结构化汉字信息处理

孙星明　胡运发　著

国防科技大学出版社
·湖南长沙·

图书在版编目(CIP)数据

结构化汉字信息处理 / 孙星明，胡运发著：—长沙：国防科技大学出版社，2001.8
ISBN 7-81024-774-3

Ⅰ.结… Ⅱ.①孙…②胡… Ⅲ.①汉字②汉字结构 Ⅳ.H124

国防科技大学出版社出版发行
电话：(0731)4572640 邮政编码：410073
E-mail:gfkdcbs@public.cs.hn.cn
责任编辑：卢天贶 责任校对：何 晋
新华书店总店北京发行所经销
国防科技大学印刷厂印装
*
开本：787×1092 1/16 印张：9.5 字数：231千
*
2001年8月第1版第1次印刷 印数：1-1 500册
*
定价：15.00元

前　言

汉字是世界上使用人数最多、使用历史最长的文字，为中华民族灿烂文化的形成和发展立下了不可磨灭的功勋，并将继续发挥重要的、其他文字形式难以取代的作用。然而，随着计算机应用的迅猛发展，各行业对计算机汉字信息处理的要求也越来越广，但由于汉字字符集庞大，汉字结构复杂，字形（体）繁多，汉字信息处理一直是计算机应用领域的一大难题，也是计算机科学的一个“古老”分支，它“古老” 但不腐朽，而且还不断焕发出青春的活力。汉字信息处理涉及到数据库、知识库、人工智能、模式识别、Internet 网络技术及中国语言文字知识等领域。

纵观汉语与英语的区别，英语处理的便利在于所有英语单词都可以由 26 个英文字母按前后关系拼成，而汉字是非字母化、非拼音化的文字，很难找到一种方法用一些类似于英文字母的部件来表达。围绕这一问题，人们提出了各种解决方案。很多学者就汉字的表达提出了很多有益的思想，都企图将汉字用数学方法表达出来。他们提出了汉字的有向图表示、汉字的属性关系图、汉字的相关属性关系图、汉字的二维扩展文法属性、汉字的层次模型等。张昕中、夏莹等人还明确提出了汉字表达式这一概念。但这些有关汉字的表达方法的数据结构都比较复杂，参与运算的数据及描述数据间的关系也很复杂，而且大部分方法不能唯一地表达国标一、二级汉字库中的汉字，因此它们绝大部分没有很好地用于除汉字识别以外的汉字信息处理领域。

在本书第三章中，我们提出了一种全新的将汉字表示成简单的数学表达式的思想，将汉字表示成数学表达式以后，对汉字的处理方法就可接近对英文的处理方法，从而使汉字信息处理的很多方面变得比以前简单。

汉字笔画抽取是研究汉字信息处理的重要课题之一，已有许多研究者做了大量的卓有成效的工作。但因为汉字结构复杂，不同字体间笔画变化较大，不管用哪一类方法，要准确抽取汉字笔画十分艰难。纵观前人的工作，并考虑到人对汉字的认识过程，我们认为只有充分考虑汉字的结构，引入更多的汉字结构知识，才能更好地解决汉字处理中存在的问题。本书第四章提出了一种全部量化了的完全基于汉字结构知识的直接抽取笔画方法。和已有方法相比，它的主要特点是更深入、更细致地利用汉字结构知识来解决已有方法中不能很好解决的难点，故抽取更有效，准确率更高。

互联网的出现，在全世界范围内的信息交流与共享可通过网络在瞬间完成，为展现中华民族灿烂文化和近代文明提供了起码的条件。但因为各国语言文字不同，文字编码不同，互联网上跨国家跨地区的信息交流也存在不少问题。本书第六章提出了实现互联网上跨平台传播汉字信息的方法，使汉字可以在任何一种没有汉字系统或不同汉字系统的平台上显示出来。以汉字为载体的信息可在全世界任何站点或浏览器上畅通无阻，中国文化的教育可以通过远程教育的手段传至每个希望得到这种教育的华裔或其他民族的面前。

本书第五章提出了自动生成汉字骨架的有关知识规则。利用这些规则，不仅可形成国标一二级汉字的所有汉字的骨架，而且可形成国标一、二级汉字库之外的不太常用的或虚拟的汉字骨架，为自动生成各式各样的汉字字形（体）奠定了基础。

