

中文信息处理丛书

汉字键盘输入技术与理论基础

陈一凡 胡宣华 著



清华大学出版社
广西科学技术出版社

中文信息处理丛书

汉字键盘输入技术与理论基础

陈一凡 胡宣华 著

清华大学出版社
广西科学技术出版社

(京) 新登字 158 号

内 容 简 介

本书全面论述了汉字键盘输入技术的原理、方法和系统。分别讨论了汉字特征信息、特征信息编码和非编码键盘输入原理、系统设计方法、系统质量评估和选用原则,并介绍了海内外各具特色的 32 种实用系统。本书内容融理论、实用、资料于一体,全书标引参考文献 150 篇,书后附名词术语中英文索引。

这是中国第一本汉字键盘输入技术的系统性学术著作。可作为有关专业研究生、大学高年级学生的辅助教材或参考书,也可供从事计算机应用和中文信息处理技术的研究人员和技术人员参考。广大关心汉字键盘输入的计算机使用者也会从本书中得到益处。本书对国内外研究汉字键盘输入技术有促进作用。

版权所有,翻印必究。

本书封面有清华大学出版社激光防伪标志,无标志者不得销售

汉字键盘输入技术与理论基础

陈一凡 胡宣华 著

☆

清华大学出版社出版

北京 清华园

广西科学技术出版社出版

南宁市河堤路 14 号

通县宏飞印刷厂印刷

新华书店总店科技发行所发行

☆

开本: 787×1092 1/16 印张: 16 字数: 380 千字

1994 年 6 月第 1 版 1994 年 6 月第 1 次印刷

印数: 00001—20000

ISBN 7-302-01448-5/TP·568

定价: 12.00 元

清华大学出版社 广西科学技术出版社
计算机学术著作出版基金

评审委员会

主任委员	张效祥
副主任委员	周远清 汪成为
委员	(按姓氏笔画排列)
	王鼎兴 李三立
	杨芙清 张兴强
	施伯乐 徐培忠
	徐家福 夏培肃
	董韫美

出版说明

近年来,随着微电子和计算机技术渗透到各个技术领域,人类正在步入一个技术迅猛发展的新时期。这个新时期的主要标志是计算机和信息处理的广泛应用。计算机在改造传统产业,实现管理自动化,促进新兴产业的发展等方面都起着重要作用,它在现代化建设中的战略地位愈来愈明显。计算机科学与其它学科的交叉又产生了许多新学科,推动着科学技术向更广阔的领域发展,正在对人类社会产生深远的影响。

科学技术是第一生产力。计算机科学技术是我国高科技领域的一个重要方面。为了推动我国计算机科学及产业的发展,促进学术交流,使科研成果尽快转化为生产力,清华大学出版社与广西科学技术出版社联合设立了“计算机学术著作基金”,旨在支持和鼓励科技人员,撰写高水平的学术著作,以反映和推广我国在这一领域的最新成果。

计算机学术著作出版基金资助出版的著作范围包括:有重要理论价值或重要应用价值的学术专著;计算机学科前沿探索的论著;推动计算机技术及产业发展的专著;与计算机有关的交叉学科的论著;有较大应用价值的工具书;世界名著的优秀翻译作品。凡经作者本人申请,计算机学术著作出版基金评审委员会评审通过的著作,将由该基金资助出版,出版社将努力做好出版工作。

基金还支持两社列选的国家高科技重点图书和国家教委重点图书规划中计算机学科领域的学术著作的出版。为了做好选题工作,出版社特邀请“中国计算机学会”、“中国中文信息学会”帮助做好组织有关学术著作丛书的列选工作。

热诚希望得到广大计算机界同仁的支持和帮助。

清华大学出版社 计算机学术著作出版基金办公室
广西科学技术出版社

1992年4月

中文信息处理丛书编委会

主任委员	陈力为
副主任委员	许孔时
委 员	(按姓氏笔画排列)
	王 选 刘 源
	何克抗 吴文虎
	苏东庄 张 普
	俞士汶 袁 琦
	徐培忠 曹右琦
	黄昌宁

中文信息处理丛书

序 言

中文信息处理技术在我国现代化及信息化建设中,越来越起着重要的作用,作为一个高新技术的重点,它已经列入国务院批准的“国家中长期科学技术发展纲领”。十几年来,我国的中文信息处理领域里,在技术的研究、产品的开发、以及产业的建立等方面都取得了显著的成绩。现在很需要把这些方面的成果加以综合并且提炼出来,以便推广应用,并且作为一个起点,再上一个新台阶。这就是我们组织编写并出版这套中文信息处理丛书的目的。

在这套丛书即将开始出版之机,我愿向读者介绍以下两点:

第一 为什么我们要把中文信息处理技术作为高新技术的一个重点来发展呢?

