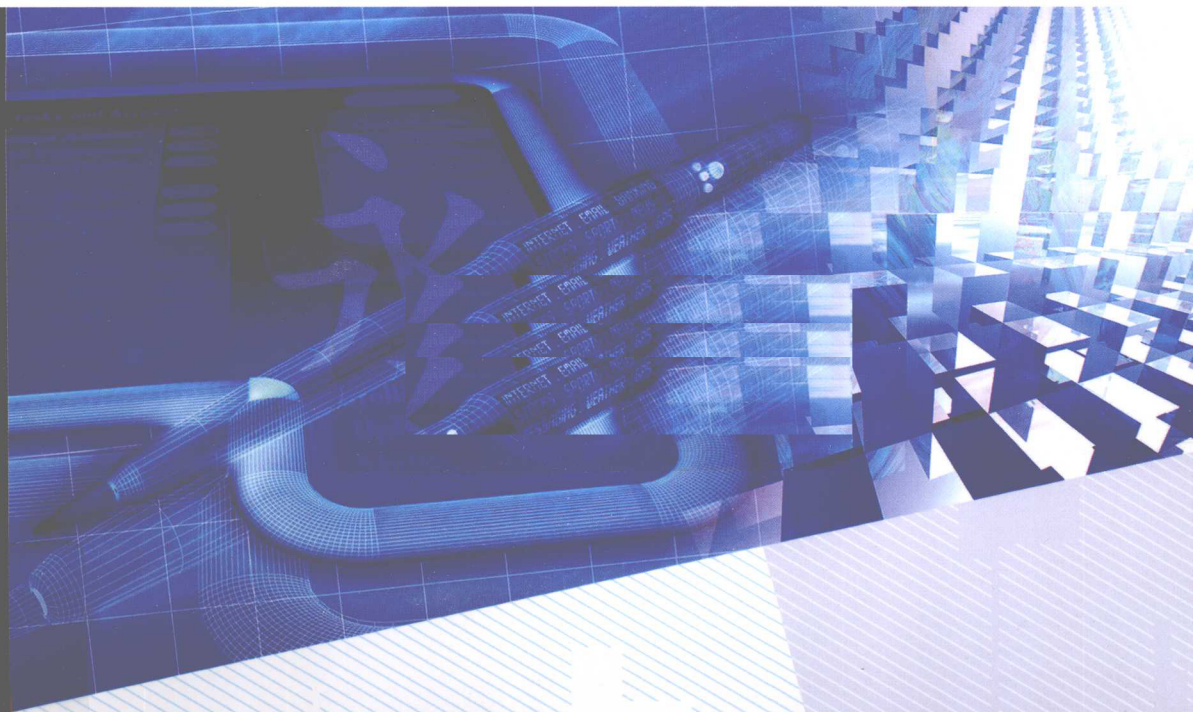
智能科学技术著作丛书

中文印刷体文档识别技术

王科俊　冯伟兴　著



科学出版社
www.sciencep.com

智能科学技术著作丛书

中文印刷体文档识别技术

王科俊　冯伟兴　著

科学出版社

北京

内 容 简 介

本书全面阐述了中文印刷体文档识别的原理、方法和系统组成。依据中文印刷体文档的特点，分别介绍了文档图像预处理、版面分析、汉字识别、公式的定位与提取、公式字符分割与识别、公式结构分析与表示、图表处理等内容的基本原理和技术实现方法，并提供了一个中文印刷体文档识别系统实例。

本书可作为研究公式识别、表格识别和汉字识别技术的参考书，可供从事图像处理、文字识别的研究人员阅读，也可作为计算机、信息工程、电子技术、自动化等相关学科专业的研究生和高年级本科生学习图像处理、模式识别技术的辅助教材参考使用。

本书还配有完整的实例代码光盘，供读者参考。

图书在版编目(CIP)数据

中文印刷体文档识别技术 / 王科俊，冯伟兴著. —北京：科学出版社，2010

(智能科学技术著作丛书)

