名家讲演录

教计算机认字——汉字识别

JIJIAJIANGYANLU

Jiangyanlu

识别

别

实验系

上海科技

9059

教计算机认字
——汉字

吴佑寿 著

上海科技教育出版社

名家讲演录

教计算机认字——汉字识别

吴佑寿　著

策　　划　卞毓麟
责任编辑　卞毓麟
装帧设计　汤世梁

出版　上海科技教育出版社
　　(上海冠生园路393号　邮政编码200233)
发行　上海科技教育出版社
经销　各地新华书店
印刷　常熟市印刷二厂
开本　850×1168　1/64
印张　1.25
插页　2
字数　26000
印次　1999年8月第1版　2000年6月第2次印刷
印数　5 001－8 000
书号　ISBN 7－5428－1967－4/N·247
定价　3.80元

目 录

一、为什么要教计算机认字 1

二、计算机是怎样认字的 6

三、汉字识别的困难 12

字量大 笔画多 字根多

字型多 字体多 相似字多

四、笔式输入——联机手写汉字识别 23

IBM 的识别系统 汉王笔

文通笔 书写板及其他

五、光符阅读器——脱机汉字识别 41

脱机汉字识别原理

Casey 等的实验系统

北信印刷汉字识别系统

TH－OCR 印刷汉字识别系统

六、脱机手写汉字识别 67

七、让电脑"像"人脑那样识字 71

八、结束语 74

一、为什么要教计算机认字

文字是信息的载体，是人们表达、交流思想，传播知识、情报，保存资料、典籍的媒介。方块汉字已有数千年的历史，也是世界上使用人数最多的文字之一。对于中华民族灿烂文化的形成和发展，对世界文化和科学技术的影响，汉字都有着不可磨灭的功绩。

但是方块汉字也有着突出的弱点。它一字一形，结构复杂，而且字数繁多，字体不一，书写印刷都十分不便。在相当长的历史期

此为试读，需要完整PDF请访问：www.ertongbook.

间,汉字书写工具主要是称为"文房四宝"的笔墨纸砚。在人类逐步进入信息社会、通信非常发达、计算机广泛应用的时代,方块汉字的这些弱点就显得更为突出。摆在我们面前的问题是:大量的资料、文献、典籍需要整理、传送、利用或保存;许多书刊、文章、法规要求能自动检索查阅或翻译成其他文字;办公及管理自动化要求能及时编制、处理、传送各种文件、检索并获取有关情报资料,迅速作出预测与决策等等。此外,电子信函、电子出版、电子商务之类的应用也日益普及。凡此种种依靠落后工具和人工方法是难以办到的。如何利用当今科学技术的成果,特别是计算机等先进工具,使汉字这一人类的瑰宝继续并更好地发挥作用,就成为我们必需解决的问题。

汉字识别,通俗地讲就是教计算机"认

字”，其目的是把汉字自动转换成计算机内部的编码，以便于对汉字所携带的信息作进一步处理，如查询检索、提取摘要、出版、翻译、建立数据库等等。

电子计算机是西方国家发明并发展起来的。它的基础自然是西方的语言文字，其键盘也是由西文打字机衍变而来，对于习惯于用打字机打字的西方人来说，用键盘往计算机输入西文是很自然和简单的方法。在我国，情况迥然不同。我国虽然是活字排版印刷的发源地，但是用键盘“打”方块汉字，在相当长一段时间内却被人们视为“畏途”，有的人甚至认为方块汉字将要消亡，被拼音文字所代替！解决这个问题的关键就是研究计算机如何适应汉字信息处理的需要，使它不但能取代“笔墨纸砚”等传统工具，还能实现诸如信息交换、处理、存储、翻译等功能。在信

息高速公路迅猛发展的今天，这个问题的重要性已是不言而喻了。

用计算机对汉字信息进行处理的系统，大体分为汉字输入、汉字信息加工与处理，以及汉字输出三个部分。汉字输入是把方块汉字转换成为计算机便于处理的代码，这是汉字信息处理的基础，也是汉文信息处理系统的"瓶颈"。

近20年来我国学者在这些方面做了很多富有创造性的工作，取得了极其显著的成就，破除了方块汉字无法与计算机相结合的思想障碍。汉字键盘输入已可以和拼音文字的键盘输入相媲美；汉字计算机自动识别研究也已取得突破性进展，并在海内外推广应用；目前市场上出售的微机已普遍具有汉字输入和汉字信息处理功能；各种汉文数据库也不断建立；汉字照排系统已成功地用于各

种印刷系统，并远销海外。在基础研究方面，利用计算机对汉字字频加以统计并分析其分布规律，也取得了丰硕成果；关于词切分、书面汉语分析与理解、机器翻译等方面的研究工作也不断深入，特别是各种标准也已经或正在逐步建立之中，对汉字信息处理起着重要作用。

但是我们仍应看到，我们对汉语、特别是书面汉语的研究还很不够，还有很多问题和困难。尽管如此，方块汉字也不会被拼音文字所代替，不会消亡！只要我们继续努力，承载中华民族五千年灿烂文化的汉字必将继续发挥其作用，并再放异彩，为全人类作出更为辉煌的贡献。