我们日常工作中的信息,绝大部分是以语言文字作为媒介,传播交换和记载的。因此随着计算机的推广应用,由数据处理,信息处理发展到知识处理,对语言文字的处理的要求的深度和广度,越来越高。这个问题在西方国家并不突出。因为计算机在诞生之日开始,就是以处理西方语言为基础的。换言之,他们无须经过呼吁和宣传,很自然地随着计算机的推广应用的发展,都会主动地研究和解决自己国家使用计算机如何不断地适应自己国家的语言文字问题。可惜,我们的汉语与西方语言的差别很大。能够处理西方语言的计算机,面对汉语,却显得无能为力。例如:

- 西方语言为拼音文字,而汉语是表意文字。西文字符只有 20 余个,而汉语文字仅常用的就有六、七千个,总数超过五万。这是一个根本性的问题。仅这一个差异就引起了处理汉语的计算机与处理西方语言的计算机一系列的差异,需要我們自己去解决。包括键盘输入、汉字打印、显示、内部代码、汉字识别、程序语言的数据类型、数据库的检索和排序等等。

- 西方的书面语言,词与词之间有空格。而汉语的词与词之间无空格。于是词的切分问题就成了计算机处理汉语的首要问题。

- 西方语言的同音词很少,而汉语的同音词很多。例如,JI 音汉字就有一百多个。辨析同音词就成了汉语语音处理的关键。

- 西方语言多有形态变化(例如:多数、少数,过去、现在,男、女等等),而汉语缺少形态变化。计算机对汉语的处理(例如,机器翻译、人机接口等)无法利用形态,只能在语法、语义上找出路。

- 汉语的语法尚未形成规范化,而且人们习惯于非规范化的语法。于是语义的研究的重要性比西方语言重要得多。例如,“吃饭”“吃大碗”和“吃食堂”的理解只能靠语义来解决。

- 汉语的自动(计算机)处理是多学科和跨学科的研究工作,特别需要计算机与语

言学的密切结合，而且要依靠语言学的长期积累的研究成果。但我国语言学界多着重汉语教学，对象是人，而不是机器，因此对其丰硕的研究成果要经过改造、深化、量化，甚至要从头开始。要清醒地认识到它的艰巨性，要持续不懈地抓下去。

以上只是几个突出的问题。还有很多其它问题，不再赘述。这些语言上的特点造成了计算机处理汉语的很多障碍，每前进一步都会遇到新问题，使我们不得不花费自己很多力量去解决。

再就计算机的发展趋势而言，计算机产业面临转型期。多媒体和笔记本式计算机将成为热门产品。这些产品的核心技术无不与中文信息处理技术有关。因此，加强中文信息处理的研究更为必要。

第二 中文信息处理技术包括哪些科目呢？

大体上包括下列一些科目：

- 词的切分和频率统计
- 汉语句型和短语的研究及频率统计
- 汉语语义的研究
- 键盘和非键盘汉字输入技术及处理系统
- 汉语语料库的开发及应用
- 汉字的机器代码，程序设计语言的数据类型
- 汉语开放系统的接口规范
- 语声输入与合成
- 汉字识别
- 字形生成
- 汉语分析及理解
- 汉语生成
- 人机接口
- 机器翻译
- 情报检索
- 自动标引和抽词，自动文摘
- 全文检索
- 电子印刷出版系统
- 汉语辅助教学
- 电子词典

以上这些科目，有些是基础研究，有些是技术研究，也有些可以直接转化为产品。这些科目的分类并非学科分类，不过是按照编者本人日常接触的项目，把它们罗列出来而已。其分类的科学性、正确性和完整性尚待商榷。必须指出，有些基础性研究虽然看不到直接的经济效益，但它的研究成果则是其它研究工作所必需，而且要先行。

到目前为止，在上述这些项目中，有些已经产业化，例如电子印刷出版和少数几个汉字输入系统；有些项目已经商品化，正向产业化迈进；很多项目已经实用化。但每个领域都有很多问题等待我们去解决。今后的工作只能加强，不能削弱。使我们中文信息

处理的每个领域，每个项目都沿着实用化、商品化和产业化的道路奋勇前进。我相信我们这套丛书必将在促进中文信息处理技术的发展方面发挥它应有的作用。这套丛书大约十册左右，将在‘八五’期间陆续出版。

