微型机汉字信息处理技术及其应用

刘尚威 主编

微型机汉字信息处理技术及其应用

刘 尚 威 主编

冶 金 工 业 出 版 社

内 容 提 要

本书较系统地介绍了微型机汉字信息处理的基本原理和系统设计方法，内容包括汉字库，汉字的编码与输入、汉字的显示与打印，中西文编辑系统，中西文数据库管理系统，应用程序的开发等。书中还扼要地介绍了我国目前部分微型机汉字系统的应用情况和国内外流行的IBM-PC机及其汉字系统。

本书可供科研和工程技术人员参考；亦可作为大专院校计算机、管理工程、自动化等专业的教学参考书。

微型机汉字信息处理技术及其应用

刘尚威 主编

冶金工业出版社出版
（北京北河沿大街嵩祝院北巷39号）
新华书店北京发行所发行
轻工业出版社印刷厂印刷

787×1092 1/16 印张 14字数 325千字
1988年月9第一版 1988年9月第一次印刷
印数00,001～8,000册
ISBN 7-5024-0309-4
TP·12 定价3.35元

前　言

计算机汉字信息处理是计算机科学中的一个新学科。各种计算机汉字系统已成管理现代化的重要工具。

计算机特别是微型机的应用日益广泛，经济效益明显。但是，在我国和使用汉语的其它一些国家和地区，大部分信息要用汉语表示，不解决汉字处理问题，计算机的广泛使用便受到极大限制。为此，研究与应用汉字信息处理技术具有重大现实意义。

为了促进我国计算机应用工作的发展，获取更大的经济效益，促进四化建设事业，满足从事计算机技术工作的科研、工程技术人员、大专院校有关专业师生应用与学习的需要，我们在广泛收集有关汉字处理技术资料的基础上，结合我国近年来研制微型机汉字处理系统的部分成果编写了本书。

本书着重论述了汉字信息处理技术的基本原理及应用。此外，还介绍了我国目前常用的一些微型机汉字信息处理系统及其应用情况。

参加本书编写工作的有陈松乔（第一、七章）、曹宝奎（第二、三章）、杨明泰（第四、五、六章）、陈美术（第八章）、赵显富（第九章）、刘尚威（第十章）。

陈松乔协助刘尚威对全书进行了审校。

由于汉字信息处理技术发展很快，加上我们水平有限，书中不妥之处在所难免，敬请读者批评指正。

著　者

1984年11月

目　录

第一章　绪　论

第二章　字　库

第三章　汉字输入

第四章　汉字的显示与打印

第五章　中文编辑系统

第六章　微型机汉字信息处理系统的设计

第七章　中西文数据库管理系统

第八章　管理系统应用软件的设计

第九章　常用微型机汉字系统及其应用

第十章　IBM-PC微型机及其汉字系统

第一章 绪 论

第一节 研究汉字信息处理技术的意义

信息是现实世界物理状况的反映，信息科学是现代科学的三大支柱之一。当今的一切社会活动都离不开信息的收集、组织、存储、加工、抽取、传输和利用，因此人们称现代是“信息化时代”和“知识化、信息化、信息爆炸的时代”。从原理上讲，任何信息都可以经转换而用数据表示。由于现代社会信息量多，内容复杂，要求传输快、处理准确，使得信息处理工作离不开电子计算机。没有计算机的广泛、成功应用，就谈不上实现我国的四个现代化。

现代计算机除继续在科学计算、生产过程控制方面发挥显著作用外，还越来越多地用于数据处理。在发达国家，用于数据处理方面的计算机已占所拥有的计算机总数的90%左右。可以利用计算机进行企业事业的组织管理工作，比如办公室日常事务的处理，工矿企业的经营决策，生产的计划与调度，情报检索乃至照相排版印刷、机器翻译等。使用计算机，可以大大提高管理水平和促进生产发展。

在我国和使用汉语的其他一些国家和地区，大部分信息要用汉语表示，不解决汉字处理问题，计算机的广泛应用便受到极大限制。因此，我国的信息革命和现代化必然离不开计算机汉字信息处理技术。

汉字信息处理技术的研究任务，大致可分为两大类，一类是使已有的计算机（或新设计与制造的）具备较强的汉字信息处理能力；另一类是使用具有汉字信息处理能力的计算机，实现对企事业单位的管理或完成其他工作。

