分类号_____ 密级_____

UDC _____ 编号 10736

西北师范大学

专业硕士学位论文

简牍文字提取与识别研究

研 究 生 姓 名：张兰云

校内指导教师姓名、职称：杨得国 教授

校外指导教师姓名、职称：

专 业 学 位 类 别：工程硕士

专 业 学 位 领 域：计算机技术

二〇一七年五月

Slips of Character Extraction and Recognition Research

Zhang LanYun

郑 重 声 明

本人的学位论文是在导师指导下独立撰写并完成的，学位论文没有剽窃、抄袭、造假等违反学术道德、学术规范和侵权行为，否则，本人愿意承担由此而产生的法律责任和法律后果，特此郑重声明。

学位论文作者（签名）：张芝马

2017年 5月 31日

同意定稿

学位论文使用授权书

本论文作者完全了解学校关于保存、使用学位论文的管理办法及规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，接受社会监督。本人授权西北师范大学可以将本学位论文的全部或部分内容编入学校有关数据库和收录到《中国博士/硕士学位论文全文数据库》进行信息服务，也可以采用影印、缩印或扫描等复制手段保存或汇编本学位论文。

本论文提交□当年/□一年/□两年/□三年以后，同意发布。

若不选填则视为一年以后同意发布。

注：保密学位论文，在解密后适用于本授权书。

作者签名：张芝云

导师签名：

2017 年 5 月 31 日

西北师范大学研究生学位论文作者信息

论文题目	简牍文字提取与识别研究		
姓　　名	张兰云	学　　号	2015221317
专业名称	计算机技术	答辩日期	2017.05
联系电话		E_mail	
通信地址(邮编)：			
备注：			

目 录

摘　要

中国的古文化博大精深，源远流长。其中古汉字是我们了解中国历史，政治，经济，人文风俗的载体。在最初时简牍是最主要的文字载体，直到纸的出现，才结束作为文字载体的历史，因此简牍文字对民族文化的继承和发展具有重要的意义。简牍文字具有异体字多，局部形变，大量噪声等特点，对于人工的简牍文字研究不仅需要大量的人力物力，并且进度缓慢，因此运用先进的现代科技的数字图像处理、模式识别等技术，进行简牍文字提取与识别的研究有重大的意义。

本文以数字化图像处理和模式识别作为理论指导，首先对简牍文字图像的样本进行灰度化、二值化、去噪、细化、反色以及归一化等预处理，并在图像去噪中提出新的自适应加权中值滤波算法，采用分数阶微分的方法对图像噪声进行检测，对含噪声的图像依据噪声点的数量自适应确定滤波窗口大小，并自适应计算权值，对噪声点进行加权中值滤波。结果能达到更好的去噪效果并保留了更多的文字细节，有利于后续的文字识别。然后进行特征提取，特征提取作为文字识别中必不可少的步骤，本文提出了双弹性局部密度特征和整体广义密度特征，部件结构特征这三个特征融合的方法，弥补了单一特征的不足，较好的满足了高稳定性，准确性。最后在标准 BP 神经网络和改进 BP 神经网络中对简牍文字进行识别研究，实验结果表明多特征融合具有较高的识别率从而也证明了多特征融合的有效性。

关键词：简牍文字；新的自适应加权中值滤波；多特征；BP 神经网络

Abstract

China's ancient culture is profound and profound. Among them, ancient Chinese characters are the carrier of our understanding of Chinese history, politics, economy and humanistic customs. At the beginning of the bamboo slips is the most important text carrier, until the emergence of paper, before the end of the text as a carrier of history, so bamboo slips of national culture inheritance and development of great significance. The text of the slips has many characteristics, such as different characters, local deformation and a lot of noise. It is not only a lot of manpower and material resources, but also slow progress, so the use of advanced modern technology, digital image processing, pattern recognition technology The research on the extraction and recognition of slips and phrases is of great significance.

In this paper, digital image processing and pattern recognition as a theoretical guide, first of the bamboo text image samples gray, binarized, denoising, refinement, anti-color and normalization and other pre-processing, and image denoising A new adaptive weighted median filtering algorithm is proposed. The method of fractional order differential is used to detect the image noise. The noise-free image is used to determine the size of the filter window according to the number of noise points, and the weight is calculated by adaptive. Point weighted median filter. The results can achieve better de-noising effect and retain more text details, is conducive to follow-up text recognition. And then the feature extraction and feature extraction are the indispensable steps in the character recognition. This paper presents a method to fuse the three characteristics of the double elastic local density and the generalized generalized density and the structural features of the component, which makes up the deficiency of the single feature, Better meet the high stability, accuracy. Finally, the recognition of the slips in the standard BP neural network and the improved BP neural network is carried out. The experimental results show that the multi - feature fusion has a high recognition rate and also proves the validity of the multi - feature fusion.

Key Words: Slips characters； New adaptive weighted median filter； Multi-feature； The BP neural network

1 绪论

1.1 研究背景

历史在我们心中永远带着神秘的面纱，我们对它充满了憧憬与向往，而能让我们窥探到一丝真相的就是展示在博物馆里的文物，这些古文物是曾经历史岁月的见证，是历史发展的遗迹，更是现在人们心目中珍贵的民族文化。中华民族拥有着上下五千年的悠久历史，源远流长，中国作为享誉在世界上的文明古国，有着丰富的人文遗产。这些经考古人员挖掘出的历史文物是我们祖先智慧的结晶和辉煌历史的见证，深入研究这些古文物有利于中华民族文化的传承与发扬，更快的实现我们中华民族的伟大复兴。

自二十世纪以来，我国激励着众多的研究人员在考古发掘和研究历史文献上进行探索，在政策上给予了极大的支持，因此在这方面也取得了巨大成就。根据对我国挖掘和保护的历史文物进行了第三次的普查得到了最新的数据，目前我国统计不可移动文物的数量达到了将近 77 万，其中包括 2352 处属于在全国中都要作为重点保护的对象，被大众认可的具有历史气息比较有名的名镇名村有 350 处，被国家鉴定审核公布的具有厚重的历史感的名城有 118 处 [1]。这些众多的文物会为我们展现一幅幅历史的画卷，为了充分了解历史的政治，人文风貌，考古工作者谨慎进行着文物修复工作。但是仅仅依靠传统的修复技术，不仅不能满足现在的需求，而且会在一定程度上对文物造成两次伤害。随着科学技术的发展，提高历史古文物保护的有效性，是一个比较紧迫而且又富有挑战性的问题。

在众多的古文化研究领域，古文字占着举足轻重的地位，通过古文字的演变我们见证了历史的发展历程。其中简牍是最早的文字载体，但是在中国考古研究中，它是在发现了甲骨文、卜辞、敦煌的文书等这些古文物之后才被人们挖掘出来的，因此也填补了中国缺失的古文献资料。并且它为我们提供了大量的历史信息，为历史研究学者提供了可靠的参考文献，具有极高的史料价值[2]。

1.2 研究意义

简牍在被挖掘出土之前埋藏在地下度过了漫长的历史时间，并且简牍实质上就是最早书写文字用的竹片或者木片，这些材质都是易受腐蚀的有机物，在简牍文物出土时一定要采取保护措施，不然就会被氧化从而加快腐烂的速度。其实大多数出土的简牍是不完整的，断裂的，它们处于地下时已经在一定程度被腐蚀了，并且由于是堆砌摆放更出现了一定程度上的粘连。因此写在简牍上的文字大多数变得模糊不清。这就使得人工对简牍文字进行修复时存在极大的不便，如果人工

操作不当，甚至会对这些简牍文物造成二次伤害。但是对这些简牍古文物的研究是全人工进行的，对粘连在一起的简牍进行揭剥处理，为了能数字化永久的存储简牍的信息对每一枚简牍进行拍照处理，为了更好地保存简牍文物减缓腐蚀的速度需要进行细致的脱水处理。对这些简牍进行深入研究的人员需要将每一枚简牍登记入库编码整理，因此这些传统的简牍文物处理方法需要耗费大量的人力物力，同时还需要简牍专家细心的在大量的简牍图片中对残缺的破损的断裂的简牍进行人工拼接，完整的修复简牍竹片并对其中文字的有效信息进行提取的工作，其巨大的工作量是可想而知的。

当今，将快速发展的计算机信息化技术应用于简牍这些古文物的研究是寻求创新技术的契机，更是弘扬民族文化的一个重大突破方向。运用先进的数字化图像处理技术对简牍图像进行处理，运用模式识别技术对简牍文字进行识别可以降低投入的费用，提高工作的效率，使研究工作化繁为简，并且提高了文献释读整理的速度，使研究者能更好的对简牍文物进行管理和保存。

本文主要是为了能够减少对简牍文物的伤害，便于考古工作人员的研究，节省人力物力，减省繁琐的工程步骤。研究目标为能够提取简牍图像文字的信息和识别简牍文字，将先进的计算机技术运用到了简牍文物保护中，不仅为考古工作者减少了工作量，更加提高了工作的可靠性和准确性。

1.3 国内外研究现状

1.3.1 数字图像处理技术研究现状

当今，随着数字图像处理技术的发展，应用数字化图像技术的领域越来越广泛。因此将数字图像处理技术引入到文物保护工作中是国内外文物保护工作者的新的探索历程。大多数从事研究的工作者首先将文物实物数字化图像后，利用计算机进行存储，这样就可以防止对文物的二次伤害，实现了对古文物的保护。

文物数字化最早开始于 1978 年，美国 OCLC 和 RLIN 它们创建了大型的数字化的图书数据库，它的功能就是将世界的历史文献做了数据汇总，从而方便人们进行资料的查询，检索。之后 1994 年“美国记忆”项目正式启动，在 2000 年实现了全国的民众都能够进入国会图书馆获取查询将近 500 万的文献资料[3]。欧盟在 1999 年启动多国合作项目“内容创作启动计划”，在 2000 年完成了 25 万幅图像数字化工程，进入 21 世纪对提出的 3D-Murale 项目、ARCHEO-GUIDE 项目进行深入的探索，从而对本国的古文物研究做出了重大贡献，在现实中对于这一方向的研究起着引导的作用[3,4]。20 世纪末，日本运用三维扫描实现了虚拟

重建大佛主殿，并制定了“全球数字化博物馆计划”，支持网络浏览、编辑以及博物馆教育功能[5]。

20 世纪 70 年代后期，随着计算机网络技术，数字化图像处理技术，虚拟现实技术等先进技术的兴起被我国大多数的从事科学技术研究的人员接纳与吸收并拓宽到新的领域。考古学与古文物保护利用这些技术进行了全新的探索。从 1997 年开始，古代敦煌壁画项目在潘云鹤院士以及鲁东明教授领导下探究出了在多颜色空间上对壁画进行分割，实现了准确的边缘处理以及通过学习人为绘制的样本，修补了壁画中严重缺失的部分，并仿真模拟了壁画色彩的演变过程，最终实现了壁画的真实感虚拟展示[6]。随着计算机技术的发展，数字化图像处理更加深入的进入我们的视野，我国文物保护也更多地采用数字化技术，如数字博物馆的建立，数字图书馆的建立。也有更多的学者利用数字化图像处理技术在文物方面进行研究，如桂恒[7]提出对扫描成像的碑帖文字进行骨架化和笔画提取，基于 Voronoi 图的框架下利用三角节点中的垂直二分线，实现文字骨架的提取并在进行插值的情况下实现文字的笔划修复，通过该算法最终实现了碑帖文字的数字化复原。

在对简牍图像文字处理上，张伟[8]通过建立损害较小的简牍文字库，对每一个简牍文字完整的轮廓曲线用 Canny 边缘算子进行提取，并利用文字书法特征进行了文字轮廓填充，通过人机交互的模式实现了简牍文字轮廓的修复并且使用了虚拟现实技术对简牍文字进行了在网络平台三维虚拟展示；刘瑛[9]根据简牍文字图像的特点，用八邻域灰度差值算法，垂直投影法切分算法，改进细化算法对简牍图像进行了预处理，在简牍文字的特征提取上对笔画穿透算法进行了扩展；覃庆炎[10]主要是依照高斯领域针对使用高、中、低三种尺度的 Retinex 算法，比较简牍图像增强的效果，进行了实验分析，采用多尺度 Retinex 算法与其它数字图像处理技术相结合既使图像的色感有了较好的一致性，并且又对阴暗的细节有效的进行了增强处理，最终在图像增强的方面得到最好效果。张娜[11]首先提出在 HSV 空间实现简牍图像增强，并把线性变换、中值滤波和高斯-拉普拉斯这三种方法组合起来达到简牍图像去噪的效果，在图像分割中主要根据的是亮度信息，最后利用水平垂直投影对简牍文字进行识别修复。张阳洁[12]针对简牍图像的特点，利用直方图变换引入了自适应的模糊阈值分割算法并将遗传算法和最大类间方差法相结合，最终达到了简牍图像分割的良好效果。

