标，并且取得了很好的细化效果。

第八章 特征提取

8.1 常用特征

良好的文字特征应具有以下4个特点:

（1）可区别性。对于属于不同类的图像来说，它们的特征应具有较明显的差异。

（2）可靠性。对于同类的图像，特征值应该比较接近。

（3）独立性好。所选择的特征之间彼此不相关。

（4)数量少。图像识别系统的复杂程度随着系统维数(特征的个数)迅速增长。

目前常用的文字特征分为结构特征和统计特征，而统计特征按照特征提取区域的不同分为全局统计特征和局部统计特征。

8.1.1 统计特征

与结构特征相比，统计特征具有良好的抗噪声、抗干扰的性能，其鲁棒性主要体现在统计特征的抽取和模式匹配方法上。统计特征建立在二值或灰度值点阵图像基础之上，通常是对汉字点阵信息进行非线形变换后提取。尽管统计特征与汉字的本质特征有较大区别，但由于其具有多样性、互补性、易于提取等优点，它在手写体汉字识别领域中占有十分突出的地位[32]；与此同时，正由于统计特征没有反映出文字的结构信息，所以单纯使用统计特征无法区分出结构相似的文字。因此在现代的文字识别系统，都是将统计特征和结构特征结合起来进行文字的识别。

常用的全局统计特征有:

（1）正交变换特征。正交变换特征是对输入汉字点阵进行各种正交变换，取变换系数作为特征，常用的变换有Fourier变换、DCT变换、Walsh变换、Rapid变换、K-L变换等。

（2）不变矩特征。不变矩特征是一种线性特征，因其在尺度、平移和旋转等条件下的稳定性而被广泛用于模式识别领域，常用的不变矩有Hu矩，Zemike矩，Legendre矩等。

（3）笔画密度特征。笔划密度特征就是以固定扫描次数从不同方向扫描文字，计算和笔画相交的次数，形成笔划密度特征。通常取水平、垂直和两对角线四个扫描方向，每个方向取16个特征，共形成64个特征。这种特征描述了汉字的各部分笔画的疏密程度，提供了比较完整的信息[33]。

（4）笔划方向特征。笔划方向特征是以笔划在不同方向上的贡献度为基础。首先以汉字点阵图像的每一个黑像素P为中心，分别求该点沿8个量化方向至笔划边缘的距离，然后将相反两个方向的距离相加，求出规范化距离值。

局部统计特征有：

（1）背景特征[34]。从汉字点阵的背景着手进行分析所提取的特征称为背景特征.通常选取位于汉字图像两对角线上的背景点求取笔划密度，常见的背景特征有:四方向背景笔划密度特征、修正的四方向背景笔划密度特征、四方向背景笔划斜率特征等。

（2）Gabor特征。Gabor函数能够较好地兼顾信号在时域和频域中的分辨能力，用Gabor函数形成的二维Gabor滤波器在分析数字图像中局部区域的频率和方向信息方面具有优异的性能，因此，在计算机视觉和图像分析等领域得到广泛的应用。

（3）网格特征是指把$m\times m$像素构成的文字图像分割为$n\times n$个子块，在这些小区域中分别提取反映各子块特点的统计值。

8.1.2 结构特征

结构特征提取方法是人们最初用来进行手写汉字识别研究的方法，一般需要先抽取笔段或基本笔划作为基元，由这些基元再构成部件，由部件的组合来描述汉字，最后再利用形式语言及自动机理论进行文法推断，即识别。

（1）特征点特征。特征点是最基本的结构信息，包括端点、交叉点、断点和拐点。

（2）笔段信息。一般把笔段分为横、竖、撇、捺四种，通常用笔段的起点坐标、笔段长度和倾斜度进行描述。

（3）笔划信息。

（4）部件。汉字是由各种部件所组成，但部件的提取很困难，所以较少使用。

8.2 特征提取

文本数据的半结构化甚至于无结构化的特点，使得表示文本数据的特征向量高达几万维甚至于几十万维。高维的特征对分类机器学习未必全是至关重要的，有益的。高维的特性可能会大大增加机器学习的时间而仅产生与之小得多的特征子集一样的分类结果[35]。因此，在进行文本分类中，特征选择显得至关重要。通过映射(或变换)的方法可用低维空间来表示样本，这个过程叫特征提取。映射后的特征是原始特征的某种组合，所谓特征提取在广义上就是一种变换。

