

微机 汉字信息处理系统 实用大全

陆均良 陆璇辉 编著

上海科学普及出版社

微机汉字信息处理系统实用大全

陆均良 陆璇辉 编著

上海科学普及出版社

(上接封二)

可编程序控制器应用指南

A/D、D/A 集成电路应用手册

Intel 80486 微处理器技术参考手册

激光打印机的原理、使用与维修

微机和外设维修技巧

微机中文信息处理实用技巧

微机应用和编程技巧

微型计算机初级操作技术培训教程

微型计算机操作指南

家庭电脑及其应用

个人电脑新知

微机汉字信息处理系统实用大全

开关电源的设计和使用

555 集成电路实用电路集

前 言

微机汉字信息处理技术是微型计算机应用的一个重要分支,随着我国经济建设的不断深入,对汉字信息处理技术的要求也越来越高,原来那种手工方式的汉字信息处理已极不适应现代社会的发展需求。微机汉字信息处理技术就是利用微型计算机的运算速度快、存储容量大的优越性,对汉字信息进行快速录入、快速处理、快速分类、快速传递、快速排版等一系列汉字信息处理的一种科学方法。在经济建设的各个领域都要涉及到汉字信息的处理,因而微机汉字信息处理技术是现代社会发展中所必需的一门实用技术。

微机汉字信息处理可分为精密型汉字信息处理系统和通用型汉字信息处理系统两大类,前一类对汉字信息处理的要求较高,适用于报社、书籍出版的编辑、排版等;后一类相对来说要求较低,适用于一般的数据处理、办公日常事务处理等,因此,后一类应用面较广,微机汉字信息处理技术的推广应用,也主要取决于后一类的自身发展。本书的主要目的,就是介绍后一类的通用型汉字信息处理系统。

全书共分八章,前三章是基础部分,介绍微机汉字信息处理系统的概况及发展前景,同时也介绍了微机操作的基础和常用命令,后五章从 Wordstar 开始,介绍国内最具有代表性的几种汉字处理系统的应用,如 WPS 汉字处理系统、时代汉字处理系统、M-6403 处理系统以及 2.13 系列汉字处理系统。全书内容由浅入深,系统性和实用性较强,是计算机应用技术人员必备的一本参考书,也可作为汉字信息处理技术应用的培训教材。

在本书的编写过程中,杭州自动化技术研究院的叶绍华高级工程师和杭州大学的张森教授以及西湖电子集团公司职工大学校长吴震方高级工程师,副教授时海良老师给予了很大的帮助,并提出了不少宝贵意见,在此表示深切感谢!同时西湖电子集团公司的冯曙文同志对本书的编辑和录入做了大量的工作,在此也表示感谢!

由于编写人员水平有限,书中难免出现一些错误和不妥之处,敬请广大读者批评指导。

编 者

1993 年 7 月于杭州

目 录

第一章 汉字信息处理系统概论	(1)
第一节 汉字信息处理的意义和任务	(1)
一、汉字信息处理的意义	(1)
二、汉字信息处理涉及的范围	(1)
三、汉字信息处理的任务	(2)
第二节 汉字信息处理系统的构成和分类	(4)
一、汉字信息处理系统的构成	(4)
二、汉字信息处理系统的分类	(5)
第三节 汉字信息处理系统的现状和发展前景	(5)
一、我国汉字信息处理技术的现状	(5)
二、汉字信息处理技术发展前景	(8)
第二章 微型计算机操作基础	(10)
第一节 硬件的一般性知识	(10)
一、计算机硬件系统组成	(10)
二、微机系统的硬件组成	(11)
第二节 磁盘操作系统及其命令	(12)
一、目录和路径类	(13)
二、磁盘初始化命令 FORMAT	(15)
三、文件操作类命令	(15)
四、常用 DOS 命令列表	(19)
第三节 汉字操作系统介绍	(21)
第四节 键盘的操作技巧	(22)
一、正确的击键姿势	(22)
二、正确的键入方法	(22)
三、键盘指法分区	(23)
第三章 汉字结构及汉字的录入方法	(25)
第一节 汉字结构	(25)
一、汉字字形	(25)
二、汉字字音	(26)
三、汉字字意	(26)
第二节 汉字输入编码方法	(27)
一、字根代码类	(27)
二、角形代码类	(29)
三、字音代码类	(29)