对于第一类任务，国内外进行了较多的研究，已经由以研究汉字输入编码为中心，逐步转移到研究汉字计算机系统，并研制出一批性能较好、使用方便、成本较低，且能同时处理汉字和西文信息的计算机系统。这些计算机系统类型不一，各有特点，但都还谈不上十全十美，或多或少还存在着一些技术问题有待解决，特别是在系统软件方面还有大量工作要做。

目前，汉字系统的实际应用还处于开始阶段。虽然一些单位已开始了这方面的研究工作，但将汉字系统投入实际使用并取得经济效益的项目还为数不多。因此，研制性能好、功能强、使用方便、成本低的汉字信息处理系统，并将之真正用于社会实际，在实现四个现代化中发挥巨大作用，是摆在我国科研和管理工作者面前的一项长期而艰巨的任务。

第二节 汉字信息处理系统的基本工作原理

在我国，使用汉语的人数最多，因此“汉字信息处理”亦可称为“中文信息处理”。汉字信息处理系统是指能输入、处理、传输和输出包括汉字信息在内的各种信息的计算机系统。汉字信息处理是一门新兴的学科，它涉及的学科范围很广，其中包括语言学、汉字编码、计算机体系结构和计算机软件等。其基本工作原理都是先将汉字转化为数字，而后经计算机进行处理，最后再在输出设备中将数字恢复为汉字。

一、汉字的数字化

计算机完成的任务千差万别，但处理对象归根结底只有基本字符集中所包含的符号、字母和数字，即ASCII码中自00H至FFH共256个代码。其中，代码“0”～“31”为控制代码，“32”～“128”为常用字符代码；“129”～“191”为图形代码，“192”～“255”则为空间压缩代码。可用这些代码所对应的基本字符组成字符串，用以编写各种程序语言程序、各种数据和文件等。ASCII字符集很小，所以集中的每个字符可用通用西文键盘上的一键一码表示。但是，当处理汉字时，由于汉字字数繁多，致使计算机的处理对象大大扩展。为此，计算机处理的符号，最低限度应在原ASCII码的基础上再加上国家颁布的GB 2312-80标准中的一、二级汉字和其它符号，共计7455个。对于这么多的符号，用一键一码表示一个符号显然不能满足需要了。因此，必须设计合理的“编码方案”，根据汉字的特点进行必要的转换，使每个汉字都能用ASCII字符集中的一个或几个字符来表示。这种用一个或几个ASCII字符表示一个汉字的编码，称为输入码。

可以将每个汉字看作是一幅图形。若用墨笔将一个汉字写在坐标纸上，坐标纸上的一些点便涂上了墨水，将这些点记为“1”；而另一些点则没有墨水，将这些点记为“0”。将这些“0”和“1”按一定规则组织起来，便形成了能表示该汉字的“点阵码”。显然点阵越大，点阵中的点越密，越能真实地反映原来的字形；但同时所需的存储容量也愈大。因此选用哪种点阵，应视具体情况而定。目前，常用的点阵码有16×16，24×24和32×32等几种。

所收集的汉字点阵码的有规律集合，称为点阵字库。字库一般存放在只读存储器（ROM或EPROM芯片）、内存储器或外存储器中。当计算机选中一个汉字时，便从字库中调出该汉字的点阵码，并将之送入显示器或打印设备。而后在程序控制下，输出设备根据点阵中的“0”和“1”，将点阵码译成原来的汉字形状。

输入码、点阵码都是数字化了的汉字，都唯一地对应于一个汉字。但是，他们所占的字节数太多，比如对应于每个汉字，一般输入码要占3个或更多的字节。16×16的点阵码要占32个字节，24×24的点阵码所需的字节数更多。因此，输入码和点阵码都不宜直接在汉字文件中使用。事实上，计算机中有两个字节便可以表示65536种状态，当然也能区分7455个或更多个汉字和字符。因此，在汉字处理技术中，常使用占两个字节的“内部码”来表示一个汉字。此时，汉字经过键盘输入而形成输入码。而后，采用一定方法将输入码映射为内部码，由计算机使用内部码，按要求进行各种传送、处置和运算。最后，根据表示结果的内部码与相应点阵码所在地址的一一对应关系，找出对应的点阵码并将之输出，得到相应的汉字字形。