ISBN 978-7-03-028760-1

Ⅰ.①中… Ⅱ.①王…②冯… Ⅲ.①计算机应用-印刷字体-文字识别 Ⅳ.①TP391.43

中国版本图书馆 CIP 数据核字(2010)第 164067 号

责任编辑：张海娜 / 责任校对：赵桂芬

责任印制：赵 博 / 封面设计：耕者设计工作室

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂印刷

科学出版社发行 各地新华书店经销

*

2010 年 8 月第 一 版 开本：B5(720×1000)

2010 年 8 月第一次印刷 印张：13 1/2

印数：1—3 000 字数：256 000

定价：42.00 元(含光盘)

(如有印装质量问题，我社负责调换)

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象包括：

• “自然智能”(natural intelligence，NI)，包括“人的智能”(human intelligence，HI)及其他“生物智能”(biological intelligence，BI)。

• “人工智能”(artificial intelligence，AI)，包括“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。

• “集成智能”(integrated intelligence，II)，即“人的智能”与“机器智能”人机互补的集成智能。

• “协同智能”(cooperative intelligence，CI)，指“个体智能”相互协调共生的群体协同智能。

• “分布智能”(distributed intelligence，DI)，如广域信息网、分散大系统的分布式智能。

1956 年，“人工智能”学科诞生，50 年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说，当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么，可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981 年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25 年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创

新能力、建设创新型国家做出应有的贡献。

祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广
人机集成智能强
群体智能协同好
智能创新更辉煌

涂序彦

中国人工智能学会荣誉理事长
2005 年 12 月 18 日

前　言

随着科技的发展，人类社会正经历从工业化社会向信息化社会的转变，信息化程度越来越高。近年来，伴随互联网的迅速普及，通过互联网这一方式进行信息传播和交换已成为人们日常工作生活的首选信息交流方式。为了促进信息交流效率，中文印刷体文档识别技术日益受到众多学者的关注。

目前，具有汉字和符号识别功能的印刷体识别软件（OCR）已在实际中得到广泛应用，但是一个中文文档中不仅含有汉字和符号，还含有特殊字符以及各种各样的公式和图表。而现阶段的中文文档识别软件尚不能对公式等这些文档内容进行识别和处理，迫切需要一种既能识别汉字又能识别和处理公式等其他文档内容的较为全面的中文文档识别系统。针对这一现状，我们开展了以公式为主的中文印刷体文档识别研究，本书就是我们近几年来在这一领域研究成果的总结。

本书作为国内第一部关于中文印刷体文档识别技术的著作，系统地分析了中文印刷体文档识别技术的各个方面，包括文档图像的预处理、版面分析、文字和符号识别、公式定位和提取、公式结构分析与表示、表格识别和文档中的图形图像处理等内容。结合作者多年来在公式识别方面取得的研究成果重点给出了公式的定位与提取和公式的结构分析的理论与方法。本书还给出了一个含有文字、公式和表格识别功能的中文印刷体文档识别系统软件实现方法及相应的实现代码。通过阅读本书，读者可以充分了解中文印刷体文档识别技术的基本原理和相关算法，有助于该项技术的进一步研究和开发。

本书的研究工作得到教育部教师骨干培训计划项目、黑龙江省科技攻关项目（gc04a114）、黑龙江省杰出青年基金项目（JC200703）和哈尔滨市优秀学科带头人项目（2007RFXXG009）的资助，并得到了哈尔滨工程大学自动化学院的大力支持。

本书是哈尔滨工程大学模式识别与智能系统研究所公式识别研究小组集体工作的结晶。王科俊统筹全书，并撰写了与公式识别技术的相关章节，其余章节由冯伟兴撰写。感谢李永华、陈卉、刘维平、高天孚、吴俊飞、王黎斌、林桂芳、李蕊等同学所付出的辛勤劳动。感谢作者家人的大力支持和理解。