最后，感谢“计算机学术著作出版基金评审委员会”把出版中文信息处理丛书列入了‘八五’出版计划。感谢清华大学出版社和广西科学出版社给予出版基金的支持。

中国中文信息学会理事长 陈力为

1992年5月 于北京

前 言

80年代初以来,数以百计的汉字键盘输入方法打开了汉字进入计算机的通道,成为汉字输入计算机的主要手段,汉字键盘输入技术成为中文信息处理领域的重要分支。汉字键盘输入技术是一门综合性技术,它涉及到语言文字学、计算机技术、人机工程学、心理学、信息论、统计决策论、人工智能、中文信息处理、计算语言学、编码理论、离散数学等学科和技术。十多年来,国内外从字输入发展到词输入,并向基于自然语言理解的智能化句输入发展,出现了一批实用的汉字键盘输入系统。

本书第一、二章介绍汉字键盘输入的基本原理,第五、六章介绍汉字键盘输入系统的设计方法、质量评估和选用原则,第七章介绍正在发展中的新型输入方法。以上各章内容中包括国内外同行近年来的研究成果,也包括作者和作者所在的科研组的研究成果。

本书第三章介绍我国大陆科技工作者设计的21种汉字键盘输入系统,其中14种做了比较详细的介绍;第四章介绍大陆以外包括香港、台湾、美国、日本流行的11种输入系统;第七章介绍了三种新一代输入系统。这些系统各有特色,拥有一定的用户,使用情况良好。各章所介绍的系统的设计者给作者寄来了详尽的技术资料,并审阅和探讨了本书初稿的有关内容。没有他们的合作,本书是难以完成的。

本书可作为计算机应用、中文信息处理等有关专业的研究生、大学高年级学生的辅助教材或参考书,也可供从事计算机应用和中文信息处理技术的研究人员和技术人员参考。本书融理论、实用、资料于一体,对广大关心汉字键盘输入技术的计算机使用者来说,也会从书中得到益处。

本书第一、二、五、七章由北京信息工程学院陈一凡执笔,第三、四、六章由大连理工大学胡宣华执笔。

中国中文信息学会理事长陈力为教授仔细审阅了全书,李公宜、张普、何厚存教授,张侃、王晓龙博士,姚世全、华绍和、张轴材高级工程师分审各章,他们提出不少宝贵的改进意见。韩布新先生和陈民女士为本书做了大量数据统计和文字处理工作。在编著过程中得到中国中文信息学会汉字编码专业委员会和台北电脑技能基金会及国内外有关专家、同行的热情支持。在此,对以上提到的各位先生和对本书提供过帮助的所有人士一并表示感谢。

由于汉字键盘输入技术正在不断发展,目前国内尚无系统的学术专著,加上作者水平所限,因此本书从内容选择到学术观点,一定存在不少缺点和错误,敬请广大读者批评指正。

作 者

1993年10月



陈一凡教授，现任北京信息工程学院信息系统教研室主任，中国中文信息学会汉字编码专业委员会委员。长期从事计算机应用和中文信息处理等教学和研究，出版有专著5本，在国际会议和国内刊物发表论文20多篇。



胡宣华高级工程师，现任高校联合科技中心开发部主任，中国中文信息学会汉字编码专业委员会委员。长期从事计算机应用和中文信息处理等教学和研究，1980年获全国科学大会奖状，有多项专利研究。在国际会议和国内刊物发表论文10多篇。

目 录

第一章 绪论	1
1.1 汉字的字量与使用频度	1
1.1.1 汉字的字量	1
1.1.2 近代汉字的使用频度	1
1.2 汉字编码字符集的国家标准与国际标准	3
1.2.1 信息交换用汉字编码字符集—基本集	3
1.2.2 信息交换用汉字编码字符集的二、四辅助集	4
1.2.3 信息交换用汉字编码字符集的一、三、五辅助集	5
1.2.4 CJK 统一汉字编码字符集	5
1.3 汉字输入计算机的方式	7
1.3.1 汉字键盘输入	7
1.3.2 汉字字形识别输入	8
1.3.3 汉语语音识别输入	11
1.3.4 屏幕拣字输入	11
参考文献	11
第二章 汉字的特征信息与键盘输入原理	13
2.1 字形类特征信息	13
2.1.1 汉字的字形与字体	13
2.1.2 笔画特征信息	13
2.1.3 字根特征信息	16
2.1.4 汉字结构特征信息	27
2.1.5 字形类特征信息的层次结构	30
2.2 字音类特征信息	31
2.2.1 汉字读音的标注方法	31
2.2.2 汉字音节的数量和同音字	36
2.3 字义类特征信息	37
2.3.1 字义与词义	37
2.3.2 词义的形式化描述方法(1)——义素分析	39
2.3.3 词义的形式化描述方法(2)——语义场	40
2.4 词语的特征信息	43
2.4.1 词语及其使用频度	43
2.4.2 词语特征信息及提取方法	44
2.5 汉语的熵与冗余度	45