二、中西兼容

任何用于处理汉字信息的计算机，均必须能同时处理西文信息，并保证原系统中运行的一切软件不加修改或稍加修改后便可以在汉字系统中运行，即必须中西兼容。关于中西兼容的含义，有多种解释，其中主要的有两种，一种为，新开发能同时处理中文和西文信息的软件，而原来的西文软件只有在作相应的修改或加上“接口”后才能同时处理中文信息；另一种为，计算机的操作系统能识别中文和西文符号，原西文软件不经修改或稍作修改后便可同时处理中文信息。

不论采用哪种解释方法，实现中西兼容均必须采取一系列关键性措施，其中最基本的是要使计算机能自动识别汉字和ASCII字符。

在计算机内部，汉字内部码和ASCII码均是用数字表示的，表面上看起来，两者没有区别。这样，为了区分它们，必须对表示内部码的字节进行标识，其方式大致可分为下列几类：

（1）标识字位方式。将每个字节的第0位（或汉字内部码两个字节的第0位）作为标识位。如果该位为“0”，表示ASCII码；该位为“1”，则表示汉字内部码。这种方式使用方便，识别两种码的功能可放至操作系统一级来完成，易于做到与原西文软件兼容。但是，由于第0位已用作标识位，故只能用七位二进制数表示ASCII码，最多只能表示128个，而不是256个ASCII字符，而且失去了第0位在许多场合中用作奇偶校验位的功能。同时，由于许多高级语言的编译程序中，常常自动消掉了该位数字，在使用这些语言时，若不修改编译系统，便失去了区分两种码的功能。

（2）标识字节方式。在汉字串或单个汉字的两头加入一对专用字节，以标识两标识字节之间的字节为汉字内部码；或在表示每个汉字内部码的字节之前加入一标识字节，以表示后面的两个字节是汉字内部码，分别参见图1-1中的（a）和（b）。

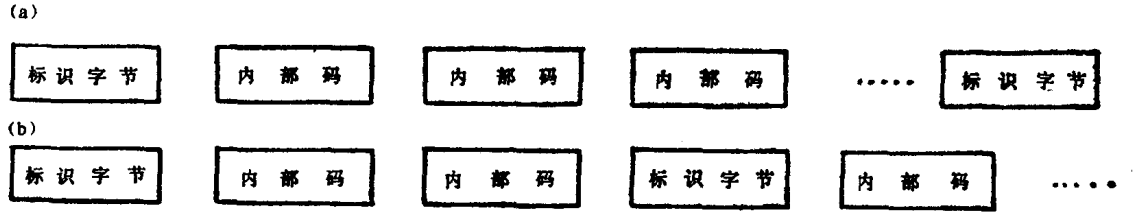

图 1-1　区分汉字和ASCII字符内部码的标识字节方式

这种方式的缺点是：对汉字文件进行插、删、改等编辑工作时，若不修改原编辑程序，便容易造成混乱，而且当ASCII码与汉字比较频繁地交替出现时，会浪费存储单元。

（3）全汉字方式。汉字内部码和所有图形及其他字符均用两个字节的“整数”表示，用该整数值范围来区分汉字和ASCII码。这种全汉字方式的优点是，表示形式整齐划一，内部码易于处理；缺点是，在处理以数字或西文字符为主体的文件时，需占用更多的存储单元，为此，在设计中西文数据库管理系统和应用程序时，应采取技术措施，使所管理的数字和西文字符仍能用原西文表示，其他汉字则使用两个字节的内部码表示。

值得指出的是，上述三种汉字和ASCII内部码的区分方法，均既不能保证原系统的西文软件不经任何修改便能中西兼容，也不能保证原西文软件的功能完全不丢失。因此，必须继续研究和完善这些区分方法，以使“中西兼容”系统的功能更加完善，含义更加确切。