1.3.2 文字识别技术研究现状

OCR（Optical character Recognition）是 1929 年德国科学家 Taushek 提出的，因此人类进入了文字符号识别的时代。首先，日本和少许发达国家主要针对印刷体数字，字母进行研究。在 1958 年，日本的 ETL 成功的研制出了印刷体的手写字母数字识别系统 [13]。多数的科研人员对于汉字识别这一领域的研究风潮开始于 20 世纪 60 年代。美国 IBM 公司的 Casey 和 Nagy，在 1996 年通过使用基础的模板匹配法进行了最早的汉字识别研究[14]。随着 OCR 技术的发展，被越来越多的人关注，大量的研究学者在这个领域进行着新的探索，20 世纪 70 年代，这个领域在日本掀起了热潮，大多数的日本研究学者针对 OCR 技术进行了大量的研究[15]，例如在 1977 年，东芝研究所研究出了单个字体的印刷体汉字识别系统，能够识别出将近 2000 多个汉字。在 20 世纪 80 年代初期，武藏野电器研究所研究成功了可以识别将近 2300 个汉字，错误率只有 0.02%的多体印刷体汉字识别系统[16,17]。这些大量的成功的研究成果更加激励了越来越多日本学者专注于有关手写汉字方面的识别的研究[18,19]。

与国外相比，对于汉字识别这方面进行深入的研究工作，我国是从 20 世纪七十年代末期才开始进行的，起步相对较晚。但是对于这方面的工作我国政府一直充分的给予了大量的支持和足够的重视，越来越多的中国学者参与到这方面的研究中，一直处于不断地上升阶段。例如在 1986 这一年，沈阳的自动化研究所的研究人员已经开发出了一种识别印刷体汉字的系统，与此同时，朱夏宁和其他的研究学者开发出了一个可以识别 6763 个的印刷体汉字识别系统[20,22]。之后在 1990 年北京大学开发了第一个识别率为 68.5%-89.91%的手写体汉字识别系统。到 1995 年，863 组织的国家一级标准的脱机手写汉字测试了 37535 个样本，最好的识别率为 89.91%[19]。众多的研究人员发表相关的论文，通过实际应用，大量的识别研究系统在不断地完善和改进中，现在，在社会上发行的产品有汉王 OCR、清华 OCR（T-H-OCR）、曙光 OCR 等系统，能被大多数人们使用才能体现它们的价值[21,22]。

随着科学技术的发展，数字图像、模式识别、机器学习、人工智能等技术被越来越多的研究学者运用到实际生活中，为人们的生活增添了越来越多的便利，应用前景越来越广阔。目前，简体汉字识别技术更多的是在商业化的道路上继续迈进。而对于考古学研究，古文献研究，古文物研究等专业领域的信息化、数字化是我们急切需求的，一个值得深入研究的方向。在 1996 年周新伦、李峰等人，

将甲骨文字符抽象为无向图，运用图论编码的方法通过三级分类的编码识别，实现了识别率达 94%的甲骨文自动识别研究[23]。2002 年中国古文字字库《古文字诂林》出版，为完善汉字大字符集编码提供了素材，是大规模汉字自动识别的基础资源[24]。2010 年华东师范大学的史小松等人，基于语料库对甲骨文字特征实现半自动构件拆分，甲骨文字形方面通过支持向量机技术进行的相似性分析系统，为甲骨文字形考释工作提供了帮助[25]。2015 年西北大学刘磊通过文本聚类对秦汉瓦当小篆文字进行单字提取，基于尺度不变特征转换（SIFT）和数学形态学结合进行特征提取，采用二叉树结构的多级 SVM 进行小篆文字识别[26]。2016 年顾邵通基于自动识别系统，依靠构造的拓扑，计算在特征库中的拓扑编码与待配准拓扑之间的欧氏距离，实现甲骨字形配准识别的研究[27]。我国将会有越来越多的学者致力于古汉字识别研究，为更好的实现可移动文物的历史价值，为广大考古工作者提供有力的技术支持，实现文物保护。

当今实际生活中应用识别技术的领域更加宽广，最广泛的是对于在人脸图像上识别技术的研究、基因表达数据分析等。由于我国古汉字种类众多，书写风格多样，文字形态多变，数量巨大，研究有一定的技术难度，研究出成果不易。因此需要我们更多的学者不畏艰辛，谨慎探索，为我们能更充分地了解中国的历史文化，过去的人文历史形态，提供强有力的技术支持。

1.3.3 分类器识别概述

分类识别是根据提取出有效的文字信息通过分类中的规则进行文字识别的过程，相同的文字样本在不同的分类器中进行识别的结果是不同的。近些年来，分类方法飞速发展，很多学者进行了大量的研究，出现了很多比较智能且有效的方法，下面就介绍一下比较常见的几种分类方法，分别是：

（1）最近邻（KNN）分类法。在 1968 年，Cover 和 Hart 两位学者对 KNN 这种新的算法进行了论文的出刊，该算法更好的解决了文本分类的问题。KNN 算法[28]主要是当一个样本需要分类时，在特征空间中根据同一类型的个体在特征空间中有聚类特性的原理，就是将多个样本进行相似聚类然后进行样本分类。KNN 算法是简单和容易实现的，尤其是当遇到的类域有交叉或重叠的样本集之间进行分类时，KNN 分类更为合适。现在比较推广的是 $k-$近邻法，为了解决近邻法计算量大的缺点，采用了快速搜索近邻法。而为了更加准确地估计错误率采用了剪辑近邻法。在此基础上，为了使计算所用的时间尽可能的减短，存储信息的要求尽可能的降低，采用压缩最近邻法在一定程度上符合了上述要求。由于使

用近邻法过程中存在着一定的决策风险甚至在不可预测的情况下决策风险很大，因此在算法中使用可做拒绝决策的近邻法，在一定程度上可以避免风险。当结果是为了错误率达到最小，在近邻法中应该考虑距离的度量，从而决定采用最佳距离度量近邻法[29]。

（2）贝叶斯分类法。贝叶斯分类法（Bayes）[30]是非规则算法，利用先前得到的经验信息和样本收集来的数据，最后对整个事件进行后验概率的确定。在贝叶斯分类法中，有最小错误率贝叶斯决策，最小风险贝叶斯决策。

其中朴素贝叶斯分类法由于分类率失误最小，计算速率快，结果精确值高等特点算是一种相对简单的分类方法。朴素贝叶斯分类法的基本原理就是对所有的样本在已知条件下，计算每一项类别的概率，最后待分类的样本就属于其中类别概率是最大的那一类。但是朴素贝叶斯的假设理论在现实生活中的运用存在相对的制约，是以在要求分类的程度比较精准的情境中，不建议使用朴素贝叶斯算法。

（3）神经网络分类法。神经网络是在 50 年代被研究学者正式提出来的，属于人工智能这一学科的比较受热捧的一个研究方向。在 80 年代中期，对神经网络的研究进入了一个高峰阶段，很多的研究人员涉足这个领域进行研究并发表了研究成果[31]。近几年来，更被从事科研的人员运用到更为广阔的领域。神经网络就是依照人脑的活动进而建造的数学模型，是由众多的能进行处理的以不同的联结方式的神经单元构建而成的，属于作用于大范围的具有自适应能力较强的一个非线性系统。

（4）支持向量机（SVM）。支持向量机[32]方法是在 1995 年由 V.N.Vapnik 教授等人提出的解决以数据为基础的非线性建模问题，主要是一种针对计算机学习的新方法。它的基本原理是以统计学习中的 VC 维理论和结合最小的风险结构为基础，在有限的样本或特征空间上，根据样本信息在可分线性下构造最优分类平面。这样就使得存在差别的样本集和分类平面它们之间存在的距离变为最大，实现了支持向量机的最优的泛化能力和较好的鲁棒性。

在线性不可分时构造最优分类平面就转变成了优化二次型的问题，需要加入松弛变量。大部分情况下，我们利用核函数通过非线性映射将低维输入空间的样本模式矢量映射到高维属性空间，这样就将非线性问题变成了线性问题，避免了在高维特征空间求解最大间隔超平面困难的问题。近几年来，它广泛的被使用于模式识别、预测预报、计算智能等科学计数领域，在文字识别领域，也有广泛的应用并加以改进，如孙莹[33]提出通过二代曲波变换得到汉字的频域特征，结合

局部特征与全局特征的多特征融合，通过改进的混合核最小方差支持向量机（LS－SVM）实现了古汉字的图像识别。

1.4 本文研究内容

本文选取《银雀山汉简文字编》和《上海博物馆藏战国楚竹书文字编》中收录的简牍图片作为本文的研究对象，建立小型字库，以图像处理、模式识别以及机器学习等技术为基础，实现简牍图像文字的提取与识别。

论文的主要内容有两大要点，一是对简牍的预处理，利用灰度化、二值化、去噪、反色、细化、归一化等实现简牍文字提取，为后续简牍文字的识别做好准备。二是对文字进行特征提取，进行简牍文字识别的研究。全文的研究内容为：对简牍文字图像样本进行预处理，在预处理的图像去噪方面，提出新的自适应加权中值滤波，为的是更好的解决降噪与保护细节的矛盾，并为后续的文字识别提供有力条件。在特征提取方面，将结构特征、整体广义密度特征和双弹性网格局部密度特征这三种特征进行融合，通过 BP 神经网络实现简牍图像文字识别。

简牍文字图像的识别要完成以下步骤：首先将简牍文字图像样本存储到计算机中，然后对简牍文字图像进行预处理，再对处理过的图像进行特征提取，最终使用 BP 神经网络进行简牍文字识别，这就是整个识别流程。如图 1-1 所示：

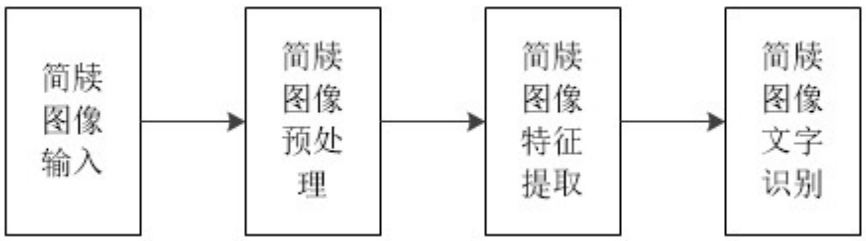


图 1-1 简牍汉字识别流程图

1.5 论文的基本框架

本文章节安排如下：

第一章：绪论。对简牍图像文字提取与识别的背景和意义进行了阐述，以及运用的数字化图像处理技术在古文物这方面的国内外研究现状，文字识别技术的国内外研究现状以及几种分类识别器的基础理论知识。

第二章：简牍文字图像预处理的一些算法的介绍。主要对简牍图像进行预处理以及效果的展示，在图像去噪处理阶段提出了新的自适应加权中值滤波算法，利用分数阶微积分确定图像中的噪声点，根据噪声数量自适应确定滤波窗口大小，运用权函数自适应计算权值，对噪声点采用加权中值滤波。最终达到较好的

去噪效果。

第三章：简牍文字特征提取的研究。在进行特征提取方面对结构特征、整体广义密度特征、双弹性局部密度特征这三种特征实现多特征融合，该方法实现了特征之间相互补充的功能，更保证了特征的鲁棒性以及更好的可区分性。

第四章：基于 BP 神经网络进行简牍汉字识别研究。利用标准 BP 神经网络和改进 BP 神经网络，针对单一特征与多特征融合进行简牍文字识别实验对比，验证多特征融合有较高的识别率。

第五章：总结与展望。对本文的研究工作进行总结，以及对进一步工作的展望。

2 简牍图像预处理

预处理是简牍图像文字进行识别的第一个重要的阶段。因为简牍长年埋藏于地下，出土时会有字迹变色、发霉、虫蛀的现象，在采集图像时，也会有光照，天气等环境的影响，因此扫描成像时会存在对比度差，汉字图像存在断点、污点有相当大的噪声点等问题。对于简牍文字图像进行预处理既去掉了不必要的信息更减少了计算的数据量。

2.1 简牍图像灰度化、二值化

对于真彩色简牍图像，进行简牍图像灰度化。图像灰度化转换是进行图像分割、边缘提取、模式识别、特征提取等处理之前的首要预处理步骤，最终将图像中的色彩信息转换成了亮度信息，保留了图像重要特征并且减少了图像数据的大小，以及后续计算量。由于本文的主要工作是对简牍图像文字进行研究，主要针对的是简牍图像文字的笔划大小、宽度和位置等信息，丰富的颜色信息对本论文的研究来说是冗余的干扰信息。因此对简牍图像进行灰度化处理，以减少数据运算量，节省空间与时间。直接使用彩色图像转化为灰度图像公式：

$$Y=0.299R+0.587G+0.114B \quad \text{（式 2-1）}$$

彩色简牍图像转化为灰度图像如图 2-1 所示：

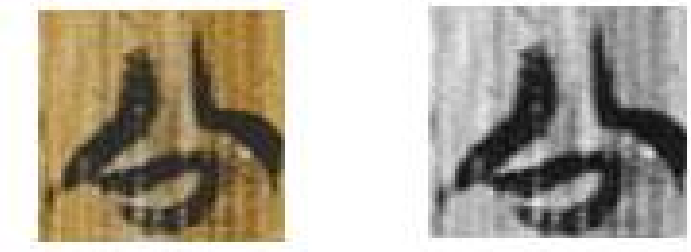


图 2-1 灰度化效果图

图像二值化，就是将原来灰度图像像素值的 256 种，转换为 0 和 1 两种取值的图像。二值化后的简牍图像不仅保留了简牍图像的主要特征，而且大大压缩了信息量，在后续的特征提取中处理速率较快。因此对简牍图像进行二值化处理，将灰度值范围在$[m, n]$的简牍图像设为$f(x,y)$，设定一个阈值$T(m \le T \le n)$，二值化处理公式为：

$$g\left(x,y\right)=\begin{cases}1 & f\left(x,y\right)\ge T\\ 0 & f\left(x,y\right)\le T\end{cases} \quad \text{（式 2-2）}$$