四、音、形结合代码类	(30)
第三节 微机处理的汉字代码	(31)
一、汉字代码种类	(32)
二、中西文兼容对汉字编码的影响	(35)
三、建立统一的汉字代码,提高计算机系统效率	(36)
第四节 微机常用汉字输入方法	(36)
一、拼音输入法	(37)
二、五笔字形输入法	(39)
三、区位码输入法	(45)
四、首尾码输入法	(46)
五、快速输入法	(46)
第五节 微机的汉字识别	(47)
第四章 汉字编辑软件 Wordstar	(49)
第一节 汉字 WORDSTAR 的运行环境及启动	(49)
一、运行环境	(49)
二、启动	(49)
第二节 进入编辑	(50)
第三节 编辑命令	(51)
一、光标和屏幕控制命令	(51)
二、常用编辑命令	(52)
第四节 非文书文件编辑	(56)
第五节 其它处理功能	(56)
第五章 WPS 汉字处理系统	(57)
第一节 Super-CCDOS 介绍	(57)
一、Super-CCDOS 的特点及功能	(57)
二、Super-CCDOS 的发展历程	(57)
三、Super-CCDOS 的运行环境	(58)
四、字库读取模块 SPLIB.COM	(59)
五、基本输入模块和显示模块 SPDOS.COM	(59)
第二节 WPS 介绍	(61)
一、WPS 系统简介	(61)
二、WPS 的一些基本概念	(62)
第三节 WPS 的使用	(66)
一、WPS 系统启动	(66)
二、WPS 主菜单的使用	(68)
第四节 命令菜单的使用	(69)
第五节 编辑文本	(71)
一、编辑方式	(71)
二、光标移动	(71)

三、插入文本	(74)
四、删除文本	(75)
五、分行与分页	(75)
第六节 文件操作	(76)
一、文件概念	(76)
二、文件操作	(78)
三、与文件有关的其它功能	(80)
第七节 块操作	(81)
一、标记块	(81)
二、块的操作	(82)
三、块的列方式	(83)
四、块的磁盘操作	(83)
五、块的取消	(84)
六、大规模块的操作	(84)
七、复制 CCDOS 块	(84)
第八节 查找与替换文本	(85)
一、查找与替换命令	(85)
二、方式选择项	(87)
三、查找字句中的控制符	(88)
第九节 设置打印控制符	(88)
一、打印字样控制符	(89)
二、打印格式控制符	(94)
三、设定分栏打印	(95)
四、打印控制符的特性及有效范围	(96)
五、打印控制命令汇总表	(97)
第十节 窗口功能及其它	(98)
一、窗口操作	(98)
二、重复执行命令	(102)
三、终止命令和暂停命令	(103)
四、计算器功能	(103)
五、执行 DOS 命令	(104)
第十一节 文本编辑格式化及制表	(105)
一、页的边界及编排	(105)
二、改变窗口显示	(106)
三、取日期与时间	(107)
四、制表格	(107)
第十二节 模拟显示与打印输出	(109)
一、模拟显示	(109)
二、打印输出	(110)

三、改变当前打印参数	(112)
四、安装新的 24 针打印机参数	(114)
第十三节 文件服务与帮助功能	(116)
一、文件服务功能	(116)
二、帮助功能	(117)
第六章 时代汉字处理系统	(123)
第一节 SD-DOS 一般介绍	(123)
一、时代汉字系统的特点	(123)
二、时代汉字系统的配置	(124)
三、安装	(124)
四、实用程序介绍	(125)
五、各种功能键	(126)
六、撤离系统	(128)
第二节 打印输出	(128)
一、汉字属性菜单输入法的使用	(129)
二、打印字样控制	(129)
三、打印修饰控制	(129)
四、打印格式控制	(131)
五、打印特殊功能控制	(133)
六、打印控制命令的特性及有效范围	(133)
七、选择打印机	(135)
八、打印输出	(136)
第三节 时代汉字处理系统概述	(137)
一、CPS 简介	(137)
二、CPS 使用环境	(137)
第四节 CPS 的操作使用	(138)
一、菜单的使用	(138)
二、进入 CPS 系统	(138)
三、编辑和修改	(138)
四、文件的打印	(139)
五、退出编辑	(139)
第五节 文件操作及其它	(139)
一、与 CPS 系统有关的文件	(139)
二、文件操作	(140)
三、文件服务功能	(140)
四、编辑状态下帮助命令	(141)
五、计算器功能	(141)
第七章 M-6403 汉字处理系统	(143)
第一节 M-6403 基础	(143)