三、汉字信息处理系统的基本组成部分

具有处理汉字信息能力的汉字系统，应至少由四部分组成，各部分之间的关系如图1-2所示。下面分别介绍这四个部分。

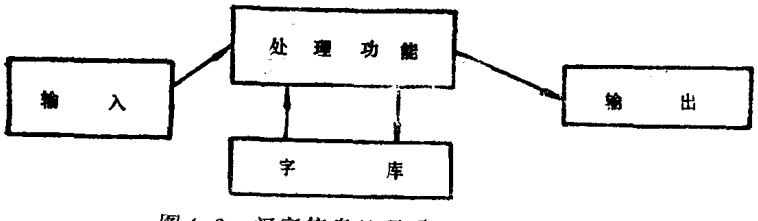

图 1-2　汉字信息处理系统的基本组成部分

（一）输入部分

输入部分的作用，是按各种编码规则将汉字和ASCII字符输入到计算机内部，并将输入码映射成内部码。好的输入方案必须具有输入速度快、编码规则简单和容错能力强等特点。

当前已研制出许多类型的输入方案，主要是自然语言识别、文字图形识别、键盘输入三大类。其中，自然语言识别类方案是让计算机直接“听懂”人的声音，而后按人的口头命令工作。由于种种原因，这种输入方案还处在初始研究阶段，远未达到实用程度。文字识别类方案则是将汉字（包括印刷体和手写体）作为图形，通过图形输入板、光学汉字阅读机等传感装置将汉字变换为数字信息，再经过信息处理、特征抽取、识别判断等一系列处理，最后将这些信息确定为特定的汉字。文字识别过程的框图如图1-3所示。

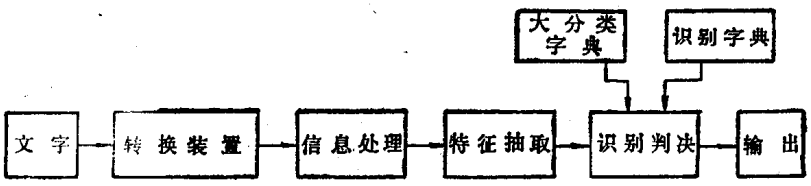

图 1-3　文字识别过程

图中，转换装置的主要作用是把文字图形转换为数字信息。通常用作这种装置的有图形输入板和光电转换设备，如飞点扫描、光导摄象管扫描、激光扫描等装置。信息处理的主要作用是消除信息中的噪音、改善信息质量和进行信息压缩。信息经处理后，由计算机抽取其中能反映文字本质形状的特征数据，如笔画的长度、角度、位置、分布以及背景等。识别判定是将抽取的特征数据与字典中标准汉字的特征数据进行比较，其特征数据与抽取的特征数据最相似的字典中的标准汉字就是所识别的文字。文字识别可对印刷体与手写体的文字进行识别。前者的识别率可达80%左右，已接近可应用水平；后者，则由于字形与标准字形相比变化太大，识别的准确率不高。因此目前还必须对手写体进行某些限制，以保证识别的精度。

目前，实用的汉字输入方法仍然是键盘输入，即通过击键将汉字的整字、字元或代表该汉字的数码和拼音字母输入到计算机中。键盘输入又可以按编码原则分为几类，这几类各有特长，分别适用于不同使用环境和人员。

一般，一个完整的汉字系统往往备有几种可供用户任意选择的输入方式。要求这几种输入方式相互之间的转换必须十分方便，以满足不同用户和使用环境的要求。

（二）字库

字库是汉字或字元点阵码有规律的集合。对于不同的要求，字库中所包含的字数也不同，但一般字库中应包含全部国标GB-2312中规定的一级和二级汉字和字符。某些特定的汉字系统，其字库中包含的字数可以更多，字体也可以有多种，而且简体字与繁体字并存。

由于字库要占用相当多的存储单元，字库中所包含的字往往不能全部存放在内存储器中，而要存放在外存储器中一部分，这样，存取速度太慢。目前处理字库存储问题所采用的方法有如下几种：

（1）采用固化字库。这种方法即将字库存入ROM或EPROM等芯片中。在这些芯片日趋便宜，容量越来越大的今天，采用这种方法，取出速度快，经济实用。

（2）采用分级字库。这种方法即将字库中的汉字分为常用字和一般字两级。常用字在系统初启时装入内存；一般字则常驻外存。对于内存容量较小的八位机，采用这种方法可以部分地缓和速度和容量之间的矛盾。