选择阈值的方法有很多种如全局阈值法、局部阈值法、动态阈值法[34]等。动态阈值法称为自适应阈值法，是基于位置坐标、图像像素和图像像素周围的像

素灰度值，自动确定不同的阈值，最常见的是动态阈值邻域平均法。全局阈值法是设定一个阈值T，将图像划分为两个部分，一部分是像素灰度值大于T的，一部分是像素灰度值小于T的。全局阈值法包含有 Ostu 法、平均灰度值法和数学期望法等。最为典型的是 Ostu 法，可以实现背景与目标的分离。局部阈值法是将图像划分为多个小的局部部分，计算每一局部部分的阈值，再根据得到的阈值对图像进行二值化处理，包括 Kamel Zhao 算法、噪声算法和 Bernsen 算法等。

Ostu 算法[35]又称为最大类间方差法，通过阈值将图像分成目标和背景两个部分，设图像具有L级灰度级，其中第i级像素为N_i个，其中i值范围为$0 \sim L-1$，则像素总数为$N=\sum_{i=0}^{L-1} N_i$，第i级的概率为$P_i=\frac{N_i}{N}$，以k作为阈值，将图像分为目标C_0和背景C_1两个区域，C_0区域像素灰度级$0 \sim k$，C_1区域像素灰度级$K+1 \sim L-1$，C_0和C_1的概率分别为：

$$P_{C_0}=\sum_{i=0}^{K} P_i \quad , \quad P_{C_1}=\sum_{i=k+1}^{L-1} P_i \tag{式 2-3}$$

区域C_0与C_1的平均灰度分别为：

$$\mu_{C_0}=\frac{1}{P_{C_0}}\sum_{i=0}^{K} iP_i \quad , \quad \mu_{C_1}=\frac{1}{P_{C_1}}\sum_{i=k+1}^{L-1} iP_i \tag{式 2-4}$$

整个图像的平均灰度为：

$$\overline{\mu}=P_{C_0}\mu_{C_0}+P_{C_1}\mu_{C_1} \tag{式 2-5}$$

总体方差为：

$$\delta^2=P_{C_0}\left(\mu_{C_0}-\overline{\mu}\right)^2+P_{C_1}\left(\mu_{C_1}-\overline{\mu}\right)^2 \tag{式 2-6}$$

本文基于 Ostu 算法对简牍图像二值化处理。效果如图 2-2 所示：

图 2-2 简牍文字图像二值化

2.2 简牍图像去噪

大多数简牍都会有大自然腐蚀的现象或者保存不当的失误以及存在破损的情况。因此获取的简牍图像上的文字一定会存在噪声，甚至处于模糊不清状态了。所以，需要在尽可能的保留文字细节的情况下进行去噪处理，这有利于简牍文字识别的研究。

在图像中的噪声是分散的像素，这些点包含像素少，对汉字信息影响较小，

通常有意义的信息是大量的像素按照一定的规则聚集而成的[36]。图像的去噪处理方法有线性滤波、非线性滤波和自适应滤波三大类。线性滤波对于颗粒噪声图像去噪效果好，但往往会模糊了图像的细节。线性滤波方法主要有平滑去噪方法，包括简单邻域平均法、邻域加权平均法等。非线性滤波在去除噪声的同时，使得图像保持清晰和不失真。非线性滤波方法主要有中值滤波，形态滤波，层叠滤波等。在本文中主要对中值滤波去噪进行了研究。

2.2.1 中值滤波

中值滤波器是在 1971 年提出的根据次序统计的非线性数字平滑滤波技术。首先被运用用到了一维信号时间序列分析中，之后这个技术被运用到了二维图像处理中[37]。传统的中值滤波就是在排序统计理论的指导下取数字图像中或者数字序列中的一点的值，对于在这个点的周围邻近区域中每个点值的中值用这一个点替代，因此在这个邻域内的像素值都是相似接近真实值的，从而达到了消除离散孤立的噪声点，而且解决了图像细节模糊的问题。

（1）一维中值滤波

设一个一维序列 $f_1, f_2, \ldots, f_n$，将窗口的长度设置为 n(n为奇数)，中值滤波后，在输入序列中顺序取出 n 个元素，$f_{i-v}, \ldots, f_{i-1}, f_i, f_{i+1}, \ldots, f_{i+v}$，其中 i 为中心点的位置，则 $v = \frac{n-1}{2}$，n 个元素大小排列，滤波输出中心点的值为：

$$y_i = Med\left\{f_{i-v}, \ldots, f_{i-1}, f_i, f_{i+1}, \ldots, f_{i+v}\right\} \quad (式\ 2\text{-}7)$$

其中，$Med\left\{\ldots\right\}$ 表示取序列中值，$i \in Z, x = n - 1/2$。

（2）二维中值滤波

设 $\left\{x_{ij}, (i,j) \in I^2\right\}$ 为灰度值，滤波窗口大小为 B，y_{ij} 表示在窗口大小为B中点 x_{ij} 的中值，则：

$$y_{ij} = Med\left\{x_{ij}\right\} = Med\left\{x_{(i+r),(j+s)}, (r,s) \in B, (i,j) \in I^2\right\} \quad (式\ 2\text{-}8)$$

2.2.2 新的自适应加权中值滤波引言

传统的中值滤波是在窗口的形状和大小预先设定的情况下进行图像去噪处理的，在滤波的同时会丢失图像细节，因此研究学者提出了许多改进的中值滤波算法用于解决降噪和保留细节的矛盾。文献[38]将加权中值滤波和定向滤波相结合，根据移动窗口方差和基方差自适应调整中心像素权值，从而提出自适应定向加权中值滤波算法，在保护细节方面优于传统中值滤波算法。文献[39]利用人眼系数视觉特性的敏感度系数确定噪声点，根据窗口内噪声点的个数自适应调整滤波窗口大小，依据相似度大小分组并给予每组像素点相应的权重，从而提出基于

相似度函数的自适应加权中值滤波。文献[40]为了确保图像中所有像素点都进行了噪声检测和滤波，运用了复制边界扩展边缘的方法，并依据噪声检测因子进行噪声进行检测，最终对含噪的图像采用自适应窗口的中心加权算法，从而提出改进的自适应加权中值滤波算法。文献[41]将分数阶积分理论用于图像处理，利用分数阶微分梯度进行噪声检测，有较高的噪声点检测准确率。

本文提出一种新的自适应加权中值滤波，该算法利用分数阶微分对噪声进行检测，根据3×3滑动窗口中噪声点的个数自适应调整滤波窗口大小，利用权函数采用加权中值滤波进行图像去噪处理，较好的解决了降噪与细节之间的矛盾。

2.2.3 新的自适应加权中值滤波算法步骤

确定噪声点是滤波算法中最关键的一步，因为关系到图像像素点的正确分类，传统的自适应加权中值算法中，是基于人眼视觉特性的噪声敏感系数来判断是否为噪声点，而本文采用分数阶微积分对图像中的噪声进行检测。新的自适应加权中值滤波算法的步骤：

步骤 1：检测图像中属于噪声的像素点。设微分阶数v=0.5，对噪声图像和 28 个方向的分数阶微分梯度模板进行卷积运算，得到多方向梯度图。根据多方向梯度图中边缘梯度无跳变，噪声梯度变化明显为依据，消除图像边缘获得噪声点位置。

步骤 2：自适应确定滤波窗口的大小。在3×3滑动窗口中，统计图像中属于噪声点的数量，根据上述获取的噪声点位置，当中央像素被确定为噪声时，统计公式为：

$$Num\left(w_{ij}\right)=\sum_{k=-1}^{1}\sum_{r=-1}^{1}N\left(i+k,j+r\right) \tag{式 2-9}$$

设自适应滤波窗口大小为F_{ij}，公式如下：

$$F_{ij}=\begin{cases}3\times3 & Num\left(w_{ij}\right)\in\left\{1,2,3\right\}\\ 5\times5 & Num\left(w_{ij}\right)\in\left\{4,5,6\right\}\\ 7\times7 & Num\left(w_{ij}\right)\in\left\{7,8,9\right\}\end{cases} \tag{式 2-10}$$

步骤 3：进行噪声像素过滤。区别于传统中值滤波，对噪声进行自适应选择权值进行中值滤波，而权值的大小取决于滤波窗口中像素灰度值与中心像素的差值，差值越小权值越大。并且针对在滤波的滑动窗口中每个像素都经过了加权滤波，非噪声同时也进行了滤波计算，导致过滤的效果并不理想，尤其处于高密度的噪声情况下。由此提出更有效的方法寻求加权系数。它消除了噪声过滤之前避免噪声滤波计算价值的负面影响和得到最好的过滤结果。就是利用权函数计算加

权系数。

$$y\left(x\right) = \frac{1}{e^{|x|}} \quad \text{（式 2-11）}$$

其中 x 表示滤波窗口中像素灰度值与中心像素的差值。因此在噪声点为$\left(m,n\right)$，滤波窗口大小为Fw_{mn}时计算加权系数的过程如下：

$$\mathrm{Med}\left(Fw_{mn}\right) = Mid\left\{f\left(m+s,n+t\right)\right\} \quad \text{（式 2-12）}$$

$$sum = \sum_{s=-L}^{L}\sum_{t=-L}^{L}\frac{1}{e^{|Med(Fw_{mn})-f(m+s,n+t)|}} \quad \text{（式 2-13）}$$

$$Wet\left(m+s,n+t\right) = \frac{1}{e^{|Med(Fw_{mn})-f(m+s,n+t)\cdot sum|}} \quad \text{（式 2-14）}$$

步骤 4：加权中值滤波在滤波窗口中的中心像素，噪声点过滤后对应的灰度值为：

$$\mathrm{g}\left(m,n\right) = \sum_{s=-L}^{L}\sum_{t=-L}^{L}f\left(m+s,n+t\right)\cdot Wet(M+s,n+t) \quad \text{（式 2-15）}$$

最后得到滤波器的输出结果，用来替换噪声点的灰度值规则如下：

（1） 如果$Min(P_{M\cdot N}) < g\left(m,n\right) < Max\left(P_{M\cdot N}\right)$则$f\left(m,n\right) = g\left(m,n\right)$；

（2） 如果$n = 1$那么$f\left(m,n\right) = f\left(m,1\right)$否则$f\left(m,n\right) = f\left(m,n-1\right)$；

在规则（2）中，为了确保输出值$g\left(m,n\right)$在滤波器加权时不能产生新的噪声点。当$f\left(m,n\right)$=$\mathrm{Min}\left(P_{M\cdot N}\right)$或者$f\left(m,n\right)$=$\mathrm{Max}\left(P_{M\cdot N}\right)$时，则$f\left(m,n\right)$的值需要替代$f\left(m,n-1\right)$的值，而一行最左边$f\left(m,1\right)$则用上一行同列的$f\left(m-1,1\right)$值代替。

2.2.4 算法分析

采用中值滤波对简牍图像进行去噪处理。传统的中值滤波一般有固定的窗口大小，采用统一的7 × 7 窗口，会出现如图 2-3 所示的结果：

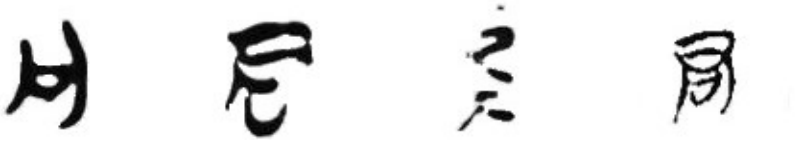


图 2-3 传统中值滤波效果图

从图 2-3 中我们可以观测到在统一的窗口下，对简牍图像进行中值滤波，一部分文字能够达到很好的去噪效果，但是更多地简牍文字在去噪过程中缺失了大量的文字笔划细节，这些文字笔画细节的缺失会影响特征提取，导致识别错误率的提高。

本文采用新的自适应加权中值滤波的方法，该方法既能够消除了噪声，并且较好的保留了图像的细节，利用此方法进行实验如图 2-4 所示：

图 2-4 新的自适应加权中值滤波效果图

从图2-4中我们可以看到在简牍图像中应用新的自适应加权中值滤波算法去噪，很好的保留了文字的细节，这有利于后续的文字识别。为了更好地体现这种方法的优越性，我们又采取文献[39]中的自适应加权中值滤波算法提取文字效果的简牍图像与新的算法进行实验对比，如图 2-5 所示：

（a）AWFM 算法效果　　（b）新的算法效果

图 2-5 效果对比图

从图 2-5 中可以看出采用这两种中值滤波，简牍文字去噪都达到了很好的效果，但是从小的细节中，新的自适应加权中值滤波算法优于文献[39]中的自适应加权中值滤波（AWFM），最明显的就是第一个字在应用传统的中值滤波中缺少了一撇，第二个字在应用传统的自适应加权中值滤波中下面的大字不如使用本文方法去噪后细节保留完整。为了客观评价图像去噪的效果，采用峰值信噪比（PSNP）和平均绝对误差（MSE）作为去噪的客观评价指标，如表 2-1 所示：