一、M-6403 的安装	(143)
二、基本使用方法	(145)
三、输入方法接口与调用	(148)
四、打印机接口	(150)
第二节 排版打印	(153)
一、文件	(154)
二、参数	(156)
三、编辑	(159)
四、作图	(161)
五、排法	(164)
六、操作	(168)
七、打印	(169)
第三节 文字编辑	(170)
一、文件功能	(172)
二、查找功能	(174)
三、字块功能	(175)
四、排版功能	(177)
第四节 表格编辑	(180)
一、文件	(183)
二、参数设置	(184)
三、表项编辑	(184)
第五节 其它功能	(185)
一、图形编辑	(185)
二、图象编辑	(186)
三、文件管理	(187)
四、造字	(187)
五、图库维护	(187)
六、安全检测	(188)
第八章 2.13 系列汉字处理系统	(189)
第一节 2.13 系列介绍	(189)
一、2.13 系列汉字系统的特点	(190)
二、2.13 系列的设计思想	(190)
三、实现方法	(191)
第二节 2.13H 汉字系统安装与启动	(192)
一、应用环境	(192)
二、文件配置	(193)
三、系统安装	(196)
四、系统启动	(202)
第三节 2.13 系统的应用	(213)

一、功能键及其使用	(213)
二、显示字库及其安装	(220)
三、特殊显示功能	(224)
四、光标闪烁	(244)
第四节 汉字的打印输出	(246)
一、常用字库技术	(246)
二、字型	(247)
三、特殊打印功能	(250)
四、屏幕拷贝	(260)
第五节 2.13 系统实用程序	(263)
一、造字	(263)
二、通用制表程序	(269)
三、分页、折页打印功能	(271)
四、查询及修改文件属性	(275)
五、其它功能	(276)
附录 A WORDSTAT 命令汇总	(278)
附录 B ASCII 码表	(283)
附录 C 五笔字型键盘字根总图	(285)
附录 D 简拼、全拼编码对照表	(286)
附录 E 图形字符代码表(GB 2312—80)	(288)

第一章 汉字信息处理系统概论

第一节 汉字信息处理的意义和任务

信息是维持人类生产活动、经济活动和社会活动的第三种资源,信息具有多种性能,例如可传输性、可转换性、可存储性、可处理性,以及可再生性等。随着科学技术的发展,信息的传输效能愈益增强,其作用范围也愈益宽广。这里我们涉及和研究的是汉字信息,以及汉字信息的处理,说得确切些就是微型计算机的汉字信息处理。

一、汉字信息处理的意义

由于电子计算机具有运算速度快,存储容量大的特点,因此我们用计算机技术来存储和控制信息,加工和处理信息。这里涉及的信息都是汉字信息,随着计算机系统功能的不断提高,应用领域的迅速扩展,汉字信息处理的概念、涵义、作用和涉及的范围也大大扩展了。也就是汉字信息处理的内容更加丰富了。例如汉字信息处理的内容可以有:书刊和报纸的自动编辑和排版;日常事务的处理;办公室自动化;文字或文书处理;文字翻译;公用咨询服务;数据通讯等。实际上,汉字信息处理技术已逐渐渗透到人类思维,生产和生活等活动的所有方面,汉字信息处理技术同科研、生产实践以至人类社会的一切活动都离不开它。以电子计算机为技术基础的汉字信息处理技术,正在促使人类的社会经济、科学技术和家庭生活发生日新月异的变革。如现代的电子词典,英译先英汉字典,电脑记事簿等都已开始进入家庭,这标志着人们已开始进入了信息化的时代。