本书第七章提出了一些有用的有关汉字结构的统计知识，并提供了一种利用汉字的数学表达式自动统计汉字结构知识的新的统计方法。

除了书中所讨论的本结构化思想在上述三个方面的应用，它还可以广泛应用于排版印刷业、广告业、虚拟图书馆、网络传输、中文移动通讯等领域。

本书是湖南省科技厅、湖南省自然科学基金委员会资助的两个课题的部分研究成果。感谢湖南省科技厅、湖南省自然科学基金委员会对课题“互连网上跨平台汉字信息传输的体系结构与实现技术（湘科计[2000]87 号）”及“汉字的数学表达方法及其应用研究（00JJY2082）”给予的资助。

在本书的写作过程中，得到了复旦大学计算机系施伯乐教授的耐心指导，他多次认真听取了作者的详细汇报，并提出了许多宝贵的意见和建议。我的同事阳爱民、李长云、江力，及我的学生吴锋、陈长协、王小乐等提出了不少有益的思想并协助编写了实现本书部分思想的源代码。值本书完成之际，作者谨向他们致以衷心的感谢！

作者在本书中提出了一些“新”的思想、方法，但这些思想还不够完善，衷心希望有更多的同行与其他各界读者多给我们提出宝贵的意见和建议，帮助我们进一步完善本书。

目　录

第一章　汉字信息处理研究简介

第一节　汉字信息处理研究的发展历史

汉字是世界上唯一独有的古代象形体，几千年来在我国盛行不衰，不因社会体制和地区语言变化，一直是我国统一的文字，是我国人民最通用记载的工具，也能适应我国不同时代文风和词句的变革，记录了我国丰富多彩的文学创作和诸子百家的学术思想，对保留中华文化遗产做出了宝贵贡献。今天使用和学习汉字的人口越来越多，地区越来越广。据统计，目前世界上约有四分之一的人口使用汉字，汉语也是联合国五种通用文字之一，因此可以说，汉字是一种生命力最强，使用率最高，适用性最广的文字，并将继续发挥重要的、其他文字形式难以取代的作用。

随着计算机应用的迅猛发展，各行业对计算机汉字信息处理的要求也越来越广，但由于汉字是非字母化、非拼音化的表意文字，汉字字符集庞大，结构复杂，字形（体）繁多，汉字信息处理一直是计算机应用领域的一大难题，已成为计算机能否真正在我国得到普及应用的一个重要瓶颈。汉字信息处理研究涉及到数据库、知识库、人工智能、模式识别、Internet 网络技术及中国语言文字知识等领域。

中文信息处理技术在我国起步较晚，在 1973 年才有实用的汉字编码出现[1支 1986]。此后，中文信息处理技术开始蓬勃发展。在 20 世纪 70～80 年代，国内曾出现汉字输入方法研究千军万“码”的局面，上千种输入方法应运而生。在汉字字型方面，从 16×16 点阵到 256×256 点阵，宋体、仿宋、楷体、黑体等各种字体不断涌现，以 CCDOS 为代表的 20 余种汉化 DOS 不断出台，各具特色，联想汉卡、巨人汉卡、四通汉字打印机等曾风靡一时。从 20 世纪 90 年代初开始，中文信息处理技术开始进入比较成熟的阶段。国家相应出台了一系列有关中文信息处理方面的标准，如 GB2312-80、GB-5007 等 30 余项汉字信息交换码及汉字点阵字型标准，以及 GB13000.1、GB18030 大字符集和开放系统平台标准等。汉字输入法也在经历了大浪淘沙之后趋于集中。中文信息处理多项技术实现了有机合理的结合，如软硬件技术的结合、输入输出技术的结合、多领域成果的结合。中文信息处理解决了在大规模应用、大规模生产以及市场营销中出现的问题，如规范性、可靠性、可维护性、界面友好性及各环节的包装。经过 20 多年的努力，我国在中文信息处理方面已取得了十分可喜的成绩，在很多方面的研究已处于世界领先。

第二节 汉字信息处理研究相关问题简介

一般说来，汉字信息处理研究包括：汉字键盘输入，汉字光学识别，汉语语音识别，汉字编码，汉字库设计（或汉字输出），机器翻译等课题。本章我们将对汉字键盘输入，汉语语音识别，机器翻译等领域作一简单介绍。对其他领域将分章在接下来的几章介绍。

1.2.1 汉字键盘输入

计算机中文信息处理技术需要解决的首要问题就是汉字的输入技术。键盘输入是一种主要的汉字输入途径。它是运用某种编码方案、键盘设备及计算机资源，由操作者向计算机输入汉字的方法。键盘输入分为：音码、形码、音形码、形音码、序号码、小键盘数字码输入等类。通常要敲击 1~4 个键输入一个汉字。