表 2-1 性能指标表

滤波方法	性能指标	椒盐噪声大小					
		2%	5%	10%	20%	30%	40%
3X3MF	PSNR	82.13	80.48	78.96	77.18	66.39	75.25
	MAE	3.47	7.35	13.85	25.87	36.98	45.68
5x5MF	PSNR	81.41	80.06	78.68	76.94	75.82	75.02
	MAE	4.21	8.19	14.82	27.76	39.50	50.09
AWFM	PSNR	84.18	82.13	80.54	78.86	74.74	73.88
	MAE	2.91	5.82	10.22	24.68	38.39	46.88
本文算法	PSNR	89.46	89.07	88.67	87.46	85.87	82.98
	MAE	0.88	0.92	1.36	1.94	2.87	3.42

从表 2-1 中可以看出本文算法的性能指标优于其它中值滤波算法，甚至在高密度噪声情况下，也具有较好的去噪效果，因此本文的算法优于其它中值滤波算法，可以更好的保留文字的细节。由于大多数情况下，细节决定文字识别率的高

低。因此在本文中采用新的自适应加权中值滤波算法，对简牍文字图像进行去噪。

2.3 细化以及反色

2.3.1 简牍图像细化

细化算法，近几年来在数字图像处理和模式识别等学科中得到了越来越广泛的应用，成为了图像预处理必不可少的一部分。简牍文字属于手写体汉字，汉字的笔划粗细长短都是不同的，用一种笔划的所有像素点作为特征信息，信息量大，会导致汉字识别率降低[42]。因此利用细化算法可以去除不影响汉字连通性和结构特征的像素点，不仅最大程度上保存了汉字的特征，并且使以后的计算量和工作量得到了减少，提高了效率。常用的细化算法有：

（1） Hilditch 细化算法

表 2-2 八邻域模板

P9	P2	P3
P8	P1	P4
P7	P6	P5

Hilditch 细化算法是 1969 年 Hilditch 提出的，适用于二值图像的串行细化算法[43,44]。算法过程需要在图像上遍历每一个像素，以中心点像素的八邻域作为移动模板如表 2-2 所示。在上下左右每一个方向上搜寻目标像素点在满足条件的状况下进行删除处理，最终在整个图像上没有可以删除的像素点就可以达到文字细化的效果。设图像中任意一点P1的 8 邻域如表 2-2 所示：当目标像素值为 1，背景像素为 0，目标像素满足以下五个条件就可以删除：

a) P2+P4+P6+P8 ≤ 3；

b) N_C=1；

c) Pk $(2 \le k < 9)$中至少有一个目标像素为 1；

d) P2=1 或N_{C2}=1，N_{C2}为假设P2=0时，P1的联结数；

e) P8=1 或N_{C8}=1，N_{C8}为P8=0时，P1的联结数。

（2）Zhang 并行快速细化算法

Zhang 并行快速细化算法[45,46]是在 1984 年 Zhang 提出的，设二值图像区域为3×3，以P1为中心，如表 2-3 所示：

表 2-3 3X3 区域

P9	P2	P3
P8	P1	P4
P4	P6	P5

P1=1，表示图片中的白点，P1可以被删除的条件是以下六个即：

a) $2 \le N(\mathrm{P1}) \le 6$，其中 $N(x)$ 表示点 x 邻域中黑点的个数，$N(\mathrm{P1})$=P2+P3+P4+P5+P6+P7+P8+P9；这是为了排除孤立点和内部点；

b) $Y(\mathrm{P1})$=1，$Y(x)$为在点 x 邻域中按照P2 ~ P9的顺序，由黑变白的次数，即由0变为1的次数；

c) P2 * P8 * P4=0 or $Y(\mathrm{P2})$!=1，继续向下删除；

d) P2 * P8 * P6=0 or $Y(\mathrm{P8})$!=1，继续向右删除；

e) P6 * P4 * P2=0 or $Y(\mathrm{P4})$!=1，继续向左删除；

f) P6 * P8 * P4=0 or $Y(\mathrm{P6})$!=1，继续向上删除。

重复以上步骤最终是遍历了所有点都处于不能被删除的状态就可以了。目的就是文字图像中的汉字有了很好的细化效果。

本文中运用 Hilditch 细化，细化效果如图 2-6 所示：

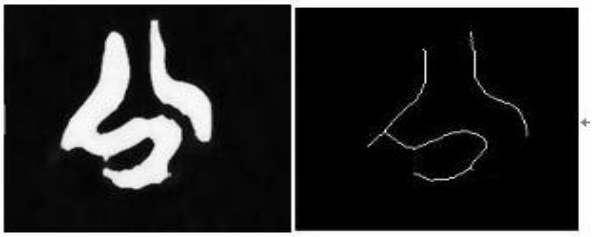


图 2-6 细化效果图

2.3.2 简牍图像反色处理

简牍图像经过二值化处理，一般可以得到两种形式，一种为黑底白字，表现形式是汉字像素值为 0，背景图像像素值为 1，而另一种为白底黑字，表现形式是汉字像素值为 1，背景图像像素值为 0，经过细化后这两种形式会有不同的效果，如图 2-7 所示：

（a）白底黑字反色效果（b）黑底白字反色效果

图 2-7 不同背景细化后的效果图

从图2-7中可以清楚的分析出使用黑底白字进行细化，文字细化效果比较好，

没有产生文字变形扭曲的现象，所以在简牍图像预处理过程中，对于经过处理的白底黑字的图像需要反色这一步骤，将其变为黑底白字。这样不仅得到了很好的细化效果，并且对于后续的特征提取和文字识别也是很有利的。

2.4 归一化

归一化处理在预处理中占有很重要的地位，对后面的有效性的，全面的特征提取产生了直接的，不容忽视的影响。由于简牍文字属于手写汉字，手写汉字的书写位置，尺寸大小皆不相同，因此需要进行归一化处理。常用的归一化算法如下：

（1）重心对准法。重心对准法[47]是把在简牍图像中处于中心的像素映射到标准的几何点阵处于中心的位置，设文字图像四个边界坐标分别为：$x_{\min}$，$x_{\max}$，$y_{\min}$，$y_{\max}$，文字图像的重心坐标为$\left(x_a, y_a\right)$ 归一化映射公式为：

$$\begin{aligned} x &= \alpha\left(x - x_a\right) + 64 \\ y &= \alpha\left(y - y_a\right) + 64 \end{aligned} \tag{式 2-16}$$

其中尺度因子为$\alpha = 140 \Big/ \sqrt{\left(x_{\max} - x_{\min}\right)\left(y_{\max} - y_{\min}\right)}$。

（2）基于点密度均衡法。点密度均衡方法是将像素点投影到x轴和y轴，将原本不均衡的投影密度均衡化。像素点投影到x轴和y轴的点密度：

$$\begin{cases} X_i = \sum_{j=1}^{J} f\left(i, j\right) + \alpha_X \\ Y_i = \sum_{i=1}^{I} f\left(i, j\right) + \alpha_Y \end{cases} \tag{式 2-17}$$

其中，α_X和α_Y分别表示两个常数，两个密度的值越大则代表能够变换的强度很大，变换的获得的新的位置坐标$\left(X_m, Y_n\right)$ 是：

$$\begin{cases} \mathrm{X}_m = \sum_{i=1}^{I} X_i \times \dfrac{M}{\sum_{i=1}^{I} X_i} \\ Y_n = \sum_{j=1}^{J} Y_j \times \dfrac{N}{\sum_{j=1}^{J} y_j} \end{cases} \tag{式 2-18}$$

（3）插值变换法。插值变换归一方法是基于在归一化图像中的插值像素图像像素在对应的原始图像中像素点的位置，从而得到归一化后的新的图像中像素点的灰度值。设原始图像$f\left(x, y\right)$，归一化后的新的图像为$g\left(x, y\right)$，$g\left(x, y\right)$中任意一像素点$\left(\mathrm{a}_0, \mathrm{b}_0\right)$与$f\left(x, y\right)$中的点$\left(x_0, y_0\right)$的映射关系为：

$$\begin{cases} x_0 = \dfrac{width'}{width} a \\ y_0 = \dfrac{height'}{height} b \end{cases} \quad (式\ 2\text{-}19)$$

在上述公式中，$width$、$height$代表原始图像$f(x,y)$的宽度、高度。$width'$、$height'$分别为归一化后图像$g(x,y)$的宽度、高度。当映射的值为整数时，则不需要进行内插变换，即$g(a,b)=f(x_0,y_0)$，若不是整数，则需要进行内插变换。

内插变换常用的三种方法分别是最近邻插值法，双线性插值法，双三次插值法。最近邻插值法依照四舍五入的原则，需要在原始图像中能够找到的真实点属于最接近的虚拟的点，采用这种方法简单容易实现但是最后的效果相对较差。双线性插值法是基于在2×2邻域中像素通过加权平均法计算，根据简牍图像像素周围的4像素的灰度值在水平和垂直两个方向上进行插值，最后出来的结果相对于最近邻插值法体现的视觉效果比较好，双三次插值法是选用4×4领域进行插值，插值核需要进行三次函数的运算，插值结果相比于前两个算法要好，但是由于时间复杂度的增加计算量很大，实现不容易。

（4）基于线间隔密度法。基于线间隔密度法[48]主要是根据文字图像上任意相邻的两条的笔划之间的间隔的长度或者是它们之间存在间隔大体分布的密度计算得到的函数，使用这两种算法能够反映简牍文字结构的特点信息。因此设$m(x,y)$，$n(x,y)$为简牍文字图像上任意一点在水平和垂直方向上的线间隔，定义两个函数：

$$F_M(x,y)=\frac{1}{m(x,y)} \quad (式\ 2\text{-}20)$$

$$F_N(x,y)=\frac{1}{n(x,y)} \quad (式\ 2\text{-}21)$$

$F_M(x,y)$，$F_n(x,y)$的值用很小的常量定义，则特征密度均衡函数为：

$$M(x)=\sum_{y=1}^{Y} F_M(x,y) \quad (式\ 2\text{-}22)$$

$$M(y)=\sum_{x=1}^{X} F_N(x,y) \quad (式\ 2\text{-}23)$$

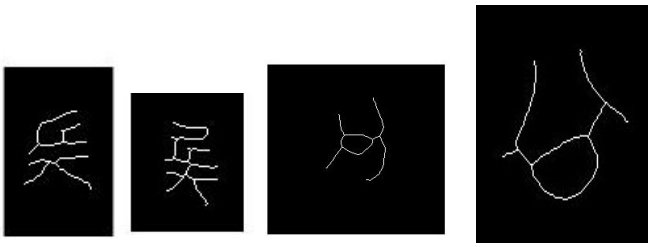
对应新的坐标位置(a,b)为：

$$a=\phi(x)=A\sum_{x=1}^{x} M(i) + x_0 \quad (式\ 2\text{-}24)$$

$$b=\phi(y)=B\sum_{y=1}^{y} N(j) + x_0 \quad (式\ 2\text{-}25)$$

其中，A，B是$\phi(x)$和$\phi(y)$的常系数，(x_0,y_0)是归一化图像的起始位置。利用插值变换算法对简牍图像进行归一化处理，归一化尺寸大小为64×64，最终

得到的预处理效果与不进行归一化的预处理效果进行对比如图 2-8 所示：



（a）不进行归一化

（b）归一化

图 2-8 效果对比图

从图 2-8 可以看出，归一化可以使简牍文字的宽和高等比例的缩放，并且简牍文字图像被归一化到相同的大小，相同的位置，并不会改变图像的重要特征，反而最大限度的突出了提取特征。