我国的汉字是一种表意文字或象形文字,字量多,字型复杂。在构成的汉字信息处理系统中,除了通用硬件设备以外,还需配备汉字输入输出设备以及相应软件处理系统,使软件处理既能适合汉字处理,又能适合西文处理。我国的科技工作者通过艰苦的努力,已研制出十多种适合不同应用范围的汉字信息处理系统,使我国的汉字处理发生了革命性的变化。特别是改革开放的今天,我国的经济建设已经离不开这些微型计算机处理的汉字信息处理系统。这些系统担负着我国经济建设中的信息收集、信息的加工、信息的传输。使这些信息能更好地为我国经济建设服务。

二、汉字信息处理涉及的范围

汉字信息处理技术是一项综合性的技术,其核心技术是计算机技术,其涉及的技术范围包括以下几个方面:

1. 汉字字量

在汉字信息处理技术中,需要使用多少个汉字是个重要问题,我国汉字多至 5~6 万,国标 GB2312 中收集一级常用汉字 3755 个,二级常用汉字 3008 个,共计 6763 个。因此,一个汉字处理系统应收集多少个汉字字量,应根据实际的使用要求来确定。

2. 汉字字体

特别在印刷排版的汉字处理系统中,对字体种类的要求较高。一般来说,汉字字体可分为宋体、仿宋体、楷体、黑体等。而每一种又有方体、长体、扁体的区别。如 M-6403 汉字处理系统可进行处理几十种汉字的字体。

3. 汉字索引

可以从不同的角度检索汉字,例如以笔画、偏旁或部首来检索,也可以用汉字发音简捷的法则准确地查得某个汉字或它的标准编码。

4. 使用频度

对于不同的汉字,其使用频度的差别是很大的。同一汉字在不同专业领域中使用时,其频度也有差异。我们根据汉字的频度,可以把汉字分为常用字,次常用字,稀常用字,罕用字等几个等级,在建立不同种类的汉字处理系统中,必须根据使用频度来选用字库中所收容的汉字。

5. 字形分解

字形分解就是把汉字分解为部首、字首、笔画、位点。位点看成是组成汉字的最小单位。分解字形是为了找出汉字结构的规律。通过对字形的分析研究,可以选取最少量的字根,合理地组成所需的汉字,在以字形特征为基础的汉字编码方法中,特别需要注重字形分解的研究工作。

6. 汉字词组

在汉字信息处理技术中,对词的研究很重要。在汉字输入方案中,对于使用频度特别高的词,可以用软件方法设定,用一个键位代表一个词,也可以根据需要改变某个键位所代表的词,也可以把词组组成词典,存入系统的存储器中,以便建立索引或进行查找。

7. 计算机及硬件

信息的处理离不开计算机,除了计算机的硬件设备以外,还需要汉字键盘,汉字显示终端,汉字输出设备等。根据汉字系统规模的大小和应用特点,还可配置其它外设或适当的外存设备。

8. 计算机软件

汉字信息处理系统的软件除了通常的操作系统、高级语言编译程序外,还需要配上汉字操作系统、汉字处理(或管理)程序,以及汉字字库等。而且要求所配的软件要中西文兼容。

三、汉字信息处理的任务

汉字的特点是字量大,字形复杂,因此要建立一个汉字系统,首先要解决汉字的输入、存储和输出问题。因此一个汉字信息处理系统需解决如下几个问题。

1. 汉字输入编码问题

虽然可用许多种方法实现汉字编码,但要得到一种功能上最佳,并且适用面很广的汉字编码方法却非易事。

目前,汉字输入编码可粗略地分为两大类。一类称为整字编码法,实质上就是把汉字按某种规则排定先后次序,用其序号作为汉字代码;另一类是组合编码法,按照所采用的具体方法不同,组合编码可以分成许多种,例如按照字形特征的编码,称为形符法;按照汉字发音特征编码的,称为音符法;有形音结合的编码;也有按照汉字的其他特征编码的。