二、计算机是怎样认字的

计算机认字的原理很简单：在计算机中有一个“字典”，将待识的汉字与字典中的每个标准汉字逐个相比较，和待识字相同的标准汉字就是待识字。

图1是计算机认字原理示意，其中的“字典”通常叫做特征库或模板库。建立字典的方法是：先将已知的标准汉字库中的汉字输入计算机，逐一抽取能代表每一个字的特征或模板，组成特征库(模板库)。这一过程叫

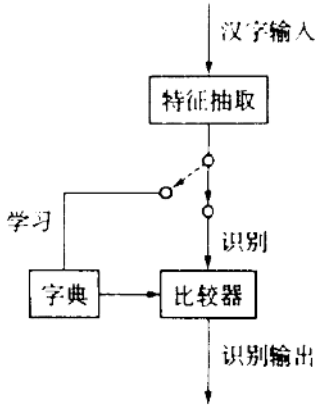

图 1　计算机认字原理

此为试读,需要完整PDF请访问: www.ertongbook.

做训练或学习过程。计算机认字时，把待识的汉字输入后也抽取这个汉字的特征，然后把它和特征库中每一个标准特征逐一比较，与待识汉字特征相同(或最相似)的标准汉字就是待识汉字。这一过程叫做识别过程。

用来识别汉字的特征，基本上分为结构特征和统计特征两大类。特征库是识别系统的核心，所采用的特征的优劣是决定识别系统性能的关键。这个问题在后文还将详细讨论。

汉字识别系统通常分为手写汉字识别系统和印刷汉字识别系统两大类(图 2)。按输入方式不同，手写汉字识别又可分为联机和脱机两种。联机手写汉字识别采用一块专用的与计算机连接的书写板，人在书写板上书写字符时，字符信号即时直接输入计算机。这种方法也叫笔输入方式。脱机手写汉字识

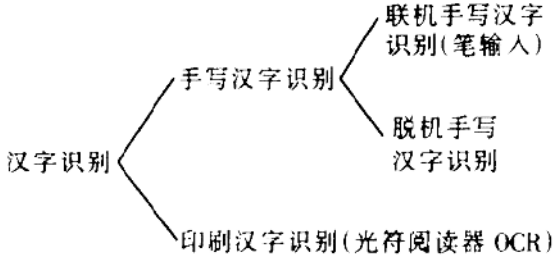

图 2　汉字识别的分类

别是把汉字先写在一般的纸张上，用光电转换器(如扫描仪或摄像机)把字符图形变为电信号，输入计算机进行识别。脱机印刷汉字识别同样使用这种装置。它称为光符阅读器(Optical Character Reader)，简记为 OCR。

衡量识别系统优劣的主要性能指标是正确识别率和识别速度；从实用角度看，还应考虑系统的复杂性、可靠性和价格等等。

对识别系统识别率和识别速度的要求，很难有一种统一的、严格的标准，主要根据实际应用的需要来确定。但是作为一种输入手段，它的性能应该可以和其他输入手段(如人工键入)相比拟。目前专业人员操作的汉字键入错字率约为 10^{-3}～10^{-2}的量级，键入速度最高达 200 字/分，平均速度在 50 字/分左右。作为参考，这些指标应该是汉字识别系统必须达到的最低要求，在某些需要大量输

人的场合(如数据库的建立),对识别系统性能的要求还应更高。至于联机手写汉字识别的速度主要受人书写速度的限制,一般的识别速度约为每秒钟1~2个字。

这本小册子主要是讨论汉字识别的原理、方法及应用。为此,我们先扼要介绍有关汉字的一些基本常识和汉字识别的困难,这对较深入地讨论汉字识别问题是有益的。

三、汉字识别的困难

汉字集合的字量大、字形复杂。这是汉字识别之所以十分困难的根本原因。

字量大

按照我国汉字基本集 GB2312－80 的标准，第一级汉字为 3755 个，第二级为 3008 个，总计为 6763 个。因此，我国的汉字识别系统至少应能识别最常用的 3755 个汉字。如果考虑还能识别常用的二级的 3008 个汉字，并能用于我国香港和台湾等地区，则识别字量应

是6763个简体字和5401个繁体字之总和，共一万多个汉字。也就是说，识别系统的"字典"至少必须有一万多个汉字，以及相应的标点符号和一定的英、日文字母，才能满足实际应用的需要。

实际上汉字识别系统的字典标准模板的数目，比上述所说的汉字字量还要多。这是因为印刷汉字有多种字体，常用的有宋体、仿宋体、黑体和楷体，以及魏碑、小篆及其变体与变形等等。不同字体的同一个汉字的拓扑结构虽然相同，但它们的点阵图形却不完全一样。目前计算机的智能不高，往往不能适应这种变化，不能直接从拓扑结构相同与否来确定它们是不是同一个汉字，而把不同字体的同一个汉字看作是不同的字。实践表明，在各种印刷字体中，楷体汉字的点阵图形和其他字体的差别最大。能兼容宋体和黑体

此为试读，需要完整PDF请访问：www.ertongbook.