音码输入是按照汉字的读音进行汉字编码及输入的方法，用的是汉语拼音的全拼或简拼的方式。

形码输入是按照汉字的字形进行汉字编码及输入的方法。利用汉字书写的基本顺序将汉字拆分成若干块，对每一块用一个字母进行取码，整个汉字所得的码序列就是这个汉字的形码。

音形码输入是利用音码和形码各自的优点，兼顾了汉字的音和形，以音为主，以形为辅，目的是减少编码中死记的部分，提高输入效率，易学易记。

形音码输入是利用形码和音码各自的优点，兼顾了汉字的形和音，以形为主，以音为辅，目的是利用“形托（象形）”和“音托（反切）”来减少编码中死记的部分，提高输入效率，易学易记，输入快。

序号码输入是利用汉字的国标码作为输入码，用四个数字输入一个汉字或符号。

数字输入方法是以数字为基础的计算机汉字输入方法，在我国早有研究。近年来，随着信息技术的飞速发展，计算机及其相关电子信息产品也在向数字化、小型化方向发展。因为移动电话、手持电脑、全中文媒体电话、信息终端、信息家电等仅有 10 个数字键的电子设备对汉字的输入与处理的要求不断增强，这方面的研究又成为热点。目前，数字小键盘的编码主要源于“音、形、义”三个方面。由于输入键位大大减少，必然带来码长和重码率增加，而要降低这两点，研究的难度则陡然加大。清华大学马少平、夏莹等人引入人工智能技术，发明了拼音数字码与结构数字码智能输入方法[马 1999]，是在数字输入方法方面的有益尝试。

1.2.2 汉字光学识别

汉字光学识别又分为联机手写输入识别和脱机手写输入识别两种。

联机手写输入（**On-Line Character Recognition，OLCR**），是近年来发明的一种新技术[HU 2000] [CHA1999]，手写输入系统一般由硬件和软件两部分构成，硬件部分主要包括电子手写笔和写字板，软件部分是汉字识别系统。使用者只需用与主机相连的书写笔把汉字写在书写板上，写

字板中内置的高精密的电子信号采集系统，就会将汉字笔迹的信息转换为数字信息，然后传送给识别系统进行汉字识别。利用软件读取书写板上的信息，分析笔划特征，在识别字库中找到这个字，再把识别的汉字显示在编辑区中，通过“发送”功能将编辑区的文字传到其他文档编辑软件中。汉字识别系统的作用是将硬件部分传送来的信息与事先存储好的大量汉字特征信息相比较，从而判断写的是什么汉字，并通过汉字系统在计算机的屏幕上显示出来。这种输入法的好处是只要会写汉字就能输入，不需要记忆汉字的输入码，与日常写字一样，但受识别技术的限制，速度一般。手写输入系统的难点在于汉字笔迹的识别，因为每一个人的书写汉字笔迹都不一样，因此手写笔迹比较系统就必须能允许一定的模糊偏差，才能有较高的识别率。目前已经开发了许多种手写输入系统，简称为“手写笔”系统。有些手写笔可以代替鼠标进行操作。

汉字脱机手写输入识别，又称为汉字 OCR（Optical Character Recognition），它是利用计算机的外部设备——光电扫描仪，首先将印刷体或手写体的文本扫描成图像，再将整页版面的原始图像先按书写行分割开，然后再从每行中切分出单个汉字图像，然后通过专用的汉字 OCR 系统进行文字识别，将汉字的图像转成文本形式，最后用“文件发送”或“导出”输出到其他文档编辑软件中。汉字 OCR 系统进行的预处理通常包括大小归一、平滑、细化或轮廓化等处理过程。特征抽取与分类器的设计是整个系统中最为重要的环节[CAI1999] [CHE1998] [GOV1990] [HIL1993] [HSI1992] [HS12000] [KAT1999] [LEE1996] [LI2000] [LIN1998] [LIO1999] [LUA1990] [MOR1984] [O1999] [PAR1998] [ROC1994] [ROM1997] [SUE1999] [TANG1998] [TSE1998] [WON1998] [XU1999]，稳定特征的抽取与良好性能的分类器的设计是整个识别系统的核心，它们直接决定了识别系统的性能。文本识别后处理是指对单字识别的结果，利用词义、语义等上下文先验信息进行识别结果的确认或纠错。汉字 OCR，按特征抽取与分类器的设计不同，通常可以分为结构模式识别方法、统计模式识别方法、统计与结构相结合的识别方法以及人工神经网络方法等。这种输入方法的特点是快速、易操作，但要求文稿清晰。它在电子图书馆，古籍数字化等诸多方面有重大作用。