特征选择主要用于排除确定的特征空间中那些被认为无关的或是关联性不大的特性。于是经常会使用特征独立性假设以简化特征选择，以达到计算时间和计算质量的折衷。因此，目前在对文本的特征空间所采取的特征选择算法一般是构造一个评价函数，对特征集中的每个特征进行独立的评估。

8.2.1 点特征提取

用于点特征提取得算子称为有利算子或兴趣算子。自七十年代以来出现一系列各不相同、各有特色的兴趣算子，知名的有Moravec算子、Hannah算子与Foistner等。

下面以Moravec算子为例说明点特征提取：

Moravec算子的基本思想是，以像元的四个主要方向上最小灰度方差表示该像元与邻近像元的灰度变化情况，即像元的兴趣值，然后在图像的局部选择具有最大的兴趣值得点（灰度变化明显得点）作为特征点，具体算法如下：

（1）计算各像元的兴趣值IV(interest value)，例如计算像素点(x,y)的兴趣值，先在以像素点(x,y)为中心的$n\times n$的影像窗口中，计算四个主要方向相邻像元灰度差的平方和，取其中最小者为像素点(x,y)的兴趣值。

（2）根据给定的阈值，选择兴趣值大于该阈值的点作为特征点的候选点。阈值的选择应以候选点中包括需要的特征点，而又不含过多的非特征点。

（3）在候选点中选取局部极大值点作为需要的特征点。在一定大小的窗口内（可不同于兴趣值计算窗口），去掉所有不是最大兴趣值的候选点，只留下兴趣值最大者，该像素即为一个特征点。

8.2.2 笔段特征提取

通过链码跟踪的方法得到笔段轮廓的点列，然后从笔断点列中提取特征点并利用特征点切分出笔段，最后用笔段的轮廓线代替骨架线表示文字的笔段[36]。

8.2.3 基于像素数量的粗网格特征与外围特征

将归一化后的文字图像等分为8×8个网格，统计各网格内黑像素的数量，取得一个64维的网格特征。

外围特征提取步骤为：

把归一化后的汉字点阵等分为8行。

计算每一行从图像左边缘至第一项由黑变白的长度（或该行的面积）。

再计算每一行从图像左边缘至第二次由黑变白的面积。

仿照上面步骤，提取其他3个边缘的特征。

8.3 本文采用方法

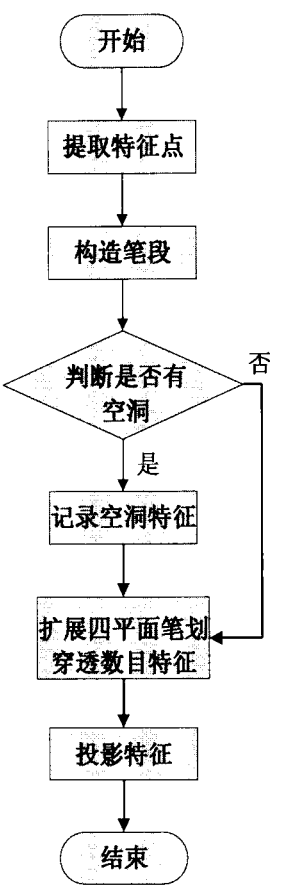


图 8-1 程序流程图

8.3.1 特征点提取

在特征点提取子模块，软件对每一幅文字图像的每一个文字像素点，求其周围8邻域的黑像素点个数，当黑像素点数目为1时，表示该点为端点；当黑像素数目为2时，该像素点可能为折点，也有可能为连接点；当数目为3时，表示该像素点为歧点，在印刷体文字识别过程中，歧点的出现是由于笔划相交现象，在进行手写体文字识别时，由于手写体文字本身存在笔划宽度分布不均的现象，导致细化后的文字图像横不平、竖不直，在文字笔划发生变形的位置同样会出现歧点；当数目为4时，表示该点为交叉点。由于无法一次性的在软件中准确的提取文字特征点，因此，在软件中运用链表记录每一个临时特征点的位置及其所属类别，准确的特征点数目及位置将在构造笔段模块获得。