汉字整字编码输入方法的优点是直观,操作者容易学习和掌握,没有重码问题。但它所用的键盘是通称的整字大键盘,其体积大,造价高,输入速度低。对于整字汉字键盘,除了盘

内字以外,还要解决外字的输入问题。目前多数用户直接送入汉字代码,或者用汉字字根组合编码的方法实现外字的输入。大键盘只能在少数场合使用,无法普及推广。

当前,汉字编码输入方法很多,研究的汉字输入方案有几百种,实用的有拼音输入法,五笔字型输入法,区位码法,首尾码输入法,快速输入法以及自然码输入法等。由于用不同的输入法得出的汉字键盘码差别很大,故为便于不同的汉字信息处理系统相互交换汉字代码,需要确定一个统一的码制,称为标准汉字交换码。我国已颁布的作为国家标准的 GB2313《信息交换用汉字编码字符集——基本集》,就是一种标准交换码,在该基本集中,除了有二级汉字外,还有几种外文字母、数字和符号,其总数为 7445 个。在国际交换码中,每个汉字、字母、数字和符号用两个字节来表示。

2. 汉字字模的存储问题

计算机信息处理用的汉字字模,按用途可分精密型字模和通用型字模两种。精密型字模用于编辑排版系统,这种汉字字模,对字形、字体、字号变倍等都有严格要求,必须适合出版行业所定的规格。通用型汉字字模适用于一般的汉字信息处理系统,应用面广。对于通用型的汉字字模。常用的点阵结构有以下几种:

简易型:15×16 点

普通型:24×24 点

提高型:32×32 点,48×48 点

在计算机的汉字信息处理系统中,存储汉字字模的存储体系叫汉字字模库或汉字字模发生器。其存储形式目前有二种,一种是存储整字字模信息的汉字字模库,这种字模库的特点是单位存储量的成本低,读取信息的速度快。另一种是存储压缩信息的汉字字模库,这种字模库采用的是信息压缩存储技术,其压缩原理可采用只收汉字字根的汉字字模库;利用向量组字法产生汉字字模库;用霍夫曼树型压缩法存储汉字信息等。目前用得较多的方法是矢量法和字根法二种。这种字模库的优点是存储汉字量大,一般 40~60K 字节的 EPROM 组件可存储产生 7000~20000 个汉字,缺点是需用软件方法组成汉字,所以输出速度较低。

3. 汉字输出问题

汉字输出主要包括汉字显示和汉字打印,由于汉字字模的点阵密,对显示器和打印机的技术要求比显示和打印西文的同类设备要高。因此,汉字的输出问题主要是汉字的显示技术和汉字的印刷打印技术。目前汉字显示用的显示器有 600×200 点阵发展到 640×350 点阵,800×600 点,1024×768 点阵等。汉字打印机也从原来的 9×9 点阵打印机发展到 16×16 点阵打印机,24×24 点阵打印机,以及 32×32 点阵打印机。目前还有更高级的激光打印机,这些打印机所印刷打印的汉字已经同印刷体所印的汉字没有什么区别了。在汉字信息处理系统中,用户可以根据要求和不同的需要,选用合适的显示设备和印字输出设备。

4. 汉字终端技术

汉字终端技术是汉字信息处理中的重要设备,它同普通的显示设备不一样,这种终端具有通信功能,采用标准的 RS232 串行接口,可以进行远程控制。目前汉字终端有二种,简易型汉字终端和智能型汉字终端。前者具有汉字的输入输出功能,汉字代码转换功能和汉字编辑功能,自身还带有汉字字模库;后者除了以上的功能外,还有文件管理功能,表格处理功能和图形处理功能等。这种终端可以脱机使用,依靠本身的软件资源作为独立的工作站。

5. 汉字、西文兼容问题

计算机只能处理 0 和 1 两种信号,因此汉字信息处理也必须先把汉字量化,即数码化,使其转变成计算机能直接处理的信息。然而,计算机是首先在欧美等西方国家得到发展,因而西文字母先为计算机所接受,接着又用西文来替代机器码,发展成为汇编语言,进而又发展成各种接近语言的高级程序设计语言。这些长期积累的经验和技术都是针对西文系统,并用西文表达的。对于汉字信息处理系统的主要内容是汉字,要发展汉字信息处理首先碰到的是和西文系统的兼容。也就是说,原有的西文系统字符代码同汉字系统的汉字代码要协调一致实现信息的处理,特别使西文的高级程序设计语言也能适应处理汉字信息。目前通过软件手段已能顺利地解决汉字和西文的兼容问题,使实用的汉字信息处理系统既能处理汉字信息,又能处理西文信息。如目前流行的有 CCDOS 系列,Super-DOS 系列,SD-DOS 系列以及 2.13 系列等。