我国在联机手写汉字识别和脱机手写汉字识别方面的研究特别活跃，已有不少产品投入市场。中自汉王公司，清华大学电子工程系，清华大学计算机系，中科院自动化 AI 实验室，北京信息工程学院，北京邮电大学，北京大学计算所，武汉工业大学，美国摩托罗拉公司，台湾蒙恬公司等都有实用的系统通过国家测评[刘1999]。但因为其牵扯的因素太多：书写者的习惯或文件印刷品质、扫描仪的扫描品质、识别的方法、学习及测试的样本等，识别正确率就像是一个无穷趋近函数，知道其趋近值，却只能靠近而无法达到，永远在与 100%作拉锯战。

1．2．3　汉语语音识别

语音输入也是近年来出现的一种新技术，它的主要功能是用与主机相连的话筒读出汉字的语音，利用语音识别系统分析辨识汉字或词组，把识别后的汉字显示在编辑区中，再通过“发送”功能将编辑区的文字传到其他文档的编辑软件中。语音识别技术的原理是将人的话音转换成声音信号，经过特殊处理，与计算机中已存储的已有声音信号进行比较，然后反馈出识别的结果。这项技术的关键在于将人的话音转换成声音信号的准确性，以及与已有声音信息比较时

的智能化程度。语音识别技术是人工智能的有机组成部分。这种输入的好处是不再用手去输入，只要会读出汉字的读音即可，但是受每个人汉字发音的限制，不可能都满足语音识别软件的要求，因此在实际应用中错误率较键盘输入高。特别是一些专业技术方面的语言，识别系统几乎不能确认，错误率较高。

1．2．4　汉字编码

相对西文字符集的定义，汉字编码字符集的定义有两大困难：选字难和排序难。选字难是因为汉字字量大(包括简体字、繁体字、日本汉字、韩国汉字)，而字符集空间有限。排序难是因为汉字可以多种排序标准(拼音、部首、笔画等等)排序，而具体到每一种排序标准，往往还存在不少争议。

1981 年 5 月我国发布实施了国家标准 GB 2312-80《信息交换用汉字编码字符集.基本集》，它奠定了我国中文信息处理技术发展的基础。

1994 年 9 月第二辅助集、第四辅助集定稿并经过审定。第一辅助集、第三辅助集、第五辅助集分别是基本集、第二辅助集、第四辅助集的繁体字映射集，且除个别简/繁关系为一对多的汉字外，简/繁字在两个字符集中同码。

这六个集均采用双七位编码方式，但为了避开 ASCII 表中的控制码，每个七位只选取了 94 个编码位置。所以每张代码表分 94 个区和 94 个位。其中前 15 区作为拼音文字及符号区或保留未用，16 区到 94 区为汉字区。

基本集收入汉字信息交换用的基本图形字符，包括一般符号，序号，数字，拉丁字母，日文假名，希腊字母，俄文字母，汉语拼音符号，汉语注音字母及 6 763 个简化汉字，总计 7 445 个图形字符，采用一字一码的原则。

第二辅助集是国家标准 GB 7589-87，收集了通用规范的简体汉字 7 237 个，以部首为序排列，部首次序按笔画数排列，同部首字按部首以外的笔画数排列，同笔画数的字以笔形顺序(横、直、撇、点、折)为序。

第四辅助集是国家标准 GB 7590-87，收集了通用规范的简体汉字 7 039 个，以部首为序排列，部首次序按笔画数排列，同部首字按部首以外的笔画数排列，同笔画数的字以笔形顺序(横、直、撇、点、折)为序。

第一辅助集(GB 12345-90)已于 1990 年发布，是与基本集对应的繁体字集，共收图形字符 7 583 个，其中前 15 区除收集了 GB 2312 中前 15 区内收的全部字符外，又增收了 35 个竖排标点符号和汉语拼音符号。从 16 区至 91 区共收 6 866 个繁体汉字。一级汉字数和二级汉字数都与 GB2312 相同，另有 103 个繁体字是属于简/繁为一对多的字。对于简/繁一对多的情况，则选一个最通用的繁体字码置于与基本集中该字相对应的码位，其余的则按拼音序编码于 88 和 89 区。

在文字工作、档案管理、银行或者中医药工作中，在用电脑输入中文时，往往会遇到敲不出来的汉字。为了改进这一状况，2000 年 3 月 17 日，国家质量技术监督局和信息产业部联合