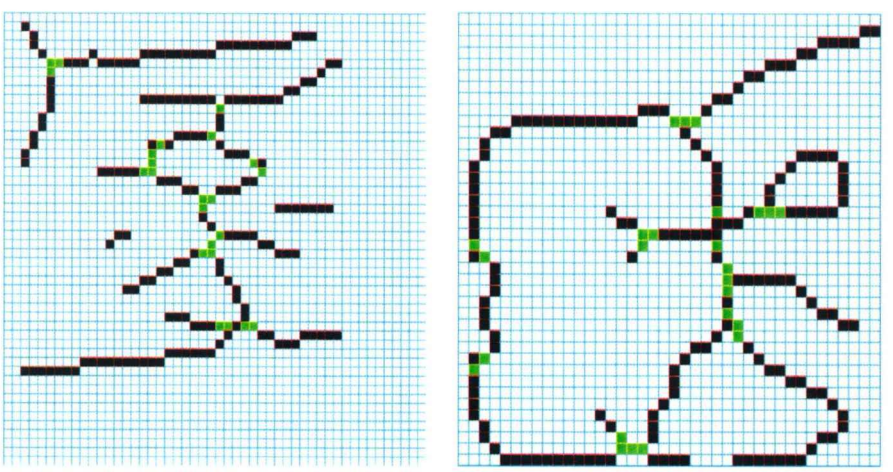


图 8-2 歧点效果图

图8-3为歧点提取效果图。由效果图可知，在文字笔划相交的位置，以及笔划出现转折的位置会同时出现多个歧点，这些歧点对于特征提取具有以下危害：（1）一般情况下认为，在文字笔划相交的位置只有一个歧点，多个歧点的出现增加了歧点特征的数目，造成歧点特征不正确；（2）会对笔划特征的提取造成困难，因此需要在处理过程中针对多个歧点相邻的情况进行处理。

在研究过程中，软件先对文字图像进行扫描，将所有的歧点存入临时位图中（歧点处像素设为0，否则为1），将歧点与其8邻域像素与以下六个模板进行比较，若满足其中一个模板，则只保留模板中心点的歧点，将其相邻歧点删除。

1	1	1
0	0	1
1	0	1

1	1	1
1	0	0
1	0	1

1	0	1
0	0	1
1	1	1

1	0	1
1	0	1
1	1	1

1	1	1
0	0	0
1	1	1

1	0	1
1	0	1
1	0	1

图 8-3 删除模板

然后找出临时位图中所有相邻的歧点，对于每一对具有相邻性的歧点（设为P_1、P_2），计算每个歧点（如P_1）与距离其最近的歧点的距离（不包括P_2），分别为D_1、D_2，若$D_1 < D_2$，删除点D_1，否则，删除点D_2。

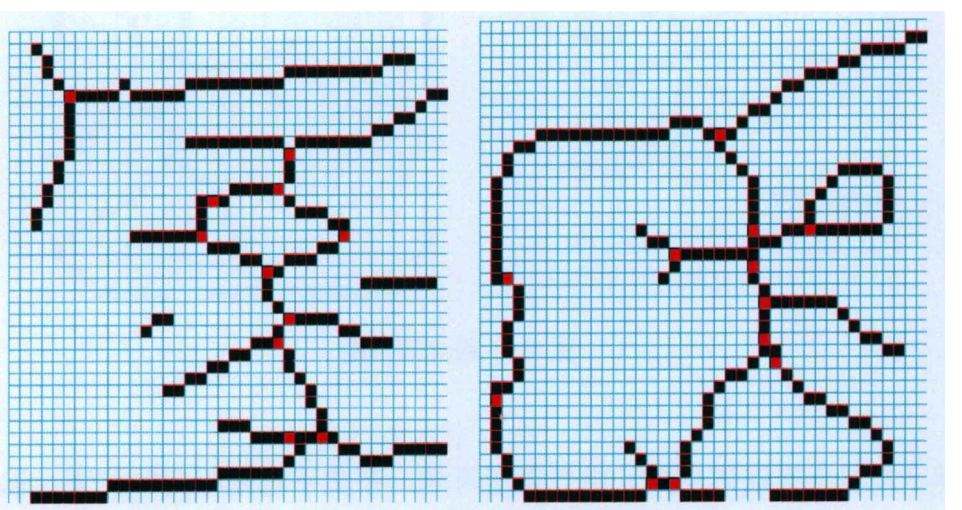


图 8-4 新歧点效果图

8.3.2 构造笔段

从特征链表中提取文字的端点（设为P_1），从文字端点开始寻找其8邻域的黑像素点(设为P_2)，判断P_2是否为连接点，若为连接点，将P_2点位置存入笔划链

表中，并将P_1灰度置为100（表示已经处理过该像素点），再以P_2点做为起始点继续寻找P_2点8邻域的黑像素点，依次类推；若不是连接点,，将P_1 、P_2及其周围8邻域点与8个连接模板相比较，若满足其中一个连接模板，将P_2位置存入笔划链表中，将P_2点灰度置为100，并将连接模板中灰度标记为0的像素点作为起始点，寻找其8邻域的黑像素点，依此类推，直至没有满足条件的像素点为止，计算笔段方向。当所有端点所连接的笔段搜索完毕后，再从特征链表中提取文字的折点、歧点和交叉点，在这些特征点的8邻域中寻找黑像素点，如果没有黑像素点，表示该特征点所连接的笔段已全部被搜索完毕，可以跳过；若存在有黑像素点，则同文字端点一样，寻找特征点的连通域，并将其所经历的黑像素置为灰色。由此模块可以得出文字的总体笔段数目，横、竖、撇、捺四种笔段的数目及其位置。