第二节 汉字信息处理系统的构成和分类

一、汉字信息处理系统的构成

汉字信息处理系统的组成包括硬件和软件两大部分。

(一)硬件组成

汉字信息处理系统的硬件包括主处理机、常规外部设备和汉字外部设备。主处理机是通用的电子计算机。根据实际需要选择合适的主机类型。常规外部设备主要包括外存储器,例如磁盘机和磁带机。汉字外部设备包括汉字输入键盘、汉字打印机、汉字显示终端设备等。

在汉字信息处理系统中,汉字字模库和汉字显示终端是重要的组成部分。这些设备决定了系统的汉字部分工作方式。通常汉字字模库在系统中有三种设置方式,一种是汉字字模库作为一个独立的外部设备,这样可以把字模库给多台设备所共享,特别是当汉字字模库价格较高时,这种设置方法有利于降低系统的造价。缺点是系统效率较低。第二种是字模库直接同打印机相连,这样系统只传送代码信息,可以提高系统的效率。第三种是字模库设置在汉字显示终端上,目前大部分的汉字终端均配有字模库。

因此对于一个汉字信息处理系统,所选配的主机类型,字模库设置方式,汉字设备的要求,也就决定了该系统处理性能,工作方式以及成本价格。

(二)软件组成

汉字信息处理系统的软件包括系统软件和应用软件两大类。

1. 系统软件

系统软件包括能兼容汉字和西文信息处理的操作系统。该操作系统除了保留通常西文系统全部功能的条件外,还包括各种汉字设备的驱动模块,能直接调用汉字输入输出管理程序以及汉字编辑程序。汉字输入输出管理程序也是系统软件的一部分,它包括汉字输入输出接口程序,以及汉字输入、换码、访问汉字字模库和汉字输出等程序;对于输入计算机的代码信息,当系统识别出它是汉字信息时,便根据输入码的不同编码方式,将其转换成标准汉字代码,供加工处理。当输出汉字时,先把标准码变成汉字字模库中对应的地址码,然后读出字模信息,以供显示或打印。除此以外,能直接处理汉字信息的各种高级程序设计语言都属于系统软件范围。

2. 应用软件

应用软件的范围很广,可根据汉字信息处理系统的性质和用途,对各种西文系统的应用软件进行汉化,使其能直接处理汉字信息,这样便就成了汉字信息处理系统的应用软件了。例如,西文 Wordstar 文本编辑软件,经过汉化处理后,就成为汉字 Wordstar 的文本编辑软件,因此除了用户自己设计各种应用程序外,应尽量利用西文系统已有的一些应用程序包,这样便可迅速扩大汉字信息处理的应用范围。

二、汉字信息处理系统的分类

从汉字信息处理系统输出汉字文字的质量要求来看,目前可把汉字信息处理系统分成两大类:即精密型汉字编辑排版系统和通用型汉字处理系统。前一种类型用于正式出版的书籍、报纸的编辑排版;后一种用于一般的汉字文件处理,统计报表,文书处理以及数据处理等。汉字信息处理技术的推广应用取决于后一类型的发展。

1. 精密型汉字编辑排版系统

这一系统最重要的关键技术就是高精度汉字字模的存储和版面输出。对于精密汉字字模,不仅每个字的点阵信息量大,而且由于字量多,需要多种字体和字号,从而使总的字模信息很庞大。同时还要兼顾到适当的字模输出速度。这类系统的输出技术大部分采用激光扫描输出,在扫描的同时,输出版面在感光底片上记录成象,因而这种设备又称为照排设备。整个系统包括:排版用计算机、相应的外部设备、编辑、改错用的联机汉字显示终端、汉字数据采集用的汉字终端、校样印刷机、照排设备、字模自动制作设备,以及图片输入设备等。目前这样的系统有:华光Ⅳ型书出版系统;北大方正 91 系统;大屏幕报刊组版系统;四通 4S 系统;α-100 桌面系统,以及科印书刊排版系统等。