本书要求读者具有数字图像处理、模式识别等相关知识。由于作者水平有限，书中不妥之处在所难免，敬请读者批评指正。

作　者

2010年4月

于哈尔滨工程大学

目　录

《智能科学技术著作丛书》序
前言
第 1 章　绪论 …… 1
1.1　中文印刷体文档识别基本原理 …… 1
1.2　中文印刷体文档识别研究现状 …… 2
1.2.1　印刷体文档的汉字识别 …… 2
1.2.2　印刷体文档的公式识别 …… 4
1.2.3　印刷体文档的表格识别 …… 6
1.3　中文印刷体文档识别中的难点 …… 6
第 2 章　中文印刷体文档图像预处理 …… 8
2.1　中文印刷体文档图像采集 …… 8
2.1.1　文档图像采集 …… 8
2.1.2　文档图像显示 …… 8
2.1.3　文档图像格式 …… 9
2.2　中文印刷体文档图像特点 …… 12
2.3　二值化处理 …… 12
2.3.1　图像灰度化 …… 13
2.3.2　图像二值化 …… 13
2.4　平滑去噪 …… 18
2.4.1　邻域平均法 …… 18
2.4.2　中值平均法 …… 18
2.4.3　噪声直接去除法 …… 19
2.5　倾斜校正 …… 20
2.5.1　图像倾斜检测 …… 20
2.5.2　图像倾斜校正 …… 26
第 3 章　版面分析 …… 30
3.1　版面结构 …… 30
3.2　版面分析方法 …… 31
3.2.1　基于连通域的版面分析方法 …… 33
3.2.2　二分法 …… 34

3.2.3　基于组合特征的版面分析方法 …… 36
3.2.4　基于神经网络的版面分析方法 …… 37
3.2.5　基于最近邻连接强度和行列可信度的版面分析方法 …… 38
3.3　版面理解 …… 44
3.3.1　文字区域 …… 44
3.3.2　图片区域 …… 44
3.3.3　表格区域 …… 45
3.3.4　版面结构表示与存储 …… 45
3.4　版面重构 …… 51
第 4 章　印刷体汉字识别 …… 52
4.1　文本区域预处理 …… 52
4.1.1　文本增强 …… 53
4.1.2　字符分割 …… 53
4.1.3　字符细化 …… 54
4.1.4　字符归一化 …… 55
4.1.5　文本区域处理效果图 …… 57
4.2　印刷体汉字的特征提取 …… 58
4.2.1　印刷体汉字的统计特性 …… 58
4.2.2　印刷体汉字的常用特征 …… 62
4.3　印刷体汉字识别的实现方式 …… 65
第 5 章　公式的定位与提取 …… 71
5.1　印刷体文档公式的特点 …… 72
5.2　基于投影的公式定位和提取 …… 72
5.2.1　独立行公式的定位 …… 72
5.2.2　内嵌公式的定位 …… 74
5.3　基于 Parzen 窗的独立行公式定位和提取 …… 75
5.3.1　待分类文本行的特征数据提取 …… 75
5.3.2　Parzen 窗方法 …… 76
5.3.3　公式定位与提取效果 …… 77
5.4　基于字符宽度中心矩的公式定位和提取 …… 78
5.4.1　文本区域基本数据获取 …… 78
5.4.2　含公式的文本行提取 …… 79
5.4.3　文本行中公式判别 …… 81
5.4.4　独立行公式的定位 …… 83
5.4.5　内嵌公式的定位 …… 83