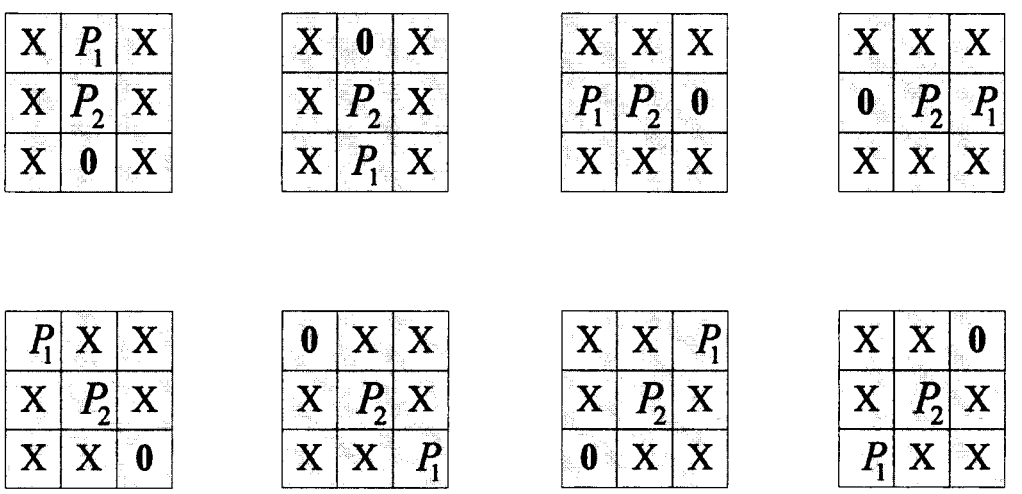


图 8-5 连接模板

8.3.3 判断是否存在有孔洞

将文字像素信息存入40×40的数组内。在文字图像最左侧一列从上至下寻找白像素，找到第一个白像素（设为P）后将P坐标放入一个像素链表中，并将P点灰度置为210，然后在P的8邻域内寻找白像素点，若白像素点满足图8-2所示的四个保留模板之一（图中1表示该点为白像素点，0表示黑像素点，X表示不关心该位置像素点灰度），保留该白像素点灰度，否则，将白像素点坐标存入像素链表

中，同时将其灰度值置为210，然后从像素链表中抽取第二个像素坐标，在该像素的8邻域内寻找白像素，如果存在有白像素，且白像素点不满足四个保留模板，将白像素的灰度置成灰色并将其坐标存入像素链表中，再从像素链表中提取第三个像素坐标；如果在第二个像素坐标的8邻域范围内没有白像素存在，则直接提取像素链表的第三个像素坐标进行处理，依此类推，直至像素链表尾。然后在文字图像的最右侧一列、最上一行、最下一行做相同的处理，如果在这些边界处不存在有白像素点，则直接结束处理。处理效果后如图8-3所示。