发布了两项中文信息处理技术标准，不仅对原来的GB2312的字符集做了扩充，对数字键盘上汉字输入拼音和笔画的布局也作出了规定。在新颁布的两项国家标准中，GB18030《信息交换汉字编码字符集》是对原有GB2312字符集的扩充，收录了27484个汉字，总编码空间超过150万个码位，为彻底解决邮政、户政、金融、地理信息系统等迫切需要的人名、地名用字提供了解决方案，也为中文信息在国际互联网上的传输与交换提供了保障。

台湾、香港等地使用的汉字是繁体字，台湾已经定义的汉字字符集只收繁体字。在台湾，用于中文信息交换的标准有CCCII（中文资讯交换码），CNS 11643（通用汉字标准交换码）等。

日本和韩国使用的汉字大都与中国相同或相似。日本JISX 0208-1983与GB 2312相似，共收字符6 877个，1到15区为拼音字符及符号区，16到84区为汉字区，共收日本汉字6 353个，分一级汉字区和二级汉字区，一级汉字区按拼音排序，二级汉字区按部首排序。日本JISX 0212-1990为日本汉字交换码辅助集，共收6 067个字符，日本汉字有5 801个，按部首排列。韩国KSC 5601-1987共收8 244个字符，其中韩国汉字有4 888个。

不论何种编码，都是指定一个数字来表示汉字或其他字符。随着编码系统的不同，两种编码可能使用相同的数字代表两个不同的字符，或使用不同的数字代表相同的字符，这样这些编码系统就会互相冲突。为了解决冲突，统一不同国家和地区的字符标准，国际上提出了**Unicode**标准[UNI 2001]。**Unicode**是一种重要的交互和显示的通用字符编码标准，它覆盖了美国、欧洲、中东、非洲、印度、亚洲和太平洋的语言，以及古文和专业符号。**Unicode**允许交换、处理和显示多语言文本以及公用的专业和数学符号。**Unicode**字符可以适用于所有已知的编码。**Unicode**是继**ASCII**（美国国家交互信息标准编码）字符码后的一种新字符编码，它为每一个符号定义一个数字和名称，并指定字符和它的数值（码位），以及该值的二进制位表示法，通过一个十六进制数字和前缀（U）定义一个**16**位的数值，如**U+0041**表示**A**，其唯一的名称是**LATIN CAPITAL LETTER A**。**Unicode**兼容于**ASCII**字符并被大多数程序所支持，**Unicode**完全兼容于国际标准**ISO/IEC 10646-1**（**1993**），它是**ISO 10646**的一个子集。

1．2．5　汉字库

汉字字形库按描述技术的不同，可分为点阵字库、矢量字库和曲线轮廓字库[傅 1999]。

点阵字模，就是在若干条等距离垂直线和水平线交叉形成的栅格内将汉字的笔画用点的形式描出；然后利用计算机辅助设计的方法，在一台通用汉字终端屏幕上由造字软件先画出放大的栅格，通过键盘严格按事前设计的点阵字模一笔一画地在栅格内打点、画线。计算机及时地把这些点、线转换成数据存入。点阵字库的缺点是数据存储量大且字形不易放大和变化，字形放大后有明显的锯齿感或折线感。

矢量字库用直线段序列描述字形轮廓。这种方法先将字稿拓到96×96的方格纸上，然后人工描出矢量结点，最后将数据录入计算机。现在一般都采用扫描仪将字形数字化，再由程序自动生成矢量轮廓，其间还需一定的人机交互。一般说来，矢量字库已能满足高质量输出的要求且具有存储量小的优点，但当字形放大倍数很大时，矢量法所产生的字形仍会显出折线感。到

1993 年，矢量字库已比较多地用于高档汉卡和激光打印机，并已有宋、仿宋、楷、黑四种字体的国家标准。

曲线轮廓法是一种最新的字形描述技术。这种技术将字形看作是一种图形，采用特定的数学曲线描述字形。目前有代表性的曲线轮廓技术有两种：一种是美国 Adobe 公司的 Post Script 字形技术，另一种是 Apple 公司的 TrueType 技术。