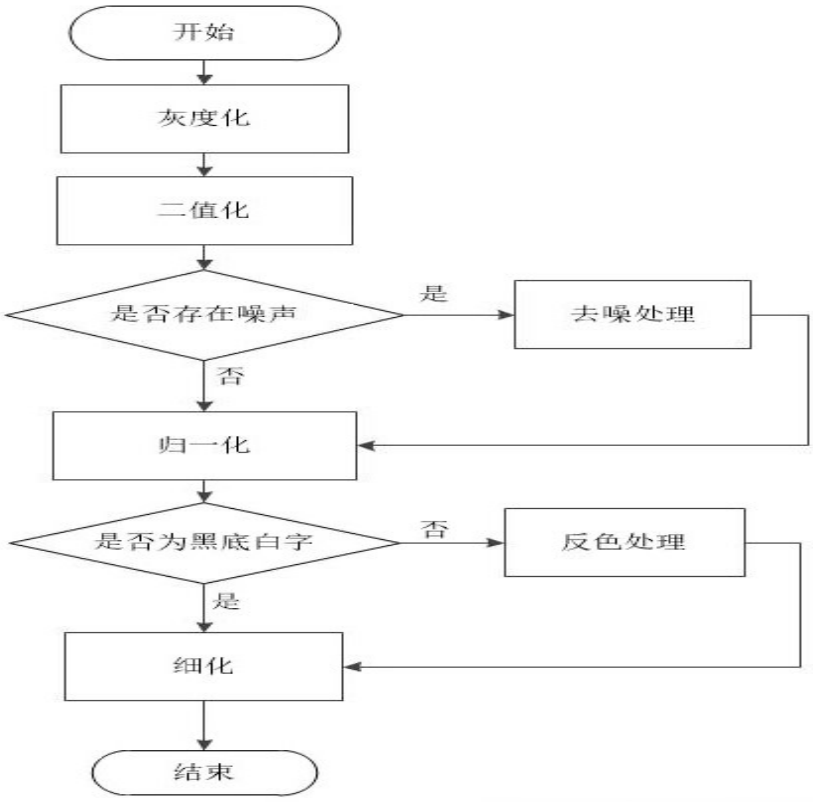


图 2-9 预处理流程图

本文的预处理流程如图 2-9 所示，在收集样本时，样本具有多样性，收集的

样本在质量也会有高低之分，通过预处理系统，就可以实现简牍图像一系列的预处理算法。

2.5 小结

本章主要介绍了对简牍文字图像进行预处理，并对所有使用的预处理算法得到的效果进行了展示，预处理的目的是最大程度上保留较多的有效的文字信息，这样有利于后面的特征提取。在图像去噪方面，根据简牍文字的特点，运用新的自适应加权中值滤波算法，取得了良好的去噪效果，并最大程度上保留了文字细节。整个简牍文字图像预处理的效果决定着文字特征提取的有效性的高低，更是影响着最终识别效果的优劣，因此预处理这一过程不仅为后续工作做了准备，更是进行简牍文字识别的基础，是必不可少的重要部分。

3 基于结构与统计的多特征融合方法研究

人们能够识别不同的事物，因为它们各有特点。为了使计算机能够识别事物，只要能够提供足够的功能和正确可行的识别方法，计算机就能够完成识别[49]。因此特征提取是图形图像领域中一个至关重要的课题，它是指利用先进的计算机技术提取出图像中各个像素点的信息，然后依照一定的判定标准来确认这些信息是属于哪一类图像的特征，对它们进行分类和归属。特征提取算法可以消除图像中一些冗余的干扰因素，并且所选取的特征具有稳定性和代表性，它的最终结果是按照一定的规则将图像上的像素分为不同的类别。因此，为了对简牍图像进行识别，特征提取是一个必不可少的步骤。

3.1 简牍图像特征提取的目的

通过对简牍图像进行预处理，在这个过程中不仅消除了在图像中存在的冗余信息，更是获得了规范化，标准化的简牍图像点阵。然而对于我们来说这个简牍图像的点阵包含的数据量过大，例如在本文预处理流程中，最终将图像大小归一化为64×64，那么需要处理的数据就是 4096 个，这种属于大类别数据集的处理，需要耗费相当大的时间和精力，并且结果不一定准确，因为在这个简牍图像点阵中包含了很多的无关的信息，例如属于背景的像素点，这些信息是不能进行识别的，而且在一定的程度上还会干扰识别。在不能进行直接识别的情况下，越来越多的研究工作者在文字特征提取方面做了大量的研究。

因此，在本文整个简牍文字图像识别过程中特征提取是无可替代的重要的一部分。以数学理论为基础，特征提取就是把图像点阵中冗余的信息进行剔除，提取出有用的图像信息，进而把不具有类别的汉字变成能够可分的汉字，从而使用模式识别与分类技术进行文字识别。特征提取的目的是找出汉字的性质，反映汉字的本质，尽量缩小同一汉字的样本。提取的特征应该有四个特点[50]：

(1)区别性：汉字样本分为不同的类别，那么不同类别的汉字特征值一定具有明显的差异性。

(2)可靠性：属于同种类别的汉字，它们所具有的特征值也是相对相近的。

(3)独立性：不同类别的汉字的特征是没有相关性的。

(4)数量少：汉字识别系统中提取特征的个数的越大，那么整个系统的复杂度也会随着增大。

3.2 特征提取的常用方法

3.2.1 统计特征提取方法

为了提取简牍图像有效的文字信息进行了非线性处理是属于统计特征的提取过程，已经预处理过的图像才能进行。由于提取特征区域的不同进而分为局部和全局两大类[51,52]。局部统计特征就是在图像被划分的基础上，对每一个图像部分的特征信息进行统计，最终提取出的特征就是总的统计的结果。全局统计特征就是针对整个图像进行特征提取。统计特征提取在提取过程中具有较高的稳定性，抗干扰能力强。被广泛使用在汉字识别领域的统计特征提取具有运算速度快，便利性强的优点。同时存在一定的缺点，对于文字的结构信息的特征不能有效的提取，从而造成文字重要特征信息的丢失，影响文字识别的效率。

3.2.2 结构特征提取方法

为了能对代表简牍文字特征的结构特征进行提取，例如提取文字的轮廓等有效信息的过程算法就是结构特征提取。在所有的汉字中都有端点、拐点和交叉点以及“横”、“竖”、“撇”、“捺”，部件等都属于结构元素。因此使用结构特征提取不仅能解决手写汉字笔划不规则的问题，更能有准确的分类能力。

3.3 基于结构与统计的多特征融合

汉字特征提取的越全面，被识别的正确率才越高，单一的特征比较片面不能很好的诠释整个汉字的形体，因此本章对简牍汉字进行提取了整体的广义密度、部件结构、双弹性网格的局部密度这三种特征作为识别的依据。

3.3.1 部件结构特征

部件结构特征有左右、上下、内外、独体、左中右、上中下等等。由于部件结构主要侧重全局结构特征，相对来说比较稳定，提取比较容易。并且大多数汉字相对位置为左右、上下、内外、独体。因此本章基于这四种相对位置来进行特征提取。相对位置的判定，是通过对单字简牍图像进行水平方向，垂直方向的直方图投影为依据的。以下是四种结构的判定：

左右结构特征：如果垂直方向有笔划梯度像素数存在中间为零，在水平方向的投影直方图没有笔划像素数梯度存在中间为零的情况，则为左右结构。

上下结构特征：如果在水平方向的投影直方图存在有笔划梯度像素数存在中间为零的，而在垂直方向的投影没有笔划像素数梯度存在中间为零的情况，则为上下结构。

独体结构特征：垂直、水平两个方向的笔划梯度像素数中间不存在是零的情况但须其中一方处于连通一方相反即为独体字。

内外结构特征：垂直、水平两个方向的笔划梯度像素数中间不存在是零的情况且双向都处于连通状态即为内外结构。

3.3.2 整体广义密度特征

整体密度是图像中的前景点像素在整个图像总像素中的比例。在提取过程中算法速度快，抗干扰能力强。因此在简牍文字特征提取时采用整体广义密度，在简牍图像一个样本中它的整体密度特征是：

$$p_i=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}f\left(x,y\right)}{n^2}\left(i=1,\ldots,M\right) \tag{式 3-1}$$

均值和方差为：

$$\overline{p}_{NM}=\frac{1}{M}\sum_{i=1}^{M}p_i \tag{式 3-2}$$

$$E\left(p_{NM}\right)=\left(\overline{p}_{NM}-p_1\right)^2+\wedge+\left(\overline{p}_{NM}-p_M\right)^2 \tag{式 3-3}$$

由上面两个式子可以得到整体像素的最大变化区间为：

$$p_N=\left[\overline{p}_{NM}-\mathrm{Max}\left(E\left(p_{NM}\right)\right),\overline{p}_{NM}+\mathrm{Max}\left(E\left(p_{NM}\right)\right)\right] \tag{式 3-4}$$

p_N 代表整体广义密度特征，整体广义密度特性是它可以更好地的显示笔划长短不一，倾斜角度各不相同的汉字的整体像素密度的最大波动范围。

3.3.3 双弹性网格特征

划分双弹性网格，假设 $M\times N$ 为简牍图像的大小，水平、垂直两个方向上的网格线是 N_1 和 N_2，简牍图像中的像素值是 $f\left(i,j\right)$：

$$f\left(i,j\right)=\begin{cases}0 & \text{白像素}\\ 1 & \text{黑像素}\end{cases} \tag{式 3-5}$$

纵横弹性网格：

$$\sum_{y=I_i}^{I_{i+1}}\sum_{x=1}^{N_2}f\left(x,y\right)=\sum_{y=I_k}^{I_{k+1}}\sum_{x=1}^{N_2}f\left(x,y\right)\quad\forall i,k=1,2,\ldots,N_1-1 \tag{式 3-6}$$

$$\sum_{y=1}^{N_1}\sum_{x=I_i}^{I_{i+1}}f\left(x,y\right)=\sum_{y=1}^{N_1}\sum_{x=I_k}^{I_{k+1}}f\left(x,y\right)\quad\forall i,k=1,2,\ldots,N_2-1 \tag{式 3-7}$$

对角弹性网格：设置图像大小是 $M\times N$，图像在45°方向上投影，总的投影线M+N-1条，$f_1\left(i\right)$ 代表其中第 i 条投影线的总像素数是：

$$f_1\left(i\right)=\begin{cases}\sum_{k=0}^{i-1}f\left(i-k,1+k\right), & 1\le i\le M\\ \sum_{k=0}^{M+N-1-i}f\left(M-k,i-M+1+k\right), & M+1\le i\le M+N-1\end{cases} \tag{式 3-8}$$

当图像在135°方向投影时，一共有M+N-1条投影线，$f_2\left(j\right)$ 表示的是其中第 i 条

投影线的总像素数是：

$$f_2(j)=\begin{cases}\sum_{k=0}^{j-1} f(M+1-j+k,1+k), & 1\le j\le M\\ \sum_{k=0}^{M+N-1-j} f(1+k,j-M+1+k), & M+1\le j\le M+N-1\end{cases} \quad (式\ 3\text{-}9)$$

设N_3和N_4分别是45°方向和135°方向上的网格线，对角弹性网格是：

$$\sum_{x=I_i}^{I_{i+1}} f_1(x)=\sum_{x=I_k}^{I_{k+1}} f_1(x),\ \forall i,k=1,2,\ldots,N_3-1 \quad (式\ 3\text{-}10)$$

$$\sum_{y=I_j}^{I_{j+1}} f_2(y)=\sum_{y=I_k}^{I_{k+1}} f_2(y),\ \forall i,k=1,2,\ldots,N_4-1 \quad (式\ 3\text{-}11)$$

3.3.4 局部点密度特征

局部点密度特征属于局部统计特征，假设某个网格内的局部点密度值定义如下：

$$b=\frac{\sum_{u=x}^{x+1}\sum_{v=y}^{y+1} f(u,v)}{(u_{x+1}-u_x)(V_{y+1}-V_y)} \quad (式\ 3\text{-}12)$$

式中$x\in f_u(0,M)$，$y\in f_v(0,N)$，局部点密度矩阵定义如下：

$$B=\begin{bmatrix} b_{00} & b_{01} & \ldots & b_{0n}\\ b_{10} & b_{11} & \ldots & b_{1n}\\ \ldots & \ldots & \ldots & \ldots\\ b_{m0} & b_{m1} & \ldots & b_{mn}\end{bmatrix} \quad (式\ 3\text{-}13)$$

3.3.5 多特征融合

部件结构特征将简牍文字图像分为了四大类，分别为上下结构，左右结构，内外结构和独体结构，根据对整体广义密度的研究将密度分为：简单，中等和复杂，分类的系数为α_1，α_2。因此基于部件特征和整体广义密度特征的串行融合特征公式为：

$$\alpha_{ij}=\begin{bmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\\ \alpha_{31} & \alpha_{32}\\ \alpha_{41} & \alpha_{42}\end{bmatrix} \quad (式\ 3\text{-}14)$$

之后将64 × 64的简牍文字图像，在纵横弹性8 × 8网格下，分别从左右，上下，左下以及右下四个方向，每个方向7条扫描线，对图像进行扫描，统计每个方向扫描图像的像素值之和，得到特征矩阵$f_{x1},f_{x2},f_{x3},f_{x4}$，计算每个网格的局部特征密度，得到局部密度矩阵$B_i$。同样在对角弹性8 × 8网格下，从四个方向

扫描图像，可得特征矩阵 $f_{y1}, f_{y2}, f_{y3}, f_{y4}$ 和局部密度特征矩阵为 B_j。将局部密度特征与纵横弹性网格进行融合，融合方法为：局部点密度矩阵 B_i 分别与特征矩阵相乘，计算公式为：

$$P_1=\lambda_1 B_i \bullet f_{x1} + \lambda_2 B_i \bullet f_{x2} + \lambda_3 B_i \bullet f_{x3} + \lambda_4 B_i \bullet f_{x4}$$

$$=\lambda_1 \begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0n} \\ b_{10} & b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m0} & b_{m0} & \cdots & b_{mn} \end{bmatrix} \bullet \begin{bmatrix} k_{00} \\ k_{01} \\ \cdots \\ k_{0m} \end{bmatrix} + \lambda_2 \begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0n} \\ b_{10} & b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m0} & b_{m1} & \cdots & b_{mn} \end{bmatrix} \bullet \begin{bmatrix} k_{10} \\ k_{11} \\ \cdots \\ k_{1m} \end{bmatrix} \quad \text{（式 3-15）}$$

$$+\lambda_3 \begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0n} \\ b_{10} & b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m0} & b_{m1} & \cdots & b_{mn} \end{bmatrix} \bullet \begin{bmatrix} k_{20} \\ k_{21} \\ \cdots \\ k_{2m} \end{bmatrix} + \lambda_4 \begin{bmatrix} b_{00} & b_{01} & \cdots & b_{0n} \\ b_{10} & b_{11} & \cdots & b_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m0} & b_{m1} & \cdots & b_{mn} \end{bmatrix} \bullet \begin{bmatrix} k_{30} \\ k_{31} \\ \cdots \\ k_{3m} \end{bmatrix}$$