（3）采用RAM字库。这种方法是将字库全点阵码存于软盘中，在系统启动或初始化时一次调入内存。这种方法具有取出速快的优点，容量较大的16位机常采用这种方法。

（4）采用信息压缩技术。这种方法不存储整体汉字的点阵数据，只存储偏旁、部首及笔画等点阵，再加上由这些部件形成的所用汉字程序代码。采用这种方式，可以大大压缩字库所占用的存储空间；但形成汉字需要运行程序，这样便需要占用主机时间。所以，这种方法的实质是用时间换空间。

（三）输出部分

输出部分的作用是，在输出设备中将字库中的汉字点阵码恢复为汉字，并在显示屏上显示或按规定格式用打印机打印出来。要求输出设备输出速度快，并能确保字形美观不失真。

常用的显示设备有能显示图形的CRT终端和其它图形显示器。一般，显示器的象点越密，显示的字形越逼真，但显示设备的价格亦越高。因此，在选用显示设备时应权衡字形质量及设备价格。一般，图象密度为640H×200V或640H×400V就可以满足要求了。

打印设备采用9针、16针和24针的针式打印机。这种打印机价格便宜，但打印速度低，一般每秒钟不超过100个汉字。欲高速打印时，可以采用激光印刷机，这种印刷机每秒钟可以打印上万个汉字。

（四）信息处理功能

研究汉字系统的目的，在于管理各种中西文文件和处理各类中西文信息，因此，汉字系统必须具备较强的信息处理功能，一般应具有下列几种功能：

（1）中文文本编辑。此功能主要用于中文文本的编辑和排版。完成此功能需要各种编辑命令，其内容包括光标全屏幕移动，字、句、行、页的插、删、改，页的存储、调换、删除和合并。好的文本编辑应能行宽任选，屏幕窗口可左右、上下快速移动。此外，还应有编辑窗口定义、横编、竖编、字体设定和画线等功能。

（2）中西文文件管理。可以采用专用的文件管理系统或中西文数据库管理系统来执行此项功能。后者，以其数据相对于程序的独立性强、冗余度小、有强性、易于扩充且使用方便等优点，赢得越来越广泛的欢迎。

中西文文件管理系统应具有对文件的数据进行定义、修改、处置（包括满足多种条件的检索）的能力，并能提供操纵这些数据的专用语言，以及进行必要的统计、计算、报表打印等功能。

（3）用高级语言处理汉字信息。用高级语言处理汉字信息的常用方法有三种。一种为修改操作系统，使之能自动识别并分别处理汉字和ASCII码；另一种为修改语言的编译系统，使之能同时处理西文和汉字；第三种为不改动原来的任何软件，而仅提供高级语言处理汉字的软件接口（过程），即将处理汉字的软件接口嵌在其他高级语言程序中。

某些单位正在研制使用汉字编程的高级语言，这无疑将为没掌握英语的一般管理人员

提供许多方便。

（4）含汉字的图象处理。中文文件常常包含图象（或图形），而图象文件中又往往插入有汉字。因此，在研究汉字处理系统的同时，必须研究图象处理技术，以使各种图象能在中文文件的任何位置上出现，和能在图象的任何位置上标示汉字。只有这样才能做到图文并举，满足实用要求。

（5）可以进行数据通讯。

第三节　汉字信息处理系统的研究状况和课题

当前，成功地利用计算机进行汉字信息处理的途径，大致有如下几种：

（1）在原西文操作系统支持下，不改动原来的硬件和软件，新开发能同时处理中西文信息的软件。这种方法是以软件接口的方式来用高级语言处理汉字信息的。由于这种方法将原西文软件和新开发的西文处理软件混合在一起，故有人称之为“混合式”。这种方法不改动原来的系统软硬件，故可靠性高，不损失原系统的任何功能；更重要的是，采用这种方法时，不论操作系统的版本如何发展（一般是几个月或半年更换一次版本），所开发的中西文处理软件均可以不经修改便照常使用，从而大大减小了软件开发和维护的工作量。这种方法的缺点，是用原西文软件处理中文信息时，必须加上专门处理汉字的软件接口。