5.4.6 公式定位与提取效果 …… 84
5.5 基于汉字拒识的内嵌公式定位和提取 …… 85
5.5.1 内嵌公式的定位 …… 85
5.5.2 公式定位与提取效果 …… 86
第 6 章 公式字符分割与识别 …… 88
6.1 公式字符的特点 …… 88
6.2 公式字符的分割 …… 89
6.2.1 基于轮廓跟踪的字符分割 …… 90
6.2.2 基于连通域的字符分割 …… 92
6.3 公式字符的识别 …… 97
6.3.1 公式字符图像预处理 …… 97
6.3.2 基于模板匹配的公式字符识别 …… 99
6.3.3 基于特征的公式字符识别 …… 100
6.3.4 印刷体公式字符识别的实现 …… 104
6.3.5 公式字符识别方法 …… 104
第 7 章 公式结构分析与表示 …… 107
7.1 公式结构分析的难点 …… 107
7.1.1 数学运算符的模糊性 …… 107
7.1.2 符号的上下文敏感性 …… 107
7.1.3 表示习惯的差异性 …… 108
7.1.4 公式的复杂性 …… 108
7.1.5 公式的多行结构 …… 108
7.2 公式结构分析前的字符预处理 …… 108
7.3 公式结构分析方法 …… 109
7.4 公式结构表示方法 …… 120
7.4.1 公式的典型表示方法 …… 120
7.4.2 实验结果 …… 124
第 8 章 图表处理 …… 129
8.1 文档中图形图像的表示与处理 …… 129
8.1.1 游程压缩 …… 129
8.1.2 霍夫曼编码压缩 …… 130
8.1.3 算术压缩方法 …… 131
8.1.4 Rice 压缩方法 …… 131
8.1.5 LZW 压缩方法 …… 131
8.2 文档中表格的分析与识别 …… 132

8.2.1　表格预处理 …… 132
8.2.2　表格直线提取 …… 139
8.2.3　表格结构分析 …… 142
8.2.4　表格字符提取与识别 …… 143
第 9 章　中文印刷体文档识别软件 HEUOCR 的设计与实现 …… 144
9.1　应用程序框架的构建 …… 144
9.1.1　框架风格 …… 144
9.1.2　数字图像处理类 …… 146
9.2　文档图像预处理 …… 152
9.2.1　图像灰度化 …… 153
9.2.2　图像平滑滤波 …… 155
9.2.3　图像阈值分割 …… 156
9.3　文档图像版面分析 …… 158
9.3.1　基本连通域提取 …… 159
9.3.2　基本连通域分析 …… 160
9.4　文本汉字识别 …… 162
9.4.1　字符分割 …… 162
9.4.2　字符识别 …… 170
9.5　公式识别 …… 178
9.5.1　公式定位 …… 178
9.5.2　公式字符分割 …… 181
9.5.3　公式字符特征提取 …… 183
9.5.4　公式字符识别 …… 190
9.5.5　公式结构分析 …… 195
参考文献 …… 199

第1章 绪 论

信息化理念已经被很多人所熟悉，人们越来越追求一种有力的、简洁的、准确无误的信息交流手段。由于人们日常生活中接收到的绝大多数信息是以图像的形式进行传递的，尤其是依托互联网的数字图书馆和远程教育的兴起，使得图像信息自动识别技术有着广泛的应用前景和重要的研究价值。中文印刷体文档识别技术就是一个典型的针对含有中文字符图像的信息自动识别技术。

1.1 中文印刷体文档识别基本原理

现有的文字识别技术一般采用光学的方式将文字图像信息采集到计算机中，因此，该类技术常被称为光学字符识别(optical character recognition，OCR)技术。经过近一个世纪的发展，OCR已经成为当今模式识别领域中最活跃的研究内容之一[1]。它综合了数字图像处理、计算机图形学和人工智能等多方面的知识，并在计算机及其相关领域中得到了广泛应用。按照识别方法，OCR识别方法可以分为如下三类：统计特征字符识别技术[2]、结构特征字符识别技术和基于人工神经网络的字符识别技术[3]。