1	0	X
0	P	X
X	X	X

第一模板

X	0	1
X	P	0
X	X	X

第二模板

X	X	X
0	P	X
1	0	X

第三模板

X	X	X
X	P	0
X	0	1

第四模板

图 8-6 保留模板

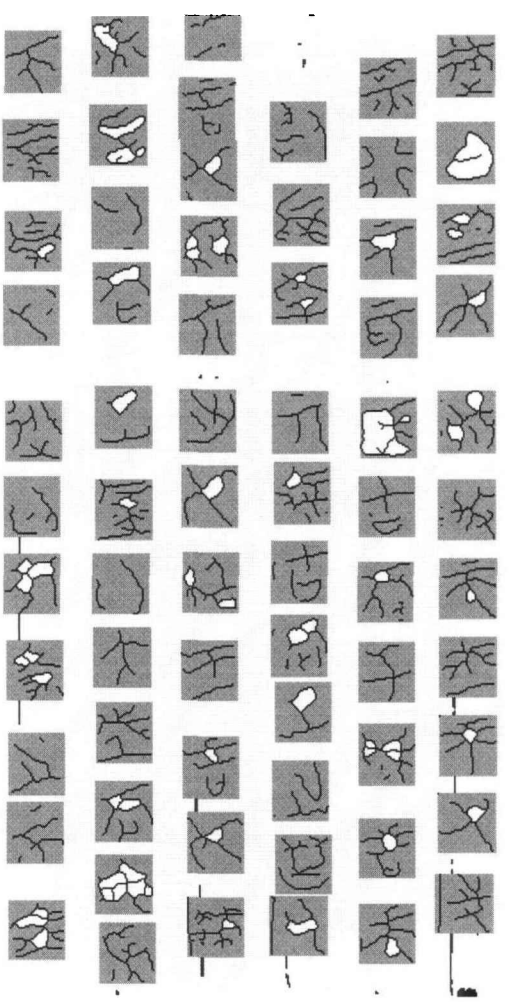


图 8-7 中间效果图

设文字孔洞特征变量的值为0，从文字图像的左上角开始从左到右、从上到下寻找白像素点，如果发现一个白像素点，将文字孔洞特征变量的值加1，同时将白像素点位置存入孔洞链表中，并寻找其周围8邻域的白像素点，若该白像素点不满足保留模板，将白像素点坐标存入孔洞链表中，并将链表中所有的像素点置成灰色（灰度值为80+20 x hole,其中hole为孔洞特征变量），然后将链表的第2个像素点坐标提取出来，做相同处理，直至链表尾。处理完成后再从文字图像的左上角开始寻找白像素点，做相同的处理过程，直至文字图像中不再存在白像素点，此时文字孔洞特征变量的值就是文字中存在的孔洞数目。图8-4为孔洞特征提取效果图。