2.5.1	符号的熵值,低阶熵	46
2.5.2	汉语字、词的高阶熵	48
2.5.3	汉语的冗余度	50
2.6	汉字小键盘输入原理	51
2.6.1	汉字小键盘输入流程	51
2.6.2	基本概念	52
2.6.3	理想的汉字特征信息键盘输入的数学模型	54
2.6.4	实际的汉字特征信息键盘输入的数学模型	55
	参考文献	58
第三章	不同类型和特点的汉字键盘输入系统(上)	60
3.1	表形码	61
3.1.1	概述	61
3.1.2	表形码的特征信息集、码长及键位分布	61
3.1.3	表形码的字编码输入方法	63
3.1.4	表形码的词编码输入方法	63
3.1.5	讨论	64
3.2	层次四角码	64
3.2.1	概述	64
3.2.2	层次四角码的特征信息集、码长及键位分布	64
3.2.3	层次四角码的字编码输入方法	65
3.2.4	层次四角码的词编码输入方法	67
3.2.5	讨论	67
3.3	大众音形码	68
3.3.1	概述	68
3.3.2	大众音形码的特征信息集、码长及键位分布	68
3.3.3	大众音形码的字编码输入方法	70
3.3.4	大众音形码的词编码输入方法	70
3.3.5	讨论	70
3.4	二维三码	71
3.4.1	概述	71
3.4.2	二维三码的特征信息集、码长及键位分布	71
3.4.3	二维三码的字编码输入方法	73
3.4.4	二维三码的词编码输入方法	75
3.4.5	讨论	75
3.5	交大李码	76
3.5.1	概述	76
3.5.2	交大李码的特征信息集、码长及键位分布	76
3.5.3	交大李码的字编码输入方法	76

3.5.4	交大李码的词编码输入方法	78
3.5.5	讨论	78
3.6	PJS/TLS 汉字输入系统	79
3.6.1	概述	79
3.6.2	系统的特征信息集、码长及键位分布	79
3.6.3	拼音—汉语变换系统 PJY	80
3.6.4	拼形汉字输入系统 PJX	83
3.6.5	讨论	85
3.7	钱码	86
3.7.1	概述	86
3.7.2	钱码的特征信息集、码长及键位分布	86
3.7.3	钱码的字编码输入方法	86
3.7.4	钱码编码应注意的问题及编码实例	87
3.7.5	讨论	88
3.8	四笔声形码	88
3.8.1	概述	88
3.8.2	四笔声形码的特征信息集、码长及键位分布	88
3.8.3	四笔声形码的字编码输入方法	89
3.8.4	四笔声形码的词编码输入方法	90
3.8.5	讨论	90
3.9	团法	91
3.9.1	概述	91
3.9.2	团法的特征信息集、码长及键位分布	91
3.9.3	团法的汉字键盘选字输入方法要点	95
3.9.4	团法部分偏旁部首的归类举例及其它	96
3.9.5	讨论	96
3.10	五十字元法	96
3.10.1	概述	96
3.10.2	五十字元法的特征信息集、码长及键位分布	97
3.10.3	五十字元法的字编码输入方法	98
3.10.4	五十字元法的全形码及形声码检索字典	99
3.10.5	讨论	99
3.11	未来码	100
3.11.1	概述	100
3.11.2	未来码的特征信息集、码长及键位分布	100
3.11.3	未来码的字编码输入方法	101
3.11.4	未来码的词编码输入方法	102
3.11.5	讨论	102