作为OCR技术的一个重要研究方向，印刷体文档识别主要针对比较正式、规范的书籍、报刊和杂志的图像信息进行采集和识别。与一般文档图像相比，印刷体文档图像存在前景信息与背景信息色差显著，文字信息形式规范等特点，这都为印刷体文档的信息处理和识别创造了便利条件。然而，各类印刷体文档中除了包含文字信息以外，还常有公式、表格以及各种各样的图形等信息，因此，若将印刷体文档中包含的所有信息都完整地识别出来，也不是一件易事。

相比普通印刷体文档识别，中文印刷体文档识别在对文档所包含的文字字符进行识别过程中，在实现常规西文字符识别的同时，还应结合中文汉字字符特点，实现对中文汉字字符的识别。一个完整的中文印刷体文档识别系统应包含如图1-1所示的诸多模块。

根据图1-1，中文印刷体文档识别系统主要包含以下几部分内容：

(1) 文档图像预处理。该部分完成对原始文档图像的预处理，使得原始文档图像能够达到识别要求。

(2) 文档图像版面分析。该部分实现文档图像中文本、表格和图形图像等不同区域的分离，并在识别出每个区域的类别后交由不同处理模块进行进一步分析

和处理。

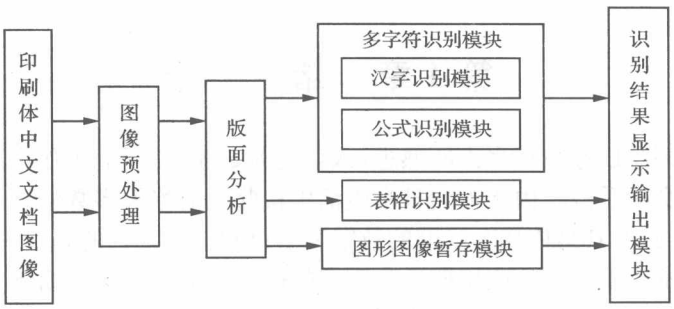


图 1-1 中文印刷体文档识别系统

(3) 文档图像中多字符识别。该部分处理通过文档版面分析得到的文本区域,包括汉字识别与公式识别两个识别模块。其中,汉字识别模块负责包括中文字符在内的所有文字字符的识别;公式识别模块则完成对文本模块中所含有公式的定位与识别。

(4) 文档图像中表格识别。该部分负责处理通过文档版面分析得到的表格区域。作为一类特殊的图像,表格具有重要的识别价值。因此,该部分应具有表格分析与识别能力。其中,对表格中字符的识别可以调用字符识别模块来实现。

(5) 文档图像中图形图像暂存模块。该部分负责处理通过文档版面分析得到的图形图像区域。这里除了实现对图形图像的暂存外,并不需要其他特殊的处理。但需要注意图形图像存储过程中在保证不失真前提下的存储效率。

1.2 中文印刷体文档识别研究现状

近些年,我国在中文印刷体文档识别的技术研究上已经得到了飞速发展,而且市面上的一些针对中文汉字字符的识别软件也已达到较高的识别率。然而,它的发展也是一个漫长而艰辛的过程。

此外,由于很多印刷体文档中的文字嵌有各种公式,因此,印刷体文档的公式识别也是印刷体文档识别中一个重要的组成部分。由于公式的结构要远比汉字的结构复杂,印刷体文档的公式识别已成为印刷体文档识别的难点问题。目前,有关公式识别的技术研究还不成熟,有待进一步发展。

1.2.1 印刷体文档的汉字识别

中国的汉字数量巨大,仅按照 GB2312—80,标准汉字就有 6763 个,其中包括一级汉字 3755 个,二级汉字 3008 个。因此,汉字识别问题属于超多类模式集合的

分类问题。

目前汉字识别技术按照字体的不同可分为[4]：

(1) 单体印刷体汉字识别(single-font printed character recognition)：仅识别某种单一印刷体字体文字或者某种打印机、照排机输出的文字。

(2) 多体印刷体汉字识别(multi-font printed character recognition)：能识别印刷出的多种字体文字，如黑体、宋体、楷体等。