上式中，不同的笔画在简牍文字样本所占的比重就是权重 $\lambda_{i(i=1\ldots4)}$ 的值，得到 8×1 维的新的特征向量 P_1。同样可将局部密度特征与纵横弹性网格进行并行融合得到特征向量 P_2，最终得到的特征向量就是将这些特征串行的融合在一起。

3.4 小结

在对简牍文字提取的过程中，简牍文字属于手写的古汉字有多样性的特点，对其特征提取时只用单一的方法可能比较片面化，不能全面的可靠的全方位的诠释它具有的有效信息，导致识别率的降低。因此进行多特征的融合方法实现各个特征之间的优势互补，进而建立最全面的简牍文字特征库，从而能更好地进行识别研究。

4 基于 BP 神经网络简牍文字识别

4.1 BP 神经网络结构与原理

BP 神经网络[53]属于一种多层前馈神经网络，它的基本原理是将信号利用某一函数通过神经元输出的正向传播以及对误差使用反向传播进行学习并选择出最优权值的过程[54]。由于 BP 神经网络能够有效地分类文字信息，结构简单，速率高，因此考虑使用此方法进行简牍文字的识别研究。BP 神经网络结构如图 4-1 所示：

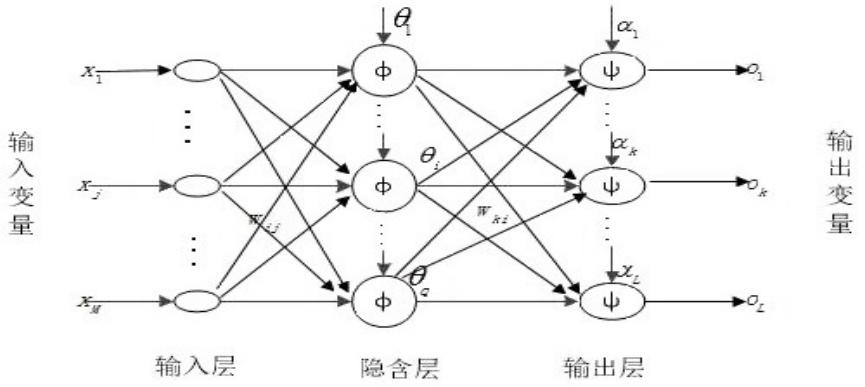


图 4-1 BP 神经网络结构

在图中，输入的矢量为 $x_1,\ldots,x_j,\ldots,x_M$；隐含层输出为 $h_1,\ldots,h_i,\ldots,h_Q$，神经网络实际输出为 $o_1,\ldots,o_k,\ldots,o_L$，训练样本的期望输出为 $d_1,\ldots,d_k,\ldots,d_L$，隐含层第 i 个神经元到输入层第 j 个神经元之间的权值为 w_{ij}，输出层第 k 个神经元到隐含层第 i 个神经元之间的权值为 w_{ki}，输入层的阈值为 θ_i，输出层的阈值为 α_k，隐含层的激励函数为 $\varphi(x)$，输出层的激励函数为 $\psi(x)$。该网络算法主要有以下两个过程：

（1）信号的前向传播过程。已知输入节点的所有变量为 x_j，根据激励函数将权值和阈值进行初始化，初始值为 0~1 或 0 附近的值，则计算隐含层第 i 个神经元的输入 N_i 的公式为：

$$N_i=\sum_{j=1}^{M} w_{ij}x_j+\theta_i \tag{式 4-1}$$

根据隐层激励函数可计算隐含层第 i 个节点的输出 h_i 的公式为：

$$h_i=\varphi(N_i)=\varphi\left(\sum_{j=1}^{M} w_{ij}x_j+\theta_i\right) \tag{式 4-2}$$

计算输出层第 k 个神经元的输入 N_k 的公式为：

$$N_k=\sum_{i=1}^{q} w_{ki}h_i+\alpha_k=\sum_{i=1}^{q} w_{ki}\varphi\left(\sum_{j=1}^{M} w_{ij}x_j+\theta_i\right)+\alpha_k \tag{式 4-3}$$

根据输出层激励函数可计算输出层第 k 个神经元的输出 o_k 公式为：

$$o_k=\psi\left(N_k\right)=\psi\left(\sum_{i=1}^{q}w_{ki}h_i+\alpha_k\right)=\psi\left(\sum_{i=1}^{q}w_{ki}\varphi\left(\sum_{i=1}^{M}w_{ij}+\theta_i\right)+\alpha_k\right) \quad (式\ 4\text{-}4)$$

（2）误差的反向传播。误差的反向传播是由输出层开始的，计算每个层的神经元的输出误差，利用误差梯度下降法来调节各个层的权值和阈值，使最后得到的输出可以最接近期望值。

设训练集中样本总数为 M，样本编号为 m，输出层第 k 个节点的期望输出为 d_k，则计算样本的二次误差函数公式为：

$$S_m=\frac{1}{2}\sum_{k=1}^{L}\left(d_k-o_k\right)^2 \quad (式\ 4\text{-}5)$$

计算整个训练样本集的总误差函数公式为：

$$S=\frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-o_k^M\right)^2 \quad (式\ 4\text{-}6)$$

利用误差梯度下降法对输出层和隐含层的权值和阈值进行修正，修正的输出层权值和阈值公式：

$$\Delta w_{ki}=-\lambda\frac{\partial S}{\partial w_{ki}}=-\lambda\frac{\partial S}{\partial \mathrm{N}_k}\frac{\partial N_k}{\partial w_{ki}}=-\lambda\frac{\partial S}{\partial o_k}\frac{\partial o_k}{\partial N_k}\frac{\partial N_k}{\partial w_{ki}} \quad (式\ 4\text{-}7)$$

$$\Delta\alpha_k=-\lambda\frac{\partial S}{\partial \alpha_k}=-\lambda\frac{\partial S}{\partial \mathrm{N}_k}\frac{\partial N_k}{\partial \alpha_k}=-\lambda\frac{\partial S}{\partial o_k}\frac{\partial o_k}{\partial N_k}\frac{\partial N_k}{\partial \alpha_k} \quad (式\ 4\text{-}8)$$

修正的隐含层权值和阈值公式：

$$\Delta w_{ij}=-\lambda\frac{\partial S}{\partial w_{ij}}=-\lambda\frac{\partial S}{\partial N_i}\frac{\partial N_i}{\partial w_{ij}}=-\lambda\frac{\partial S}{\partial h_i}\frac{\partial h_i}{\partial N_i}\frac{\partial N_i}{\partial w_{ij}} \quad (式\ 4\text{-}9)$$

$$\Delta\theta_i=-\lambda\frac{\partial S}{\partial \theta_i}=-\lambda\frac{\partial S}{\partial N_i}\frac{\partial N_i}{\partial \theta_i}=-\lambda\frac{\partial S}{\partial h_i}\frac{\partial h_i}{\partial N_i}\frac{\partial N_i}{\partial \theta_i} \quad (式\ 4\text{-}10)$$

λ 为学习速率，一般取值0 ~ 1。根据前向传播函数可得到的公式：

$$\frac{\partial S}{\partial o_k}=-\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-0_k^M\right) \quad (式\ 4\text{-}11)$$

$$\frac{\partial N_i}{\partial w_{ki}}=h_i,\ \frac{\partial N_k}{\partial \alpha_k}=1,\ \frac{\partial \mathrm{N}_i}{\partial w_{ij}}=x_j,\ \frac{\partial N_i}{\partial \theta_i}=1 \quad (式\ 4\text{-}12)$$

$$\frac{\partial S}{\partial h_i}=-\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-0_k^M\right)\psi'\left(N_k\right)w_{ki} \quad (式\ 4\text{-}13)$$

$$\frac{\partial h_i}{\partial N_i}=\varphi'\left(N_i\right) \quad (式\ 4\text{-}14)$$

$$\frac{\partial o_k}{\partial N_k}=\psi'\left(N_k\right) \quad (式\ 4\text{-}15)$$

根据上述公式，将偏导函数带入修正的隐含层和输出层权值和阈值公式，最终得到的调整公式为：

$$\Delta w_{ki}=\lambda\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-o_k^M\right)\psi'\left(N_k\right)h_i \quad (式\ 4\text{-}16)$$

$$\Delta\alpha_k=\lambda\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-o_k^M\right)\psi'\left(N_k\right) \quad (式\ 4\text{-}17)$$

$$\Delta w_{ij}=\lambda\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-o_k^M\right)\psi'\left(N_k\right)w_{ki}\varphi'\left(N_i\right)x_j \quad (式\ 4\text{-}18)$$

$$\Delta\theta_i=\lambda\sum_{m=1}^{M}\sum_{k=1}^{L}\left(d_k^M-o_k^M\right)\psi'\left(N_k\right)w_{ki}\varphi'\left(N_i\right) \quad (式\ 4\text{-}19)$$

4.2 BP 神经网络的改进

4.2.1 BP 神经网络的优点

BP 神经网络以数学理论为依据在三层的神经网络中实现了任何的非线性连续函数可以以任意的精度进行逼近，从一个输入到输出映射的功能，因此具备了相对较强的非线性映射能力和解决存在复杂的内部机制问题的能力。

BP 神经网络在训练样本期间可以通过学习自动提取输出、并且输出数据之间的合理规则，将自动的学习内容记忆在神经网络的权值之中，因此它具备高度的自学习、推广、概括和自适应能力。

BP 神经网络不仅在设计模式阶段对分类样本如何正确的进行分类的问题进行考虑，而且关注训练后的那些分类如从没见过的模式或噪声污染模式，因此具备了把学习成果运用到新知识的泛化能力。

BP 神经网络的容错能力体现在即使在局部或者部分的神经元都受到破坏的情况下，仍然进行着正常的工作并且对最后的全局的训练效果产生较小的影响。

BP 神经网络算法首先设计的网络结构对于那些复杂的神经网络是相对简单的、整个过程复杂度小且计算量较少，并且能够以并行分布处理的方式存储和处理神经网络信息，具备有很快的处理速度，而且将传统的工程技术和人工智能技术结合起来用于同时处理定量信息和定性信息，具备数据融合能力，属于多变量系统。

4.2.2 BP 神经网络的缺点

BP 神经网络是当今进行训练时使用较多也比较成熟的神经网络算法的一种，从数学的角度分析实质上就是在对总误差函数的最小值问题进行求解的局部搜索优化算法。但是在 BP 神经网络算法过程中对权值的修正是沿着局部改善的方向逐渐进行的，这样就会导致 BP 神经网络局部极小化，造成神经网络训练的

失败。并且初始不同的权重，局部收敛不同的极小值，得到的训练结果也不相同。

BP 神经网络收敛速度较慢是因为采用非线性规划中的梯度下降法并且目标的优化函数也比较复杂，往往出现“锯齿形现象”，使得算法具有比较低的效率。

BP 神经网络算法是随机初始化网络参数的，神经网络的结构选择一般是凭借大多数的经验来进行的，没有统一的规定，因此当网络结构选择不对时会出现无法学习的过度拟合现象，降低了神经网络的泛化能力。

4.2.3 附加动量项

经典的 BP 神经网络算法仅仅依靠在梯度方向 d 时刻误差函数进行权值的修正，往往忽略了梯度方向 d 时刻以前的误差函数，导致收敛速度越来越慢，同时误差导致了敏感的局部曲面产生了变化，因此导致震荡的产生。面对产生的这些情况，采取了附加动量项的措施，在修改权值时，因为添加了附加动量项，导致了权值是不确定的，是随时变化的，这一措施可以解决收敛速度慢的问题，并且平缓了震荡的产生，增强了整个神经网络的稳定性，消除了敏感的局部曲面变化的影响，尽量使神经网络局部极小值问题得到了解决。附加了动量项的权值修正公式：

$$\Delta w_{ki}\left(d\right)=\eta\delta_k w_{ij}\left(d\right)+\alpha\Delta w_{ki}\left(d-1\right) \quad \text{（式 4-20）}$$

其中α代表动量常数。

4.2.4 自适应调整学习速率

在 BP 神经网络中，我们利用权值修正是为了达到对误差函数进行调整的作用，而为了能够检测到误差函数的变化这里就使用到了学习速率。学习速率能够随误差函数产生变化，就可以消除收敛速度快慢对整个神经网络的影响。在一般的 BP 神经网络中，学习速率常常会被定义为一个常数，因此它不会感知到误差函数的调整变化，不变的学习速率最终会影响整个神经网络的收敛速度。大多时候我们会通过大量的实验对学习速率进行调整，选取适合的学习速率进行实验。大量的实验证明如果学习速率相对较小，会造成整个 BP 神经网络的收敛速度变慢，但是如果学习速率很大，那么会造成整个神经网络敏感的变化，使得整个网络变得不稳定。由于定义的学习速率会造成以上神经网络的缺点，因此为了能够使神经网络进行随机变化，并且可以检测误差函数的变化，需要调整 BP 神经网络的学习速率，因此采用了自适应学习速率。自适应学习速率公式为：

$$\eta\left(d+1\right)=\begin{cases}\beta_1\cdot\eta\left(d\right) & E\left(d\right)<E\left(d-1\right)\\ \beta_2\cdot\eta\left(d\right) & E\left(d\right)>\varepsilon\cdot E\left(d-1\right)\\ \eta\left(d\right) & \text{其他}\end{cases} \quad \text{（式 4-21）}$$