3.1.2	肖码	102
3.12.1	概述	102
3.12.2	肖码的特征信息集、码长及键位分布	103
3.12.3	肖码的音形编码输入方法	105
3.12.4	肖码的纯形码编码输入方法	106
3.12.5	肖码的屏幕提示功能	107
3.12.6	讨论	107
3.1.3	郑码	107
3.13.1	概述	107
3.13.2	郑码普及型的特征信息集、码长及键位分布	108
3.13.3	郑码普及型的字、词编码输入方法	108
3.13.4	郑码标准型的字、词编码输入方法	109
3.13.5	讨论	110
3.1.4	自然码	111
3.14.1	概述	111
3.14.2	自然码的特征信息集、码长及键位分布	112
3.14.3	自然码的字编码输入方法	113
3.14.4	自然码的词编码及字、词相关智能处理	114
3.14.5	自然码人机界面的安排和辅助功能	114
3.14.6	讨论	116
3.1.5	双拼及简拼输入系统	116
3.15.1	几种流行的双拼	116
3.15.2	五种双拼的键位分布表	117
3.15.3	声韵双拼以外的简拼方式	118
	参考文献	121
第四章	不同类型和特点的汉字键盘输入系统(下)	123
4.1	仓颉输入法(台湾)	123
4.1.1	概述	123
4.1.2	仓颉输入法的特征信息集、码长及键位分布	123
4.1.3	仓颉输入法的字编码输入方法	123
4.1.4	仓颉输入法编码规则的说明、补充和实例	124
4.1.5	讨论	127
4.2	大易输入法(台湾)	127
4.2.1	概述	127
4.2.2	大易输入法的特征信息集、码长及键位分布	127
4.2.3	大易输入法的字、词编码输入方法	127
4.2.4	大易输入法编码实例	129
4.2.5	讨论	129

4.3	行列输入法(台湾)	129
4.3.1	概述	129
4.3.2	行列输入法的特征信息集、码长及键位分布	130
4.3.3	行列输入法的字编码输入方法	131
4.3.4	讨论	132
4.4	华象中文输入法(台湾)	132
4.4.1	概述	132
4.4.2	华象中文输入法的特征信息集、码长及键位分布	133
4.4.3	华象中文输入法的字编码输入方法	135
4.4.4	华象中文输入法的注音输入方法	135
4.4.5	讨论	136
4.5	呖虾米中文输入法(台湾)	136
4.5.1	概述	136
4.5.2	呖虾米中文输入法的特征信息集、码长及键位分布	136
4.5.3	呖虾米中文输入法的字编码输入方法	138
4.5.4	讨论	138
4.6	轻松输入法(台湾)	138
4.6.1	概述	138
4.6.2	轻松输入法的特征信息集、码长及键位分布	138
4.6.3	轻松输入法的字、词编码输入方法	139
4.6.4	轻松输入法编码图解及举例	139
4.6.5	讨论	140
4.7	饶氏笔形输入法(美国)	140
4.7.1	概述	140
4.7.2	饶氏笔形输入法的特征信息集、码长及键位分布	141
4.7.3	饶氏笔形输入法的字编码输入方法	142
4.7.4	取笔形和取角注意要点	142
4.7.5	讨论	143
4.8	唯物输入法(香港)	143
4.8.1	概述	143
4.8.2	唯物输入法的特征信息集、码长及键位分布	143
4.8.3	唯物输入法的字、词编码方法	144
4.8.4	讨论	145
4.9	王安三角编码法(美国)	146
4.9.1	概述	146
4.9.2	王安三角码的特征信息集、码长及键位分布	146
4.9.3	王安三角码的字编码输入方法	147
4.9.4	讨论	148

4.10 象形王码(台湾)	149
4.10.1 概述	149
4.10.2 象形王码的特征信息集、码长及键位分布	149
4.10.3 象形王码的字编码输入方法	149
4.10.4 讨论	151
4.11 中易系统(台湾)	151
4.11.1 概述	151
4.11.2 中易系统的特征信息集、码长及键位分布	152
4.11.3 中易系统的字编码输入方法	154
4.11.4 讨论	155
参考文献	155
第五章 汉字键盘输入系统的质量评估与选用	156
5.1 汉字键盘输入系统评测的意义	156
5.2 汉字键盘输入的认知模型	158
5.2.1 知觉加工,形表象与音表象	159
5.2.2 认知加工,编码过程	159
5.2.3 动作加工,键位相关击键时间当量与选键时间	160
5.2.4 长期记忆与短期记忆	164
5.3 汉字键盘输入系统的易学性评测	164
5.3.1 长期记忆的内容	165
5.3.2 训练量与统选操作员训练曲线评测法	166
5.4 汉字键盘输入系统速度素质的评测	169
5.4.1 决定输入系统速度素质的因素	169
5.4.2 选键、击键手动时间的评测	170
5.4.3 操作员竞赛性评测	173
5.5 汉字键盘输入系统易用性与规范化评测	173
5.5.1 易用性	173
5.5.2 规范化程度	174
5.6 汉字键盘输入系统其它项目的评测	174
5.7 怎样选用汉字键盘输入系统	175
5.7.1 一般使用人员	175
5.7.2 职业电脑打字人员	178
参考文献	181
第六章 汉字键盘输入系统的设计	183
6.1 概述	183
6.1.1 汉字键盘输入系统设计涉及的领域	183
6.1.2 汉字键盘输入系统的研究方法	183
6.1.3 汉字键盘输入系统的设计思想	183