2. 通用型汉字信息处理系统

这类系统的特点是实现数据处理的和一般的汉字信息处理,其使用面广,系统的成本低,不需太讲究汉字字模的质量。通用型汉字系统的字模点阵目前流行的有 24×24 点; 32×32 点以及 48×48 等点阵。汉字的输出以普通的 24×24 点阵针式打印机为主,今后也推广激光扫描式印字机。目前这类系统主要用于企事业单位的事务处理,企业信息管理,现代办公用计算机,家庭的电子记事簿,以及其它的一些汉字通讯等。现在市场上流行的有金山汉字处理系统,四通汉字处理系统,M-6403 汉字处理系统,2.13 系列汉字处理系统等,这类系统除了保持低价实用的优良性能,目前也在汉字输出质量方面做了大量的研究工作,汉字输出也可配备照排设备,以此同精密汉字编辑排版系统争夺市场。努力推广汉字信息处理系统在四化建设中的普及应用。

本书的目的就是系统介绍我国市场上流行的几种通用型汉字信息处理系统。促使我国各行各业的数据处理、办公文字处理、日常报表处理等工作早日实现计算机化。

第三节 汉字信息处理系统的现状和发展前景

一、我国汉字信息处理技术的现状

随着大规模集成电路的发展,存储器和微处理器芯片的价格越来越便宜,因而促进了我国微机汉字信息处理技术的发展,硬件上研制了一批新型的汉字输入输出设备,软件上呈现

出一大批汉字排版处理系统,目前除了进一步研究汉字基础理论的同时,还研制了各种各样的汉字信息处理设备和软件系统,使汉字信息处理技术以更快的速度向前发展。下面以汉字信息处理技术的基础工作,印刷体汉字识别技术,以及汉字设备的研制和生产等现状作简要介绍。

(一)汉字信息处理技术的基础工作

1. 汉字字符集标准的确定

在汉字使用频度统计的基础上,根据原《信息交换用汉字编码字符集——基本集》的国家标准 GB2312,今年又确立 ISO-10646 国际标准,该标准的确立将为电脑汉字的完善化,汉字信息交换,汉字软件的发展打开新的天地,对我国汉字信息处理技术的发展将起很大的促进作用。

2. 汉字词汇的研究

为了使汉字信息处理技术水平推向更高的阶段,目前正进一步开展对汉字词汇的研究和进行词频的统计工作,开展对汉字自然语言处理的研究,以此来促进设计出更先进的汉字输入设备。

3. 汉字编码输入方法的研究

汉字输入编码除了采用整字输入方法外,目前主要使用编码输入法,要设计出一种既好学、易使用、输入速度快的编码方法的确不是容易的事,但在原方案上改进的输入编码方案不断涌现,目前各种汉字编码已达 400 多种,并且还在发展,如五笔字型是一种常用的输入编码方案,其优点就是输入快,如果和“音”、“形”结合再进行编码,就可以实现易学的优点。又如改进自然码输入法等。可以肯定,一种具有综合性能指标并能为大多数用户所接受的编码方案一定会出现。

4. 汉字自动识别的研究

为了彻底改变汉字输入速度问题,目前汉字识别研究已成为热门课题而且取得可喜的进展,下一节就专门介绍这方面的现状。

(二)印刷体汉字识别技术的现状

近 20 年来,随着计算机的发展,社会进入信息时代,汉字出现了真正的难题——计算机输入输出。对输出而言,西文字母用 5×7 点阵就能表达,而有的汉字用 15×16 点阵都表达不全。对输入而言,西方用键盘输入字母十分方便,而汉字一定要按照某种编码方案编成字母串,再利用键盘输入。对于每个要与计算机打交道的人来说,增加了学习汉字编码方案的额外负担。