其中β_1，β_2代表学习速率调整比例因子，ε代表可以在范围内的反弹误差系数，ε取值大于 1。

4.2.5 BP 神经网络学习过程算法

BP 神经网络学习过程算法描述如下：

（1）将输入层与隐含层之间、隐含层与输出层之间连接的权值w_{ij}和w_{ki}，阈值θ_i和α_k，学习速率λ初始化，迭代次数为1=1，累计误差S置为 0；

（2）输入简牍文字样本特征向量$x_1,\ldots,x_j,\ldots x_M$，网络期望输出为$d_1,\ldots,d_k,\ldots,d_L$；

（3）计算隐含层输出节点和 BP 神经网络输出节点计算误差；

（4）修正输出层的权值Δw_{ki}和隐含层的权值Δw_{ij}；

（5）计算总误差S；

（6）查看是否对全部的训练样本完成了训练，如果没有完成，重返步骤（2）；

（7）总误差是不是符合精度要求，如果满足S $<$ ε，则训练结束，没有满足条件，需要重新调整各层权值和阈值，重返步骤（2）。

BP 神经网络学习过程算法流程如图 4-2 所示：

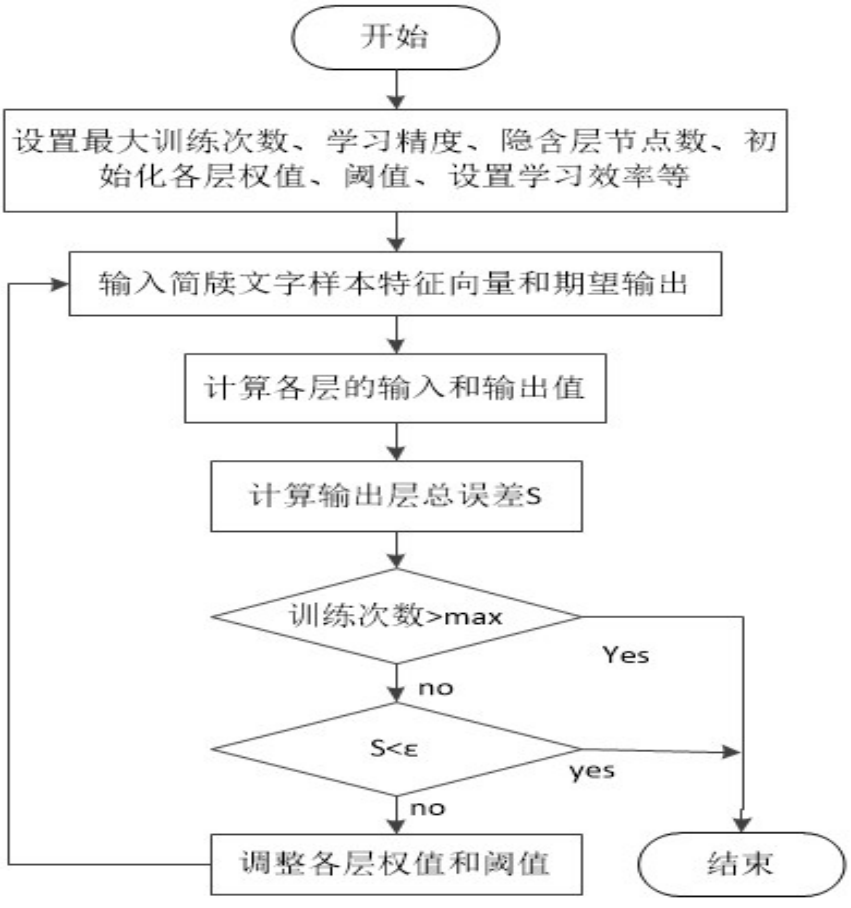


图 4-2 BP 神经网络算法流程图

4.3 实验结果与分析

本文简牍文字的样本库是选自《银雀山汉简文字编》和《上海博物馆藏战国楚竹书文字编》，简牍文字的样本有：为，者，王，公，君，有，则，也，而，所，大，多，吾，曰，兵等等 50 个汉字集，每个字采集了 120 个样本，一共 6000 个样本。仿真实验的部分文字样本如图 4-3 所示：

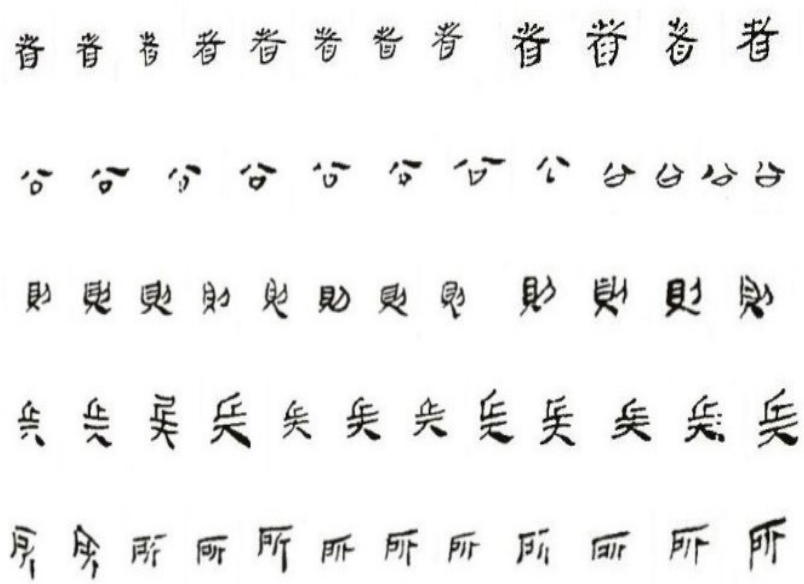


图 4-3 简牍图像样本图

（1）首先输入简牍文字图像，对简牍文字图像进行预处理。简牍图像部分样本进行了图像预处理后效果如图 4-4 所示：

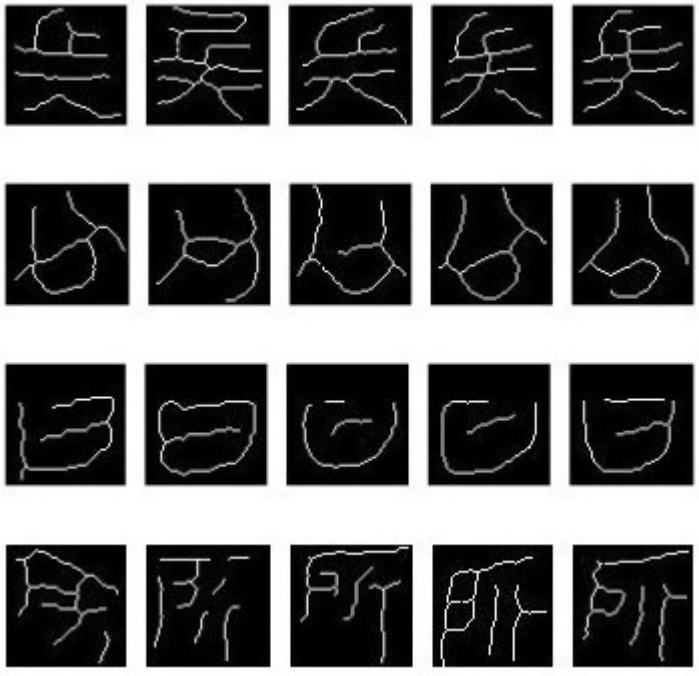


图 4-4 预处理效果图

（2）然后特征提取与多特征融合。本文识别上下，左右，独体和内外这四种结构特征，对应的特征值进行定义如表 4-1 所示：

表 4-1 结构特征值表

结构特征	上下	左右	独体	内外
结构特征值	20	30	40	50

由于样本数量有限，对广义密度特征划分类别：简单，中等，复杂。分类系数为：$\alpha_i\ (i=1,2)$为了避免出现误识和漏识的情况，添加了调整因子$\Delta_i\ (i=1,2)$。密度区间与阈值对应关系为表 4-2 所示：

表 4-2 阈值与密度关系表

阈值 / 结构特征	α_1	Δ_1	α_2	Δ_2
上下	0.018	0.003	0.028	0.006
左右	0.016	0.005	0.032	0.004
内外	0.019	0.004	0.026	0.008
独体	0.021	0.006	0.033	0.005

之后对简牍文字图像进行双弹性网格的划分，进行特征融合，得到特征矩阵S_1, S_2, S_3，S_1代表只进行纵横弹性网格的多特征融合，S_2代表只进行对角弹性网格的多特征融合，S_3代表双弹性网格下的多特征融合。S_1和S_2最终得到8×8维的特征向量，S_3为$8\times 8\times 2$维的特征向量。

（3）最后进行 BP 神经网络的设计。本文是针对简牍图像文字识别的背景，进行的 BP 神经网络设计的研究。

输入层神经元个数：本文中根据特征向量的维数，两个 BP 神经网络，一个节点数为 64 个，一个为 128 个。

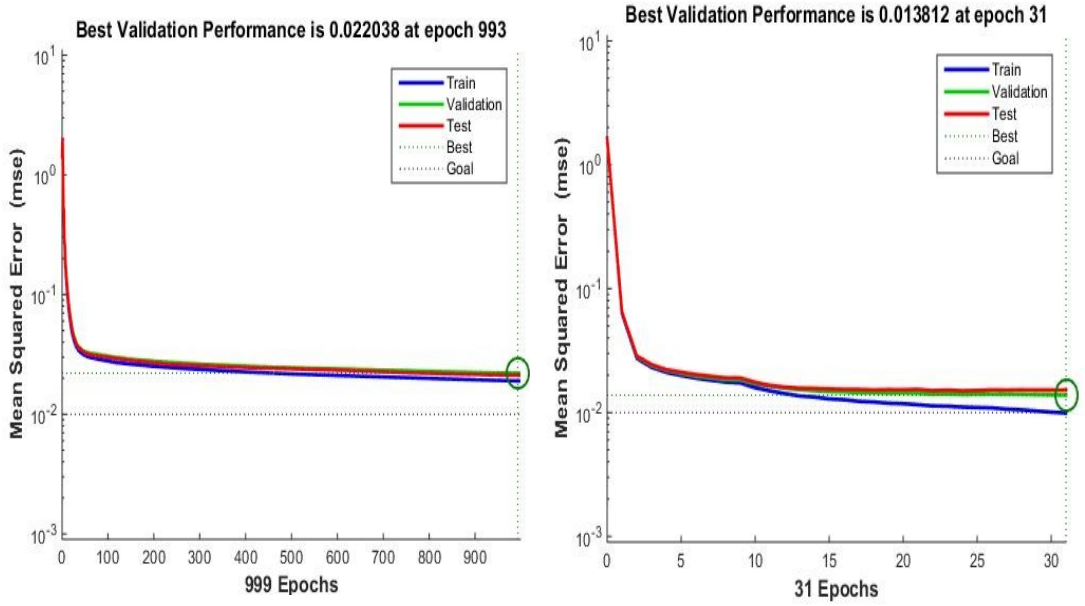
输出层神经元个数：本文使用二进制编码，识别个数 50 个，则在输出层的神经元个数是 6。

隐含层神经元个数：本文运用三层 BP 神经网络的结构，根据公式：$n=\sqrt{M+L}+a$，a 取 1 ~ 10 之间的整数。可得 $n_1=\sqrt{64+6}+(1\sim 10)$，$n_2=\sqrt{128+6}+(1\sim 10)$ 结果为隐含节点数的基础，再通过实验不同隐含节点数有不同的神经网络收敛效果，如表 4-3 所示，因此本文中两个 BP 神经网络都取隐含层个数为 20。

表 4-3 神经网络实验结果表

隐含层神经元个数	迭代次数	训练时间（秒）
5	2201	90
10	2021	71
14	1505	52
18	608	22
20	534	20
24	701	25
28	788	28
30	799	34
40	490	39

在 MATLABR2014b 应用软件环境下，BP 神经网络结构 64-20-6 下，设标准 BP 神经网络最小均方误差设为 0.01，动量值为 0，学习速率定常数 0.05。改进的 BP 神经网络在相同的设定下，动量值为 0.9，采用自适应学习速率及附加动量项相结合的方法，对同一简牍样本采用8×8网格局部密度特征提取，均方误差对比如图 4-5 所示：



（a）标准 BP 神经网络均方误差图　　（b）改进 BP 神经网络均方误差图

图 4-5 均方误差对比图

从图 4-5 中可以看出改进的 BP 神经网络收敛速度快，能达到高要求的误差，效果较好。因此考虑使用改进的 BP 神经网络进行简牍文字识别。

为了更加充分证明改进 BP 神经网络的优越性，以及进行多特征融合的准确性与有效性，在 BP 神经网络分类器结构分别为 64-20-6 和 128-20-6 下，在标准 BP 神经网络中，输入层和隐含层采用 S 型双曲正切函数为传递函数，输出层采