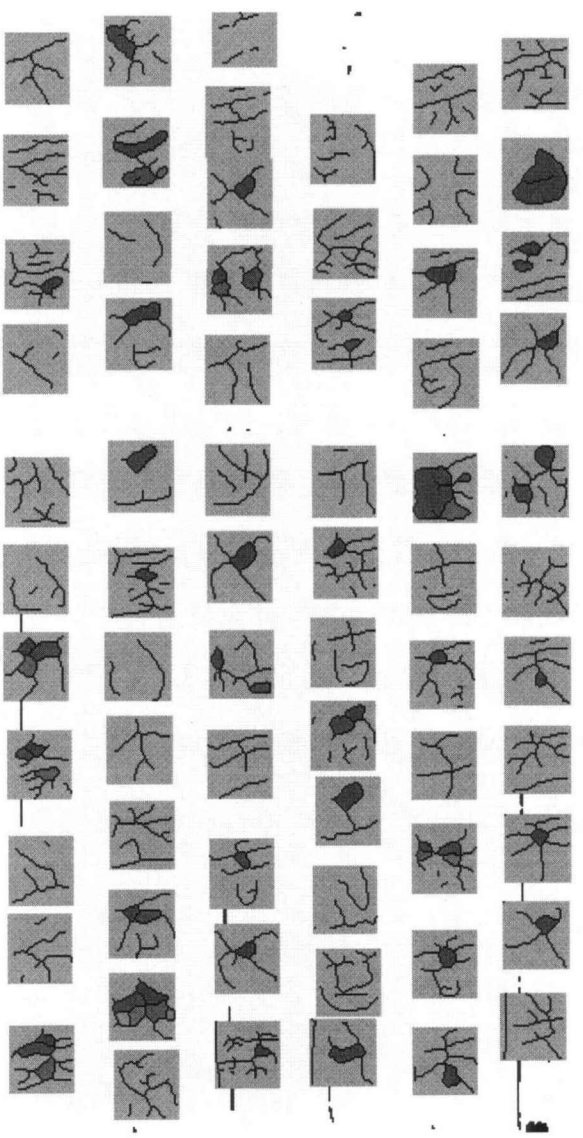


图 8-8 算法效果图

8.3.4 扩展四平面笔划穿透数目特征

基本的四平面笔划穿透数目特征从文字的水平、垂直、主对角线及副对角线四个方向划出8条扫描线，计算每条扫描线与文字笔划的相交数目，可以得出32维特征相量。

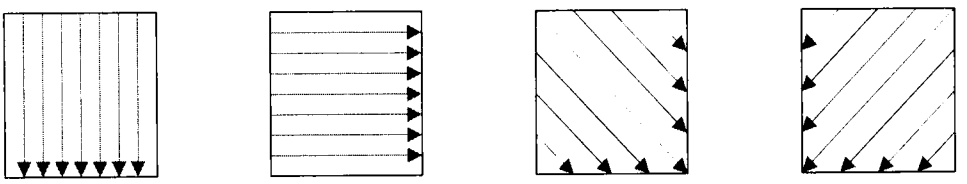


图8-9 扫描示意图

在设计特征提取的过程中，本文设计了一种扩展的四方向笔划穿透数目特征。

研究发现，虽然手写体的文字会出现各种变形，但是文字的笔划在书写顺序上变形很小，比如“小”字，无论书写“小”字的人书写习惯如何 ，当从左至右对“小”字进行扫描时，扫描线与文字笔划相交的顺序基本为先与撇笔划相交，再与竖笔划相交，最后与捺相交，因此，扫描线与文字笔划相交的顺序应该是比较稳定的特征，受文字手写变形较小。扩展四方向笔划穿透特征在进行扫描过程中，不是简单的记录下与扫描线相交的笔划数目，而是记录下与相交线相交的横、竖、撇、捺四种笔段的先后次序（在软件设计过程中，与横、竖、撇、捺四种笔段的相交次序用二进制数字表示，数字位为0，表示没有扫描到该种笔段，为1表示扫描到该种笔段。如横笔段在扫描后得到的数字为001，表示在进行该次扫描时第三次穿过的笔划为横笔段），这样得到的特征维数为 $4\times8\times4=128$ 维持维特征相量。