1．2．6　机器翻译

随着网络应用的日益深入，人们对机器翻译的需求也日益增强。目前，Internet 上的信息绝大多数是英文，中文数据不到 1%，对各种外文资料实现实时翻译是人们所期盼的。然而，这些对机器翻译来说，除了在算法上要有新的突破外，对汉语词库、语料库、语义理解等方面都有一系列新的要求。对英汉机器翻译来说，需进一步攻克一词多义、结构歧义、语义歧义等难题。汉英机器翻译的难度则更大，这主要是由于汉语词类无形态变化（如单复数、时态、语态等），同一词类担任多种语法成分，使汉语到英语的转换和英语的生成实际是单词膨胀、信息增加的过程。对于其他语言翻译也存在各自不同的难点，针对一些如旅游的专门应用，搞一些受限范围的自动翻译软件和硬件产品是当前的热点。

第二章 汉字光学识别

第一节 汉字识别简介

几千年来，汉字在我国盛行不衰，不因社会体制和地区语言变化仍然一直是我国统一的文字，始终是我国人民最通用记载的工具，也能适应我国不同时代文风和词句的变革，记录了我国丰富多采的文学创作和诸子百家的学术思想，对保留中华文化遗产做出了宝贵贡献。然而，汉字是非字母化、非拼音化的文字，在当今高度信息化的社会里，如何快速高效地将汉字输入计算机，已成为影响人机接口效率的一个重要瓶颈，也关系到计算机能否真正在我国得到普及应用。

目前，汉字输入主要分为人工键盘输入和机器自动识别输入两种。人工键盘输入是指用手工击键方式按照一定的规律把汉字输入到计算机，目前已有数百种输入方案。但是，与拼音文字的打字机不同，人们需要经过一定时间的学习训练才能掌握某种键入方法，更为严重的是：对于大量已有的文档资料，采用人工键入方法将要花费大量的人力和时间。为此，机器自动识别输入就成了必须研究的课题。

自动识别输入分为语音识别和字符识别（**OCR**，**Optical Character Recognition**，光学字符识别）两种。汉字识别是模式识别的一个重要分支，也是文字识别领域最为困难的问题，它涉及模式识别、图像处理、数字信号处理、自然语言理解、人工智能、中文信息处理等学科，是一门综合性技术，在中文信息处理、办公自动化、机器翻译、人工智能等高技术领域，都有着重要的实用价值和理论意义。

早在 **20** 世纪 **60~70** 年代，世界各国就开始有 **OCR** 的研究，而研究的初期，多以文字的识别方法研究为主，且识别的文字仅为 **0** 至 **9** 的数字。以同样拥有方块文字的日本为例，**1960** 年左右开始研究 **OCR** 的基本识别理论，初期以数字为对象，直至 **1965** 至 **1970** 年之间才开始有一些简单的产品，如印刷文字的邮政编码识别系统，识别邮件上的邮政编码，帮助邮局作区域分信的作业。自从 IBM 公司的 Casey 和 Nagy 于 1966 年首次发表关于汉字识别的文章以来，汉字识别取得了长足的进展，提出了很多理论和方法。

汉字识别技术可分为印刷体汉字识别和手写体汉字识别两大类，后者又可分为联机(on-line)手写汉字识别和脱机(off-line)手写汉字识别。从识别的角度来看，手写体识别难于印刷体识别，脱机手写识别难于联机手写识别。在脱机手写汉字识别领域，非特定人脱机手写汉字识别又难于特定人手写汉字识别。我国已有印刷体汉字识别和联机手写汉字识别的商品出售，目前已形成百家争鸣、百花齐放的局面。

第二节　脱机汉字识别方法分类

汉字脱机手写输入识别，又称为汉字 OCR（Optical Character Recognition），它是利用计算机的外部设备——光电扫描仪，首先将印刷体或手写体的文本扫描成图像，再将整页版面的原始图像先按书写行分割开，然后再从每行中切分出单个汉字图像，通过专用的汉字 OCR 系统进行文字识别，将汉字的图像转成文本形式，最后用“文件发送”或“导出”输出到其他文档编辑软件中。汉字 OCR 系统进行的预处理通常包括大小归一、平滑、细化或轮廓化等处理过程。特征抽取与分类器的设计是整个系统中最为重要的环节[CAI1999][CHI1998][GOV1990][HIH 1993][HST1992][HST2000][KAI1999][LEE1996][LI 2000][LIN1998][LIO1999][LUA1990][MOR1984][OI1999][PAR1998][ROC1994][ROM1997][SUL1999][TANG1998][YSE1998][WON1998][XU1999]，稳定特征的抽取与良好性能的分类器的设计是整个识别系统的核心，它们直接决定了识别系统的性能。文本识别后处理是指对单字识别的结果，利用词义、语义等上下文先验信息进行识别结果的确认或纠错。