(3) 手写体汉字识别(handwritten character recognition)：用于识别人写在纸上的规整汉字。手写体汉字识别按识别对象又可分为：

① 特定人手写体汉字识别(personal handwritten character recognition)：是手写体识别的一个特例，笔迹鉴别属于这一类。

② 非特定人手写体汉字识别(unconstrained handwritten character recognition)：对于任何人自由书写的汉字都能正确识别，这是手写体汉字识别的最终目的。

由于印刷体字体比较规范，因此，印刷体汉字识别相对于手写体汉字识别难度要小一些。

根据输入方式的不同，汉字识别技术又可分为联机识别与脱机识别。

(1) 联机识别(也称在线识别)：是指采用电子笔在电子板上边写边识别。联机识别时，人们在写字板上手写的字符数据(即笔触板时的触点坐标)，通过计算机的串口或并口传入计算机。然后，计算机程序对字符数据进行预处理、特征抽取、特征匹配和分类识别。这种识别方式，较适合个人和小规模数据的录入使用。

(2) 脱机识别(也称离线识别)：它是对已经写在纸上的字符通过扫描仪将其图像输入计算机，然后，计算机程序对字符图像进行预处理、字符定位、字符分割、特征抽取、特征匹配和分类识别。脱机识别由于书写和识别分开进行，适合大批量数据的集中录入。

由于可以提取更多的汉字特征，联机识别相对于脱机识别难度要小一些，如书写过程中可以直接得到汉字笔画、笔顺等特征；另外，联机识别可以规定汉字的书写顺序，如从左到右，这也使得联机识别相对比较容易。

根据印刷体汉字识别的识别过程，印刷体文档汉字识别只能采用脱机识别方式。

据文献记载，印刷体汉字的识别最早可以追溯到20世纪60年代[5]。1966年，IBM公司的Casey和Nagy在一篇文章中利用简单的模板匹配法识别了1000个印刷体汉字。1977年，东芝综合研究所研制了可以识别2000个汉字的单体印刷体汉字识别系统。80年代初期，日本舞藏野电气研究所研制了可以识别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，大都采用了基于 K-L 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，因而并没有得到广泛的应用[6]。

我国对印刷体汉字识别的研究开始于 70 年代末、80 年代初，大致分为以下三个阶段：

(1) 第一阶段：70 年代末期～80 年代末期，主要是算法和方案探索。

(2) 第二阶段：90 年代初期，中文 OCR 由实验室走向市场，初步试用。

(3) 第三阶段：90 年代至今，主要是印刷体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。

虽然汉字识别在我国研究的起步较晚，然而经过多年的努力，印刷体汉字识别技术的发展和应用已有了长足的进步。从简单的单体识别发展到多种字体混排的多体识别，从中文印刷文档的识别发展到中英文混排的文档识别。如今，各种汉字系统可以支持简、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，同时汉字识别率也已达到了 98%以上。

1.2.2 印刷体文档的公式识别

由于很多印刷体文档中除了包含文字信息以外，还常嵌有各类公式，因此，印刷体文档的公式识别也是印刷体文档识别中一个重要的组成部分。由于公式的结构要远比汉字的结构复杂，印刷体文档的公式识别已成为印刷体文档识别的难点问题。我国目前对于公式识别的研究尚处于起步阶段，其技术水平还远没有汉字识别那么成熟。

公式识别技术是文档识别领域国内外学者们研究的热点与难点问题之一。国外对公式识别技术的研究起步较早，也较为成熟。

1968 年，Anderson 在其博士学位论文中最早提出了公式处理问题[7,8]，随后 10 年陆续有几篇文章发表[9,10]。这些工作大部分只提出了理论上的处理方法，主要思路是试图构造完备的文法来描述公式的结构。虽然构造的文法能够描述类型非常复杂的公式，但是研制的实际系统只能处理包含几个符号的简单公式。