（2）根据处理中文信息所需的各种功能，局部或全面地修改（或重新设计）操作系统，比如修改操作系统的输入、输出处理部分，使之能同时适用于中文和西文的处理，这样原系统的一切西文软件，不经修改或稍加修改后便能同时处理中文信息。这种方法可以将中文处理能力溶于原操作系统中，故称之为“化合式”。从理论上讲，采用这种方法是实现中西兼容的较好途径。但如前所述，没有一种十全十美的区分汉字内部码与ASCII码的方法。因此，即使修改了操作系统，其原西文软件仍然要经相当多的修改才能用以处理中文信息，有时修改后原西文软件所具有的某些功能还要丢失。另外，采用这种方法时，要修改操作系统。因此，在修改后，操作系统的任何更新版本，一般都必须要重新修改，才能处理中文信息。这样便需要较高的维护技术，同时维修工作量也较大。

（3）将汉字库与汉字的输入输出软件纳入计算机终端，使汉字输入输出功能，全部由汉字终端和汉字打印机执行，原操作系统和高级语言的编译程序可以基本不作改动。对于大中型计算机和多用户计算机系统，采用这种方法可以使主机避免承担繁重的汉字转换和处理任务，从而可以获得更大的效益。采用这种方法的缺点，是需要对终端的硬、软件进行较大的改动，从而加大了工作量，同时系统的灵活性也较差。

尽管上述途径是可行的，但汉字信息处理技术在当前还远未成熟，必须进一步深入进行研究。为解决这方面的问题，需要认真做好下述工作：

（1）设计和生产具有中西文处理能力的新计算机系统和终端设备。这对发展我国计算机事业是必不可少的。

（2）对已有的计算机系统进行必要的技术改造，增加处理中文信息的能力。目前，我国引进的计算机数量多，品种杂，因此这项改造任务十分繁重。

（3）研究和开发中西兼容的计算机系统软件，这是一个十分广阔的领域，研究的内容非常广泛，其中包括设计和修改中西兼容的操作系统，研制和开发能力强、实用性好的

中西文数据库管理系统，设计功能强、使用方便的中西文文本编辑、字处理软件和各种排版、制表语言，研究和改造各种高级语言，使之能适用于中文信息处理，研究和开发含汉字的图象处理软件，研究中文信息的通信和中西文计算机网络等等。

（4）研究基础理论，如有关自然语言处理，特别是语声处理，以及文字和图象识别等基础理论。

（5）研究汉字系统的应用技术，将已具有中西文处理能力的计算机系统用于实际，如用于办公室的事务管理、厂矿企业的经营决策、情报检索、排版印刷、机器翻译、军事部门的军事指挥和通讯等。

第二章 字 库

汉字字库用来产生汉字字形和各种图形符号，类似于西文终端上的字符发生器；但由于汉字在字形数量和结构方面的复杂性远远超过了西文，而将用于产生或直接存储大量汉字字形和各种图形符号的部件称为“字库”。

汉字字库按结构和工作方式分类，大体上可分为模拟式和数字式两类，如图2-1所示。

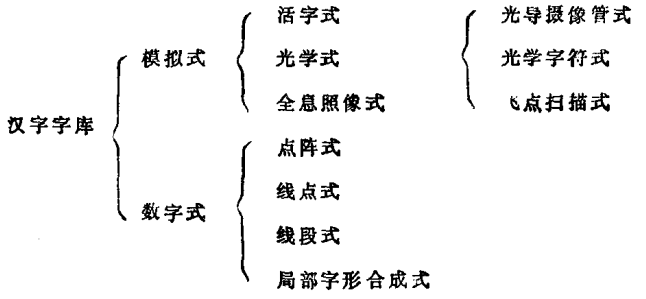

图 2-1 字库分类

在模拟式字库中，字形多存放在机械转盘和矩阵板上，采用光电扫描、偏转和摄像等手段来读取汉字。这类字库一般字形比较美观，但读取装置复杂，速度较慢，不适于一般计算机使用。

在数字式字库中，将字形看成是线或点的集合，并以数码形式将之存储于各类存储介质上。这种字库的优点是形式简单、存取方便，有利于计算机进行处理，目前各类计算机汉字信息处理系统多采用这类字库。本章的内容只限于讨论这类字库，介绍它的建立、使用和维护等，并评述其特点。

第一节 汉字字形的存储原理

将一个汉字写在一页坐标纸上（图2-2a），把被字形笔画覆盖或基本覆盖的小方格部分记为“全有”，其余的记为“全无”。这样，整幅图形便如图2-2b所示，变成一幅二维方格矩阵。将该方格矩阵中的每一个小格看成是一个坐标点，这样，由此构成的二维点阵便与计算机的存储器结构对应起来。其中，每一个坐标点对应于存储器中的一位，“有”记为“1”，“无”记为“0”。这样，整个字形就可以以二进制数据形式存储起来。