用 S 型对数函数，最小均方误差设为 0.01，动量值为 0，学习速率定常数 0.05。改进的 BP 神经网络在相同的设定下，动量值为 0.9，采用自适应学习速率及附加动量项相结合的方法。基于多特征融合对简牍文字识别，实验通过下面 4 种方案进行了识别研究：

a) 仅采用8 × 8网格局部密度对简牍文字提取识别。

b) 仅采用纵横弹性网格多特征提取。

c) 仅采用对角弹性网格多特征提取。

d) 双弹性网格的多特征提取。

基于 BP 神经网络简牍文字识别实验结果如表 4-4 所示：

表 4-4 识别结果对比表

实验方案	BP 网络类型	识别率（%）	识别时间（s）
方案一	标准 BP 神经网络	49.78	20.16
	改进 BP 神经网络	63.41	30.23
方案二	标准 BP 神经网络	57.76	43.11
	改进 BP 神经网络	76.01	58.62
方案三	标准 BP 神经网络	58.67	47.77
	改进 BP 神经网络	73.43	68.43
方案四	标准 BP 神经网络	67.29	54.89
	改进 BP 神经网络	82.03	129.51

通过实验对比可知，采取多特征融合，简牍文字识别率较高，但是如果特征向量较高，所用识别时间也比较长，但在可接受范围。因此本文多特征融合算法为有效的。并且运用改进的 BP 神经网络识别率要高于标准 BP 神经网络。因此最好使用改进的 BP 神经网络进行简牍文字识别研究。

本文识别率仍有提高的空间，是因为存在样本识别不精确的问题，由于简牍文字属于毛笔书写，笔划随性，笔迹粘连，埋藏时间长，腐蚀比较严重，汉字结构复杂等等，特征提取后的分类识别会出现一定的错误。并且由于时间所限，收集到简牍文字样本有一定的局限性，数量相对较少，也会在一定程度上影响识别率。

4.4 小结

本章主要围绕使用 BP 神经网络对简牍文字识别的研究，这是针对简牍文字的探索的一种新的方式，探寻将有效的数字化图像处理技术和文字识别技术运用到简牍文字的识别研究中，既能加快简牍文字探究的脚步，更为从事考古的研究学者提供技术上的便利。

5 总结与展望

5.1 总结

简牍是纸张出现之前中国历史上最早的文字载体，是中国古代文化中不可缺少的一部分，考古学家们做了大量的研究，是对中华民族文化研究的重要文献，对中国古文化的影响不容忽视，因简牍年代久远，出土时会有光照的影响，空气的腐蚀，对简牍文字的研究需要投入大量的人力与物力，研究进度十分缓慢。随着科技的发展，能够利用数字图像处理、模式识别技术，对简牍文字进行识别研究不仅可以节省人力和物力，并且加快研究的进程从而早日能够对中国古文化进行更深入的研究。研究的成果如下：

1.详细论述了简牍文字图像识别这方面研究的背景与意义，数字图像处理技术以及文字识别的国内外现状，并对几种分类识别技术进行了理论介绍。

2.在对简牍图像预处理中，对简牍图像进行了灰度化、二值化、去噪、归一化、反色、以及细化处理，并在去噪处理时，提出新的自适应加权中值滤波去噪算法，得到较好的去噪效果，保留更多文字的细节，较好的解决了降噪与保护细节的矛盾，从而使得最终的预处理效果较好，有利于后续的识别工作。

3.在特征提取方面，将结构特征、整体广义密度特征、双弹性局部密度特征这三种特征进行融合，在标准 BP 神经网络和改进的 BP 神经网络中输入融合的特征向量实现简牍文字识别的研究，实验最后证明了多特征融合的优越性、可行性，以及有效性。

5.2 展望

本文将 BP 神经网络用于基于多特征融合的简牍文字图像的汉字识别，取得了良好的识别效果，但对于准备数据不够充分和技术仍需改进的这些比较薄弱的地方，希望做进一步的研究。具体如下：

1.提取的特征对简牍文字的识别有很大的影响，在本文中所用的特征依然是最常用的统计特征和结构特征，在特征提取上有很大的局限性，因此从另一个角度提取新的特征加以优化，是下一步的研究重点。

2.本文的简牍文字样本较小，需要进行大字符集的收集，进行多分类器的识别研究，找出最好的分类识别效率。进行分类识别的改进，提高文字的识别率。

3.在本文中对于破损比较严重的简牍文字没有运用一定的修复方法，在一定程度上导致了识别率的降低，希望对简牍文字修复的算法进行探究。

6 参考文献

[1] 中国文物保护现状[EB/OL].（2012-04-11）[2016-12-20]．http://www.gov.cn/test/2012-04/11/content_2110883.htm.

[2] 王云庆，孙家睿．近三十年来我国历史简牍文献的重大发现[J]．文史杂志，2016(07)：91-97．

[3] 赵东．数字化生存下的历史资源保护与开发研究-以陕西为中西[D]．山东：山东大学，2014．

[4] PCIGNONI．Adapptive tetrapuzzles:Efficient out-of-core construction and visualization of gigantic multiresoulution polgonal model[J]．ACM Transaction Graphics，2004(23):796-803.

[5] 鲁东明，习常宇，刘德智.《第二届中华文化遗产数字化及保护研讨会论文集》[A]．上海远东出版社，2007：27-36．

[6] 华忠，鲁东明，潘云鹤．敦煌壁画虚拟复原及演变模拟模型研究[J]．中国图象图形学报，2002(02)：181-186．

[7] 桂恒．书法碑帖字的数字化修复研究[D]．南昌：南昌大学，2013．

[8] 张伟．简牍图像中文字修复的研究与应用[D]．成都：成都理工大学，2008．

[9] 刘瑛．OCR 技术在简牍图像数字化中的应用[D]．成都：成都理工大学，2007．

[10] 覃庆炎．简牍数字图像增强算法研究与应用[D]．成都：成都理工大学，2008．

[11] 张娜．简牍图像增强与分割研究[D]．成都：成都理工大学，2007．

[12] 张阳洁．基于阈值的图像分割技术在简牍中的应用[D]．成都：成都理工大学，2010．

[13] Iijima I，Okumura Y，Kuwabara K．New Process of Character Recognition Using Sieving of Method[J]．Information and Control Research，1963(1)：30-35.

[14] Casey R，Nagy G．Recognition of Printed Chinese Characters[C]．IEEE Transactions On Electronic Computers．1966，EC-15(1)：91-101．

[15] 朱学庆．脱机手写体汉字识别的研究与实现[D]．北京：北京大学，2000．

[16] 张炘中．汉字识别技术[M]．北京：清华大学出版社，1992．

[17] 张凯歌．基于 K-means 和神经网络算法的图像文字提取与识别[D]．云南：云南大学，2013．

[18] S.Mori，K.Yamamoto，M.Yaduda．”Research on Machine Recognition of handprinted character”[J]．IEEE Trans Patter Analysis Marc，Intell，Pam，1984：386-405.

[19] 唐方坤．基于仿生模式识别和多权值神经网络的脱机手写汉字识别研究[D]．四川大学，2014．

[20] Tang X O，Lin F，Liu J Z. Video-based Handwritten Chinese Character Recognition[C]. IEEE Transactions on Circuits and Systems for Video-based technology，2005：167-174.

[21] 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)：74-77．

[22] 刘妍．基于 Gabor 双弹性网格特征提取的手写体汉字识别的研究[D]．河北：河北工业大学，2014．

[23] 周新伦，李锋，华星城等．甲骨文计算机识别方法研究[J]．复旦大学大学学报（自然科学版）．1996，35(5)：481-485．

[24] Teow，L.N，Loe，K.F.Robust ．vision-based features and clasffication schemes for on-line handwritten digit recongnition[J]．Pattern Recognition 2002，5(11)：2235-2246.

[25] 史小松．基于支持向量机的甲骨文字结构分析研究[D]．上海：华东师范大学，2010．

[26] 刘磊．基于内容的秦汉瓦当小篆文字识别方法研究[D]．西安：西北大学，2015．

[27] 顾邵通．基于拓扑配准的甲骨文字形识别方法[J]．计算机与数学工程．2016(10)：2001-2006．

[28] Li B L，Qin L， Yu S W． An Adaptive k-Nearest Neighbor Text Categorization Strategy[C]. ACM Trans on Asian Language Information Processing，2004，3(4) ：215- 226

[29] Mark S.Nixon，Alberto S.Aguado 著，石英 ，杨高波 ，译．特征提取与图像处理（第二版）[M]．北京电子工业出版社， 2010．

[30] 柳回春，马树元．支持向量机的研究现状[J]．中国图象图形学报，2002， 7(6)：618-623．

[31] 刘迎建， 戴汝为，张立清．基于神经网络的手写汉字特征选择[J]．模式识别与人工智能，1992， 5(1)：37-43．

[32] Vapnik V．The Nature of Statistical Learning Theory[M]．Springer，1995.

[33] 孙莹莹．基于混合核 LS-SVM 的古汉字图像识别[D]．安徽：安徽大学，2015．

[34] 孙华．基于多特征融合 SVM 的古汉字[D]．长沙：中南大学，2010．

[35] OstuN． Athresholdselectionmethodfromgray-levelhistograms[J]． Automatica， 1975，11(285-296)：23-27.

[36] 周星辰．基于深度模型的脱机手写体汉字识别研究[D]．浙江：浙江大学，2016．

[37] Paulina M，Usinskas A．A survey of genetic algorithms applications for image enhancement and segmentation[J]．Information Technology and control，2015，36(3).

[38] 郭晓新，卢奕南，许志闻等． 自适应定向加权中值滤波[J]．吉林大学学报理学版，2005，43(4)：494-498．

[39]邓秀勤，熊勇，彭宏． 一种有效的自适应加权中值滤波算法[J]．计算机工程与应用，2009， 45(35)：185-187．

[40]王松林， 蒋峥．改进的自适应加权中值滤波算法[J]．传感器与微系统，2016，35(11)：128-131．

[41]杨柱中， 周激流，郎方年．基于分数阶微积分的噪声检测和图像去噪[J]．中国图象图形学报， 2014，19(10)：1418-1429．

[42]祖丽菲亚•卡哈尔．基于特征组合的联机手写维吾尔文字母识别技术研究[D]．新疆：新疆大学，2013．

[43]J.Y.Lin，Z.Chen．A Chinese character thinning algorithm based on global features and contour information[J]．Pattern Recognition，1995，28(4)：493-512

[44]hilditch CJ．Linear Skeletons from Square Cupboards[J]．Machine Intelligence，B.Meltzer and D.Michie，Eds ，Elsevier，New York，1969，4：403-420.

[45]白莹．手写汉字的细化算法研究[D]．西安：西安电子科技大学，2014．

[46]Zhang TY，Suen C Y．A fast parallel algorithm for thinning digital patterns[J]．Comm ACM，1984，27：236- 239.

[47]王岩．离线手写体汉字鉴别及识别算法研究[D]．河北：河北工业大学，2013．

[48]侯伟．基于移动平台的联机手写汉字识别[D]．西安：西安电子科技大学，2010．

[49]刘京超．面向专利发布特殊字符的识别系统[D]．河北：河北工业大学，2009．

[50]刘雨心．基于笔画的脱机手写体汉字识别与研究[D]．太原：太原理工大学，2014．

[51]马继丰．基于决策层信息融合的手写汉字识别研究[D]．西安：西安科技大学，2007．

[52]郭萍萍．SVM 多分类关键技术研究及其在车牌字符识别中的应用[D]．大连：大连海事大学，2012．

[53]Simon Haykin 著， 叶世伟，史忠植译．神经网络原理[M]．北京：北京机械工业出版社，2004．

[54]伍文源．基于 BP 神经网络的湘西方块苗文图像识别研究[D]．湖南：吉首大学，2016．

致　谢

时间在不经意间匆匆流走，两年的研究生学习生涯即将落下帷幕，仿佛一瞬间走完了这段人生的重要旅程。回首这两年来生活的点点滴滴，欢笑的，迷茫的心酸的，努力奋进的这些满满回忆的片段，我收获良多。在这里要对所有关心我，帮助我的人表示最衷心的感谢。

首先要衷心的感谢我的导师杨得国教授这两年来对我的指导和帮助，感谢他在学习上对我的认真指导，在生活上的亲切关怀。尽管学习和工作都很繁忙，但导师会抽出时间对我们的学习和科研工作进行细致的指导。导师严谨的治学态度，在为人处事上的谦和近人，以及做事情的认真负责，一直激励着我继续热情积极的学习，令我终生受益。

感谢我的同学、室友、师姐、师弟、师妹以及一个实验室的同学们。他们在这两年来在生活上和学习上给予我很大的帮助和关怀。陪我一起度过了两年的美好时光，我将永远铭记。

感谢我的父母在我漫漫求学的路上一直支持我，鼓励我，即使我并不优秀，但仍以我为荣，给我最温暖的港湾。无论我做什么决定，都是我坚强的后盾。你们无私的爱和关心是我不断的动力。

张兰云

2017 年 5 月 25 日