20 世纪 70 年代，公式识别的研究工作进展缓慢，沉寂了将近 10 年。直到 20 世纪 80 年代中后期，才有少量文章发表[11,12]。这期间的工作不再追求识别公式类型的多和全，而是专注于一种或几种类型公式的识别。但这些方法一般都要穷举所有符号之间的关系，因此计算开销很大。研究工作如此长时间的缓慢发展是由于 20 世纪 60 年代是模式识别技术初创时期，理论尚不完备，OCR 技术还不成熟，因此研究工作无法与实际相结合。

20 世纪 90 年代以来，公式识别的研究热度逐渐增加，这期间有大量的论文发表。研究工作覆盖了公式识别的主要研究领域，例如公式自动定位[13~15]、公式符号切割与识别[16~19]、公式的结构分析[20~23]等。同时，一些研究工作还考虑了实际

应用情形[24~30]，比如实现公式识别或分析结果的手动或自动纠错[31]、大规模的在真实图像样张上进行方法测试、讨论公式处理系统的性能评估问题等[32,33]。这一阶段，公式识别研究发展速度较快的原因是OCR技术已经成熟，和公式识别密切相关的版面理解的研究也已经开始。目前，个别较完整的实验系统或者针对具体应用的实际系统已见到报道，但仍然没有出现完整的、针对真实文档图像的、接近于实用的公式识别系统。

对于国内而言，公式识别技术的研究起步较晚，大部分的研究成果都是在2000年之后出现的。由于国内是在中文文档识别的基础上进行公式识别技术研究，这相对于国际上以西文文档作为基础进行公式识别技术研究更加有难度。总体来说，国内对于公式识别的研究技术水平与国际上相比还有一定差距。

下面简要介绍国内公式识别技术的部分研究成果。

公式中任意两个相邻符号之间的位置关系，包括以下几种：上下、水平(左，右)、上标(左上，右上)、下标(左下，右下)和包含。实际上，上/下标符号很少出现在主符号的左边，所以公式上/下标识别中只考虑相邻符号之间存在的上标(右上)、下标(右下)以及水平(右)这三种位置关系即可。文献[34]提出基于统计特征的印刷体公式上/下标判别方法，即利用两个相邻符号最小外接矩形的位置坐标值进行上/下标关系判别。

相比于仅进行公式字符的位置关系分析，文献[35]进一步基于神经网络进行了公式符号分割与识别研究，并利用公式符号的轮廓特征、矩特征设计了一个从文档扫描数据预处理到最终电子文档生成的一个接近实际应用的公式识别系统。该系统主要包括以下几个步骤：文档扫描输入、图像数据二值化、图像倾斜角度的调整、符号分割、利用神经网络进行符号识别以及符号重组。

文献[36]设计了公式定位及公式分析方面的技术方案。在公式定位方面，采用Parzen窗区分独立公式和普通文字行，通过检测二维结构定位内嵌公式；在公式分析方面，定义十一种基本公式类型，提出先识别公式类型，然后根据公式类型对公式进行分解，从而实现公式结构分析。

文献[37]提出了基于多候选的数学公式系统主要包括公式图像预处理，多候选公式符号分割和多候选公式结构分析等三个部分。在公式符号切分中，使用三次动态规划方法对公式图像进行多候选公式符号切分。在公式结构分析中，采用层次结构方法多候选分析公式符号间的结构关系，然后使用LaTex格式和MathType格式表示数学公式的识别结果。为了确定符号间的空间位置关系，建立了符号的空间关系模型。基于多候选的数学公式识别系统能够较为准确地确定符号间的空间关系。