8.3.5 投影特征

从水平和垂直两个方向对汉字进行投影，统计投影轴穿过的笔划数目。

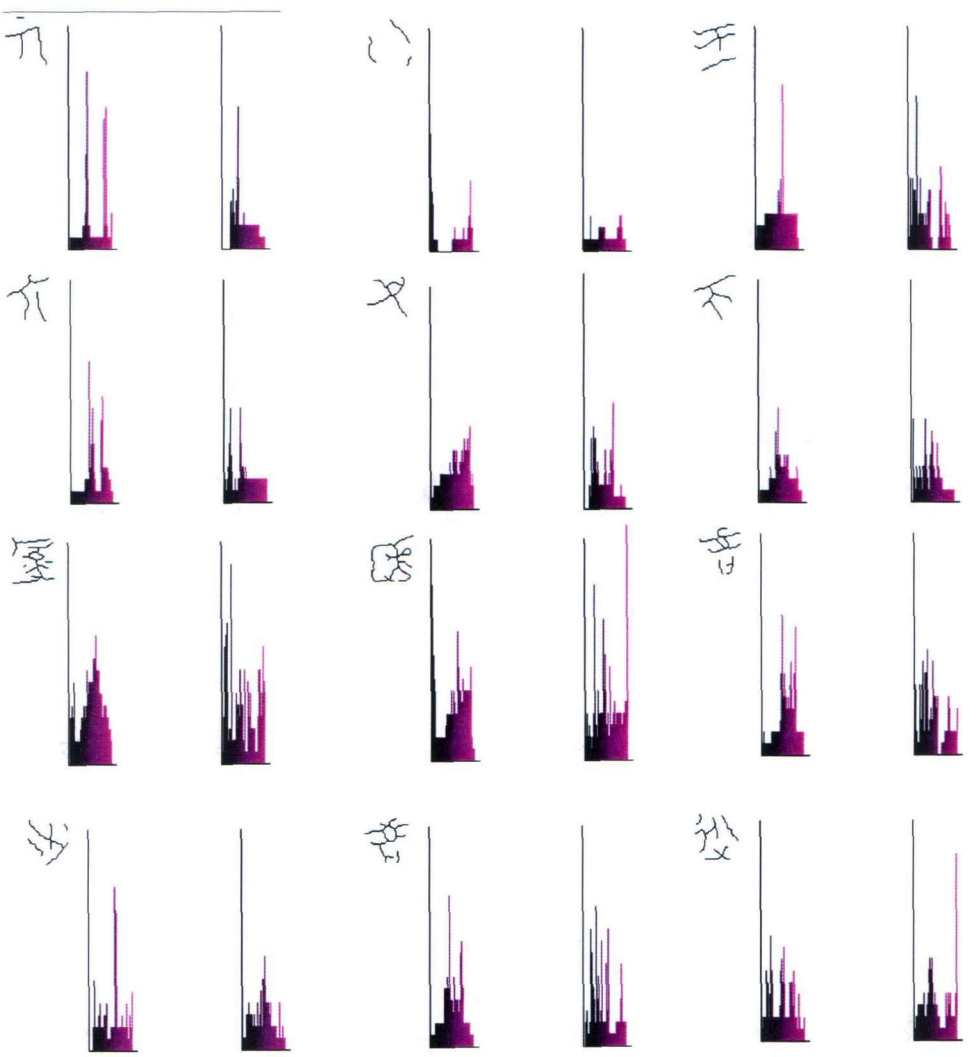


图8-10 投影特征示意图

图8-10为将投影特征放大10后所得投影特征示意图，其中每个文字具有两个投影特征图，左边为垂直投影特征，右边为水平投影特征。

结 论

总结

本文在深入研究了简牍图像的特点后，提出了一系列的简牍图像文字识别处理算法，主要工作体现在以下几个方面：

（1）在二值化处理模块，针对文字图像特点-文字的笔划边沿与文字背景之间在8邻域方向上存在灰度差异-提出8邻域灰度差值法，并运用该方法成功的将简牍的灰度图像转换为只有黑白两色的二值图像。

（2）根据二值化后简牍图像特点，提出利用投影图进行文字粗切分，利用后续补偿措施进行文字细切分的文字切分算法。经过实验验证，该种文字切分算法能够准确的将简牍上面的文字提取出来，同时具有运算速度快、能够去除边沿噪声的特点。

（3）改进已有文字细化算法，增加两个保留模板和1个标志数组。新的文字细化模板在保持原文字连通性的前提下将文字笔划细化到一个像素的宽度，并且避免文字笔划出现分叉现象。

（4）研究已有文字特征，提出了一种新的文字特征—扩展四平面笔画穿透数目特征，该特征具有特征提取容易、基本不受文字手写变形影响的特点。

（5）提出一种提取文字孔洞特征的软件实现算法，该算法具有运算速度快、提取孔洞数目准确的特点。

虽然本文的研究工作已取得很大进展，但在很多方面还存在有不足：

（1）8邻域灰度差值法还不能实现自适应图像二值化，即当简牍图像中竹简（或木牍）的颜色发生变化导致灰度化后的竹简（或木牍）灰度值有较大范围内的变化时，还不能通过算法自身进行调节，需要人为的改变调节系数α的值。

（2）没有实现分类器设计及后处理模块。

展望

针对本文所存在的不足，今后的研究工作主要集中在以下几个方面：

（1）改进8邻域灰度差值法。8邻域灰度差值法之所以不能实现自适应二值化，原因在于α是一固定值，如果能够找出竹简（或木牍）的背景颜色及文字边沿颜色与α之间的关系，则可以实现文字的自适应二值化算法。

（2）继续进行分类器设计及后处理模块研究，选择合适的文字识别分类器，并建立相应的文字特征库。

致 谢

衷心感谢导师王绪本教授对我的悉心指导。在我读研究生期间，王老师在学习、科研上给了我无微不至的关怀和照顾，也正是王老师的精心选题和悉心指导，我的论文才得以最终完成。王老师严谨的治学态度，认真勤奋、不知疲倦的的工作作风，是我学习的榜样。

感谢所有从事文字识别研究的科研人员，正是因为有了他们的智慧与经验，才有我今天的进步。

感谢张娜、杨思涵、李文超、阚瑗珂同学及课题组的其他同学，与她们的讨论中我获得了许多新的知识与灵感，感谢他们热心的帮助。

感谢我的父亲和母亲。在我读研究生期间，他们给予精神上的无私帮助。

感谢我的爱人，感谢他在我成功和失败时，对我始终如一的信任与支持。

感谢在我论文完成期间，所有支持和帮助过我的人们!

参考文献

[1] 陈　阳,中文古籍数字化的成果与存在问题[J]，出版科学,2003，4:47—48.

[2] 史睿,试论中国古籍数字化与人文学术研究[J]，国家图书馆学刊,1999,2:.

[3] 吴家驹,中文古籍数字化的进展与主要成果述评[J]，南京师范大学文学院学报,2004,3:178—183.

[4] 王国钧，邵斌，周昌乐，中文电子签名的认证方法研究，计算机应用研究，2004,21(4)：109—111

[5] 贾广威.印刷体汉字图像处理及基于结构特征相似字识别的研究：[D] 哈尔滨：哈尔滨工业大学，2003.

[6] Shunji.MORI,Historical Review of OCR Research and Developmen″,Proceedings of The IEEE, 80(7), 1992:1029—1058.