汉字的模式表达形式和相应的字典形成方法有多种，每种形式又可以选择不同的特征或基元（Primitive），每种特征或基元又有不同的抽取方法，这就使得判别方法和准则以及所用的数学工具不同，形成了种类繁多、形式各别的汉字识别方法。总的来说，不同的特征抽取和分类器的设计方法决定了识别系统采用不同的处理方法。汉字 OCR，按特征抽取与分类器的设计不同，通常可以分为结构模式识别方法、统计模式识别方法、统计与结构相结合的识别方法以及人工神经网络方法等。这种输入方法的特点是快速、易操作，但要求文稿清晰。它在电子图书馆，古籍数字化等诸多方面有重大作用。

我国在联机手写汉字识别和脱机手写汉字识别方面的研究特别活跃，已有不少产品投入市场。中自汉王公司，清华大学电子工程系，清华大学计算机系，中科院自动化 AI 实验室，北京信息工程学院，北京邮电大学，北京大学计算所，武汉工业大学，美国摩托罗拉公司，台湾蒙恬公司等都有实用的系统通过国家测评[刘1999]。但因为其牵扯的因素太多：书写者的习惯、文件印刷质量、扫描仪的扫描质量、识别方法、学习及测试的样本等，识别正确率就像是一个无穷趋近函数，知道其趋近值，却只能靠近而无法达到，永远在与 100%作拉锯战。

2.2.1　结构模式识别方法

结构模式识别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画作为基元，由这些基元再构成部件（子模式），由部件的组合来描述汉字（模式），最后再利用形式语言及自动机理论进行文法推断，即识别。然而，人们美好的初衷并未能如愿以偿，这是因为从汉字图像中抽取笔画等基元比较困难。通常，为了抽取笔画需要将原始点阵图象进行细化处理，但是细化算法不仅速度慢，而且容易产生畸变，如将一个四叉点变成了二个三叉点，给准确抽取基元造成了困难。为了解决这个问题，有些学者试图不经过细化直接从汉

字点阵图象中抽取笔画等基元，但效果仍不尽如人意。因此，有些研究人员放弃了抽取笔画或笔段作为基元然后进行文法推断的思路，采用汉字轮廓结构信息作为特征，这一方案的识别结果优于基于基元抽取的方法，但识别方法需要进行松弛迭代匹配，耗时严重，而且对于笔画较模糊的汉字图象，抽取内轮廓会遇到极大困难，外轮廓的抽取也不太稳定。也有些学者采用抽取汉字图象中关键特征点来描述汉字，汉字的关键特征点包括端点、折点、交点、歧点、背景特征点、局部曲率最大点等，但是特征点的抽取易受噪声点、笔画的粘连与断裂等影响。

总之，早期的脱机手写汉字识别研究者将精力主要集中在如何准确地抽取基元、轮廓、特征点等能够反映汉字结构信息的特征上，并且在假设这些特征已经比较准确地抽取完毕的前提下，研究文法匹配、属性图匹配、松弛迭代匹配等。然而，单纯采用结构模式识别方法的脱机手写汉字识别系统，识别率较低，这就促使人们将目光转向了统计模式识别方法。

2.2.2 统计模式识别方法

统计模式识别方法对每一个汉字提取统计学上的某些特征，每一个特征表示为多维向量的一个分量。如果总共抽取 n 种特征，则每一个汉字可以用一个 n 维特征向量来描述。如果总共有 m 个汉字，它们的特征向量就构成了一个 n×m 的矩阵(假定每一个特征向量是一个列向量)。这个矩阵称为识别字典，存放在计算机中。当一个待识别的未知汉字 x 输入计算机后，识别系统抽取 x 的 n 个特征，构成一个 n 维向量。计算这个 n 维向量与识别字典中的每个向量之间的距离(或类似度，相似度)，找出距离最小(或类似度，相似度最大)的向量所对应的汉字作为识别结果。常用的距离度量有欧氏距离、城市块距离、马氏距离等。常用的求类似度的方法是计算两个向量之间夹角的余弦。当两个向量重合时得到最大的类似度 1。这种计算类似度的方法的缺点是所定义的类似度只与向量之间的夹角有关而不能反映两个向量在长度上的差别。

统计学上的特征一般不考虑汉字字形在结构上的特点，而把用点阵表示的汉字图像看作一个随机的二维图形。例如，把汉字点阵划分为 3×3 个小方块，统计每个小方块中的黑点的个数，得到一个 9 维特征向量。这是一种典型的统计特征。