文献[38]研究了公式图像的结构理解与重现，提出了一种鲁棒的公式结构理解方法，使用公式图像识别结果、语法规则和句法规则分析公式结构，对公式的类

型进行了划分，对识别结果的错误进行自动的检查和纠正，能够自动分析公式符号的优先级和计算顺序。

汉王助教先锋是汉王科技研制的汉王文本王的系列产品，主要是用于教学工作方面使用。据报道，汉王助教先锋中内嵌了公式识别技术，该技术可以让使用者轻松地实现各种公式结构的识别[39]。

相对于上述研究，我们早在2002年即开始公式识别的技术研究[40,41]，并取得了一系列研究成果。关于公式定位，提出了利用汉字与公式字符的投影特征对公式字符进行定位，并对左右结构的汉字的定位误识问题也给出了解决方案[42]。还提出了对整个文档中的文字进行汉字识别，并利用识别中的拒识结果进行公式字符定位[43]。关于公式结构分析，提出了基于特征字符的公式结构分析方法[44]。在上述研究基础上，设计了实用的具有公式识别功能的中文印刷体识别系统[45~47]。

1.2.3 印刷体文档的表格识别

表格识别是文档识别技术中一个较新的研究内容。人们在日常工作、学习和生活中经常需要填写各种各样的表格：财务报表、商业数据统计表、税务统计表、学生成绩表等，而这些表格中的大量信息常常需要输入到计算机进行整理、归类、排序和分析等高层次的应用，因此，人们迫切需要一种表格自动识别系统来替代繁重的人工输入操作。一套高准确率、高效率和健壮的表格识别系统能够大大加快表格信息输入速度、提高工作效率，从而有着巨大的研究意义。

表格识别技术的研究与发展在时间上滞后于文字识别技术。国内开始进行印刷体文档识别研究时，大部分专家都将精力放在印刷体文字识别技术的研究上，很少有人进行表格识别技术研究。自2000年以来，随着印刷体文字识别技术的逐步成熟，表格识别技术的研究得到重视，进而出现了像清华文通、汉王等在文字识别基础上兼顾表格识别的研究机构。

对表格识别来说，它不同于一般文字识别。表格的形式变化多样，很难找到一种能较好地识别任何表格的通用方法，致使表格单元字符的识别准确率还远远低于纯文本中的字符。另一方面，在实际应用中，很多领域使用的表格，如银行、邮局、税务等，需要识别的通常是固定的某些表格单元，因此，目前很多表格研究都是针对结构已知的表格进行识别。典型的结构已知的表格识别系统包括邮政编码自动识别、金融票据识别、车牌识别等，而关于较复杂的通用表格识别的研究也有一些，但成型的实用系统较少，理论也不够完善。

1.3 中文印刷体文档识别中的难点

近些年，中文印刷体文档识别系统一直备受研究者的关注。针对该系统的研

究发展很快并取得了很多技术成果，但仍有一些难点问题有待进一步解决。

首先，尽管清晰规范的版面格式为中文印刷体文档识别提供了便利条件，但中文印刷体文档的版面内容仍然是多样的，这意味着不能对版面中的每一个图像区域采取相同的处理和识别方法，因此，文档识别的一个难点是如何将文档图像中的不同版面区域进行有效分割并进行准确的类别识别。

此外，很多现有的中文文档 OCR 识别系统，其汉字的识别率已经相当可观，但是，对于中文文档中出现的公式还无法进行识别。其中，如何将汉字与公式字符分开是一直以来的难点，特别是对内嵌在汉字中的公式字符的定位与分割更为困难。而且，就公式而言，它不像汉字是以一维形式存在的，而是由不同字符的各种结构组合而形成的二维形式。换句话说，对一个公式识别，不能单一地只把公式中的各个字符识别出来，还要对公式中各个字符之间的结构排列关系进行分析与识别，这一点也是公式识别的难点问题。

和公式识别一样，表格识别也是印刷体文档识别的一个重要研究内容。由于相对于文字和公式而言，表格具有复杂的二维结构，而且，表格内含有的字符除了常用的文字外，还可以是简单的公式，这都使得表格识别成为文档识别的另一个难点。