一个字形，可以通过逐点给出其信息或点坐标和线段起始坐标的方法来确定。前者给出了有用和无用的全部信息；而后者则仅仅给出了汉字图形部分，空白部分被压缩掉了。

逐点存储的方法又称为整字存储法。采用这种方法进行存储，过程简单而且字形美观、自然，但存储空间消耗极大。仅取字形特征进行存储的方法，称为笔画式存储法。其缺点是，实施比较困难，字形质量也受到一定影响，但具有能大幅度降低存储空间的消耗的突出优点。

整字存储法，是将每个汉字点阵看成是二进制位（bit）的集合，其数据结构可以描述成（以图2-2为例）：

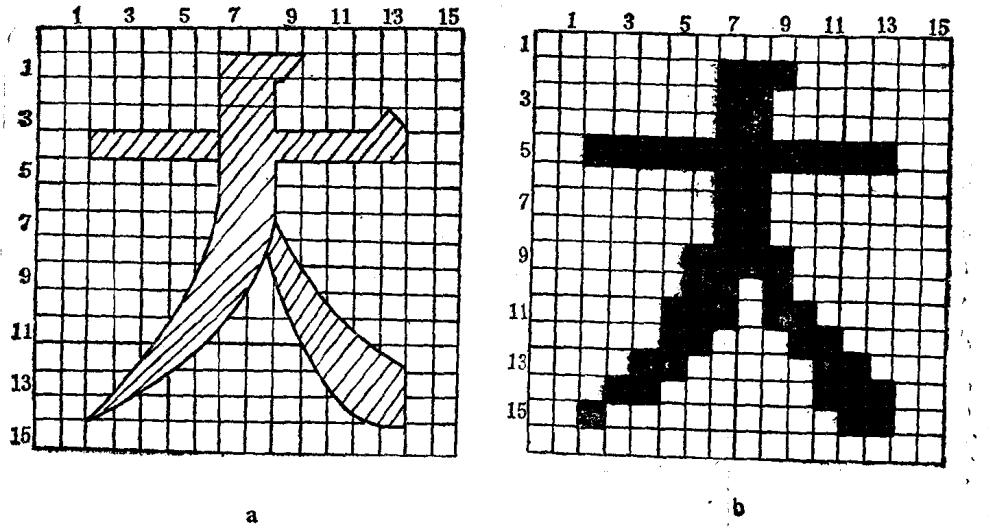

图 2-2 汉字字形的点阵表示

TYPE SET 15=SET OF 0··15;

WORD=ARRAY 〔0··15〕 OF SET15;

点阵中的一行描述成一个集合SET15；一个字有16行，由16个 SET 15 组成，每个SET15中的集合元素值标示该行中相应下标位置的取值。

如果以计算机中习惯使用的字节为单位，把横向一行分为前后两个字节，则16×16的点阵共需要32个字节，即

TYPE WORD ARRAY 〔0··31〕 OF BYTE;

如果将整个字库作为一个“文件”来存储，则每一个字形便是该文件中的一个“记录”。

借用两位十六进制数标识一个字节，并按从左到右，从上到下的顺序排列，图 2-2 所示的“大”字便可表示为0000，01C0，0180，0180，3FFC，0180，0180，0180，03C0，0340，0760，0630，0C18，181C，200C，0000。

笔画式存储法是把汉字看成是各种规格的点线集合，存储时只需把那些反映笔画位置和特征的信息保存起来。

等线体线段描述法是一种最简单的笔画式存储法。这种方法是将一个汉字看成为粗细均匀的直线段的集合，例如，用线段信息描述“大”字时，只需存储每一条线段的起点和终点信息，或起点、方向和长度信息。

如果笔画已扩充到折线和曲线，则除了起点和终点外，还要增加若干中间分量点和特征位。

在用线段描述时，一个字可以以下述形式描述：

T/PE LINE=ARRAY 〔0··3〕 OF INTEGER

WORD=SAPER ARRAY 〔1··*〕 OF LINE

其中，线段LINE由起点、终点的坐标确定。因为每个汉字的繁简不一，其组成线段的