[7] 高彦宇,杨扬.脱机手写体汉字识别研究综述[J].计算机工程与应用,2004,40 (7)：7477.

[8] tag. 图像的灰度化原理和实现 [EB/OL] http://blog.csdn.net/chenamo9651/archive/2006/07/06/886699.aspx.

[9]陈兵旗，孙明，Visual C++实用图像处理专业教程[M].北京：清华大学出版社，2004.

[10] 林宗辉.数码电视影像处理元件与滤波技术探究[EB/OL]. http://gb-www.digitimes.com.tw/gate/gb/tech.digitimes.com.tw/print.aspx?zNotesDocId=0000047354_A348HL2CSM61M7L3FPK3K .

[11] 王鲁敏.基于图像分析方法的颗粒粒度测量[J].化工装备技术，2005,26(4):65-67.

[12] 樊崇艺,宫国.，灰度图像二值化在汉字识别中的应用[J]教育信息化，2001,10月:44—46.

[13] 白 洁,杨耀权,陈余梅.基于贝叶斯算法的二值化算法[J]华北电力大学学报,2007,34(3):65—67.

[4] 李了了，邓善熙，丁兴号.基于大津法的图像分块二值化算法[J].微计算机信息，2005 ,21(08x):76—77.

[15] Kenneth.R.Castleman，朱志刚等译.digital image processing[M].北京：电子工业出版社，2001.

[16] 梁华为.直接从双峰直方图确定二值化阈值[J].模式识别与人工智能，2002，15(2):

253—256.

[17] 陈丹，张蜂，贺贵明.一种改进的文本图像二值化算法[J].计算机工程，2003,29(13)：85—86.

[18] 庄军,李弼程,陈刚.一种有效的文本图像二值化方法[J].微计算机信息，2005 21（8）:56—57.

[19] http://cslin.auto.fcu.edu.tw/scteach/scteach88/text1/2-7.htm.

[20] 赵云.手写体汉字的计算机识别研究:[D].武汉：武汉理工大学，2004.

[21] Casey R G，Lecolinet E. A Survey of Methods and Strategies in Character Segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,

1996，18(7)：690—706.

[22] 陈艳，孙羽菲，张玉志。灰度图像中字符切分方法的研究[J]。中文信息学报，2004，18（4）：44—49.

[23] 明德烈，柳健，胡家忠，李海涛.一种改进的手写汉字文本切分算法[J]华中理工大学学报，2000，28(21):87—89.

[24] 王琳琬，杨 扬，颉斌，杨 毅.基于连通域单元和穿越算法的汉字切分[J].信息技术，2004,28(4):30—32.

[25] 张捷.手写数字识别的研究与应用:[D].西安：西安建筑科技大学 ，2004.

[26] H.Ya mada,KYamamoto,T.Saito.a nonlinear nonmalizaiton method for handprinted Kanji character recogniiton-line density equalization.Pattern Recogn,pp, 1990:

1023-1029.

[27] S.W.Lee,J.S.Park.Nonlinear shapen onnalization methods for the recognition of large-set handwritten characters.,Pattern Recogn,1994,27:895—902

[28]贾广威.印刷体汉字图像处理及基于结构特征相似字识别的研究:[D]. 哈尔滨:哈尔滨工业大学,2003.

[29]王家隆.指纹图像的预处理及其改进算法:[D]. 大连:大连理工大学，2003.

[30] 陈瑞改，陈怀新.干涉条纹中心线提取与细化的新方法[J].激光杂志,2005,26(5):40—41.

[31] 杨义军.基于三角剖分的带状图像细化算法的研究:[D].济南：山东大学，2006.

[32] Lei Huang,Genxun Wan,Changping Liu.An Improved Parallel Thinning Algorithm[M]: Document Analysis and Recognition,2003.Proceedings.Seventh International Conference on,2003:780-783.

[33] 李美丽.基于数学形态学的脱机手写体汉字识别:[D].北京：北京科技大学，2007.

[34] 吴佑寿，丁晓青.汉字识别原理方法与实现[M].北京:高等教育出版社，1992.

[35] 张析中.汉字识别技术[M].北京:清华大学出版社，1992.

[36] 彭时名.中文文本分类中特征提取算法研究:[D].重庆：重庆大学，2006.

[37] 王浩军，赵南元，邓钢轶.一种现代藏文笔段提取算法[J].中文信息学报，2001,15（4）:41—46.

[38] 张宏林编著.Visual C++数字图像模式识别技术及工程时间[M].北京：人民邮电出版社，2003.