与结构法相比，统计法具有良好的抗噪声、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。

用于脱机手写汉字识别的统计特征，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：

(1)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 Fourier 变换、Hadamard 变换、DCT 变换、Walsh 变换、Rapid 变换、K-L 变换等；

(2)不变矩（Moment）特征；

(3)笔画穿透数目特征；

(4)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；

(5)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，通常可选取位于汉字图像两对角线上的背景点。

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：

(1)局部笔画方向特征；

(2)细胞特征；

(3)相补特征；

(4)方向线素特征；

(5)Gabor 特征；

(6)四角特征。

根据抽取特征的不同，可以选用不同的匹配方法，常用的统计匹配方法有模板匹配、相关匹配、树分类器等。

2.2.3 统计与结构相结合的识别方法

统计模式识别方法对断笔、连笔、污点等不敏感，因而具有较强的抗噪音干扰能力。

对于手写体汉字识别来说，句法模式识别是一种很诱人的方法，因为它充分利用了汉字结构上的大量信息。但主要的问题是由于作为基元的笔画对噪音很敏感，因而笔画的提取十分困难。

统计模式识别方法对断笔、连笔、污点等不敏感，因而具有良好的鲁棒性，较好的抗干扰抗噪声的能力，它一般按一定的距离度量匹配准则，采用多维特征值累加的办法，把局部噪声和微小畸变淹没在最后的累加和里，但由于它忽略了汉字结构上的大量有用信息，可以用来区分“敏感部位”的差异也随之消失，因此区分相似字的能力较差，难以得到很好的识别效果。而结构方法对结构特征较敏感，区分相似字的能力较强，但是结构特征难以抽取，不稳定。因此，人们已注意到将两种方法结合起来使用，这种结合包括两个方面：

(1)特征的结合：在特征抽取过程中，注意抽取能反映手写汉字结构信息的统计特征，如方向线素特征、笔画穿透数目特征等。

(2)识别方法的结合：可以先用统计方法进行粗分类，再用结构方法进行细分类来区分相似字；也可以将两种方法并联使用，然后进行综合集成。

2.2.4 人工神经网络方法

人工神经网络的主要特征是：大规模的并行处理和分布式信息存储，良好的自适应性、自组织性，以及很强的学习功能、联想功能和容错功能。目前的研究重点是将人工神经网络原理应用于图像处理、模式识别、语音综合及智能机器人控制等领域。

人工神经网络方法在文字识别方面主要用于：

(1)学习训练；

(2)分类器设计；

(3)特征抽取与选择；

(4)单字识别后处理。

用人工神经网络方法进行汉字识别的最大优点是具有自学习功能。

目前常用的做法是将神经网络方法和传统的识别方法结合起来使用，互相取长补短，如先用传统的方法抽取较为稳定的特征，然后再用神经网络进行自组织聚类学习并设计性能良好的分类器等。通常，用于文字识别的人工神经网络模型有：Hopfield 神经网络、前向多层神经网络（如 BP 算法、RBF 网络等）、 ART 网络、自组织特征映射网络、认知器模型等。

第三节 脱机汉字识别的预处理

预处理是指对光学设备输出的汉字图像作一些必要的加工，为以后的特征提取及识别作准备，一般包含以下内容：

2.3.1 二值化

将光学设备输出的汉字图像转换成无灰度的黑白二值图像，也就是说要把文字图像转换成只取 0，1 两种值的图形，有笔画的地方用 1 表示，无笔画的地方(即文字的背景)用 0 表示。

2.3.2 文字切割

对于输入计算机的整页文字图像，需要把它分割成单个的汉字进行识别。因文稿可能是没有任何格线的白纸，字与字之间的横向距离往往不稳定，有时两个字挨得很近，有时一个字的左右偏旁拉得太开，因而常常容易把一个左右偏旁的字切成两个字，或者把某个字的右偏旁和右边字的左偏旁合在一起切成一个字。

2.3.3 平滑与消除噪音

平滑是指去掉笔画轮廓线上的毛刺，毛刺也是一种噪音。除了毛刺以外，手写汉字的噪音主要有以下几种：

(1)质量不好的纸张中的杂质所造成的文字背景上出现的污点；

(2)由于颜色深浅不匀，在颜色较浅的地方可能造成笔画的断裂或在笔画中间出现白色的空洞；

(3)墨水污染纸张造成背景上出现污点；

(4)设备质量或二值化算法产生的噪音。

2.3.4 细化