汉字识别是十分困难的模式识别问题。其困难首先在于常用汉字的数量极大,相似字较多,解决对成千上万个不同汉字这种超大数量模式类别的识别问题是十分困难的,而且识别速度也必然会大大降低。这比西方拼音文字仅仅需要区分辨识数十种(如英文大小写字母共 52 类)字母要困难的多。汉字识别还有一个人们不易注意的难点,就是同一汉字字形的多变性。不仅手写体汉字如此。不同个人或不同时间书写的风格和字形都不相同;而且对于印刷体汉字,不同字体、不同字号、不同印刷方式、不同印刷情况所印刷的汉字字形也会有或多或少的变化,这都给汉字识别带来极大的困难。

要完全实现印刷汉字文本的自动输入,是一个极为繁杂的和困难的问题。因为我们实际要面对着世界上,至少是大陆所有印刷单位的各种印刷文本。对这些不同内容,采用不同印

刷方式,在不同纸张上以不同字母印出的不同字体、字号、深浅的印刷文本,都要能以相当高的辨识速度将它们自动识别以后自动输入计算机。经过多年艰苦的努力,我们对一定范围内印刷文本的自动输入取得了较好的结果。但是要实现完全的印刷文本自动输入,还有相当多困难的问题需进一步加以解决。

汉字识别是一个大字符集的超多类模式识别问题。按我国对常用汉字的统计,必须能识别 3000~6000 余字。要准确地分辨和识别这些汉字,无疑是十分困难的。近年来对这种大字符集汉字识别方法的研究取得了相当的进展。

近年来,我国印刷汉字识别的研究在“七五”计划和国家“863”计划的支持下取得了显著进展,主要表现在:

(1)解决了印刷汉字识别的基本方法。即采用汉字结构特征的统计模式识别方法是解决超大字符集印刷汉字识别的有效方法。

(2)基本上解决了印刷汉字识别的特征选择问题。即选择足够数量的、能反映汉字笔划结构和内部拓扑关系的结构特征,作为对印刷汉字的经验模式,类别判决的统计特征。

(3)基本解决了汉字模式分类过程中配比分类的优化处理识别速度和识别率之间的矛盾。即在保证识别率的前提下,利用分类字提高识别速度。

(4)一定程度上解决了印刷汉字识别前处理,即切分归一化的问题,使各种字号的汉字可以同时得到识别。

(5)一定程度地解决了后处理中字义相关匹配后纠错的问题,使印刷文本篇章的识别率有所提高。

(6)初步建立起适用的人机界面。

上述六个问题的进展,使我国在印刷汉字识别,特别是多体印刷汉字识别方面的研究居国际领先地位。我们已研制出能同时适应各种字体变化(宋、仿宋、黑、楷四大字体及其各种变形变体和各子模变化),适应各种字号大小变化,适应各种印刷方式(铅印、激光照排、激光打印、计算机针打、胶印、复印等)造成的汉字字形变化,以及适应各种印刷质量的汉字识别系统。这种多体印刷汉字识别系统具有很高的适应性和抗干扰能力,能实现几种实际印刷文本的自动输入,受到广大用户的欢迎。

还应看到,真正解决印刷汉字的计算机自动输入问题还有相当的困难要克服。首先是如何进一步减少误识,提高识别率的问题。我们知道,识别率与被识别印刷文本的质量密切相关。目前我们研制的印刷汉字识别系统对印刷质量较好的文本,识别率可以达到 98~99%。但是对印刷质量差的文本,识别率就会下降。在这方面,我们还需做的工作有:

(1)改进预处理,提高和尽量确保每一个汉字切分的正确性。

(2)进一步提高单字的正确识别率。

(3)改进后处理自动纠错能力。

(4)改善人机交互校对纠错处理的效率。

我们知道,提高识别的每一步都要付出极大的努力,但这一步我们必须要走下去。此外,进一步提高识别速度,改进印刷汉字识别的人机界面,是汉字识别系统的使用向“傻瓜”型发展,这些工作都等待着我们去完成。

总之,我国印刷汉字识别技术正在向高技术产业发展,已开始成为高技术信息产业中的一支新军。我们希望它能充实到所有中文计算机系统的输入方式中去,为我国的信息化作出