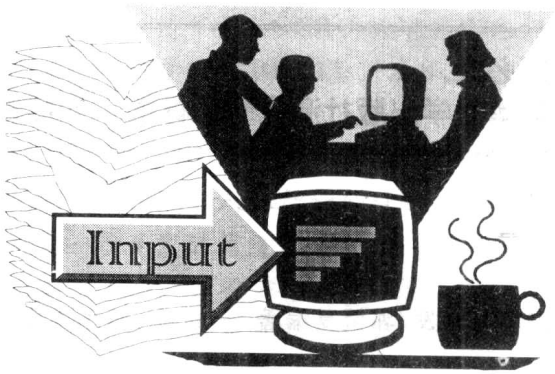
清华TH-OCR®技术应用与开发

郭繁夏 编著

清华大学出版社

清华TH-OCR®技术应用与开发

郭繁夏 编著



清华大学出版社

(京)新登字158号

内 容 提 要

本书从应用和进一步开发的角度，介绍了得到普遍关注的清华TH-OCR高性能中英文印刷文本自动识别输入系统的技术特点、应用方法和开发标准。全书分为3篇：上篇是清华TH-OCR技术简介，介绍了有关清华TH-OCR技术的基本内容；中篇是清华TH-OCR应用指南，以图文对照的方式，提供了清华TH-OCR完整的操作指南；下篇是清华TH-OCR进阶开发，深入介绍了清华TH-OCR的系统特点和使用技巧，并给出了在清华TH-OCR核心技术基础之上，进一步开发广泛应用系统的标准规范和实用范例。

本书适合于高等院校师生、计算机工程技术人员、计算机应用开发人员和所有对清华TH-OCR感兴趣的读者使用，并可以作为清华TH-OCR高性能中英文印刷文本自动识别输入系统(Windows版本:NT及NS)的使用手册。

图书在版编目(CIP)数据

清华 TH-OCR 技术应用与开发/郭繁夏编著．—北京：清华大学出版社，1996
ISBN 7-302-02138-4

Ⅰ．清… Ⅱ．①郭… Ⅲ．①文字自动识别输入系统-软件开发
②OCR-软件开发-基本知识 Ⅳ．①TP391．4②TP335

中国版本图书馆 CIP 数据核字(96)第 04106 号

出 版 者：清华大学出版社（北京清华大学校内，邮编 100084）
责任编辑：刘明华
印 刷 者：北京市海淀区清华园印刷厂
发 行 者：新华书店总店北京科技发行所
开　　本：787×1092　1/16　印张：9.75　字数：229 千字
版　　次：1996 年 4 月第 1 版　1996 年 4 月第 1 次印刷
书　　号：ISBN 7-302-02138-4/TP・1009
印　　数：0001—5000
定　　价：14.00 元

前　言

人类社会已经开始进入信息时代，“信息高速公路”热潮，席卷全球。文字是人类信息和文化最为集中最为重要的表现，在信息化的过程中，延续和发展人类文明的各种文字记录都面临着“电子化”——输入的迫切要求。因此，汉字的自动识别输入技术，有着极其广泛的应用前景。

清华TH-OCR高性能中英文印刷文本自动识别输入系统是清华大学电子工程系在“863”高科技计划、“七五”科技攻关计划、自然科学基金和军事基础研究课题的支持下，从1985年开始经十多年的科研取得的成果，整体性能居“国际领先水平”。曾多次获得国家级和省、部级奖励，1994年还被评为全国十大电子科技成果。

目前，清华TH-OCR的专业版以每年约5000套的数量销往全世界，其中在国内约3000余套，并呈现出增长的趋势；清华TH-OCR的标准版则已经授权国内外多家扫描仪厂商随扫描仪一起发行，年发行量超过10000套。清华TH-OCR作为863的高科技产品，在国内中文OCR市场的占有率高达65%以上，在海外更是独领风骚，充分显示出高科技的无穷力量。

OCR技术在学科上属于模式识别和人工智能的范畴。本书面向广大计算机用户和技术开发人员，从应用和进一步开发的角度，介绍了得到普遍关注的清华TH-OCR系统的技术特点、应用方法和开发标准。全书共分为3篇：上篇是清华TH-OCR技术简介，介绍了有关清华TH-OCR技术的基本内容；中篇是清华TH-OCR应用指南，以图文对照的方式，提供了清华TH-OCR完整的操作说明；下篇是清华TH-OCR进阶开发，深入介绍了清华TH-OCR的系统特点和使用技巧，并给出了在清华TH-OCR核心算法和技术基础之上，进一步开发广泛应用系统的标准规范和实用范例。本书上篇的有关内容，适合于清华TH-OCR系统的各个版本；中篇和下篇的内容可以作为“清华TH-OCR NT for Windows”、“清华TH-OCR NS”及其后续更新版本的用户手册。

在本书的写作过程中，得到了丁晓青教授、实验室同仁和清华文通公司的朋友们的大力支持和帮助，谨表示由衷的谢意！需要强调说明的是，本书第11章的部分内容，由实验室的陈明提供，在此，作者表示衷心的感谢！

本书的作者长期从事中英文OCR理论和技术的研究开发，但由于写作经验不足，缺点和错误在所难免，欢迎广大朋友批评指正。

郭黎夏

1996年1月于北京清华大学

目 录

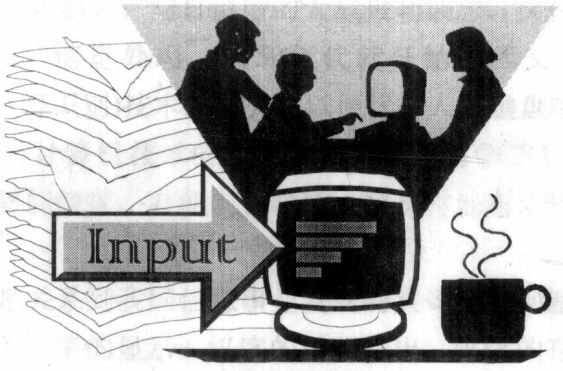
上篇 清华TH-OCR®技术简介

中篇 清华TH-OCR®应用指南

下篇 清华TH-OCR®进阶开发

上 篇

清华TH-OCR®技术简介



第1章 OCR技术与系统概述

1.1 OCR技术的发展背景

人类社会已开始进入信息时代，各类信息事业的发展将极大地影响国家的发达和民族的兴旺，因此，世界各国对信息事业和产业的发展都给予了极大的重视和关注，目前席卷全球的兴建“信息高速公路”的热潮，就是一个明证。文字是人类信息和文化最为集中最为重要的表现，信息化过程中，延续和发展人类文明的各种文字记录都面临着“电子化”——输入的迫切要求，从而实现信息的计算机处理和电子技术通讯等等。

欧美国家为了将浩如烟海、与日俱增的大量报刊杂志、文件资料和单据报表等文字材料输入计算机进行信息处理，从50年代就开始了西文OCR(Optical Character Recognition，即光学字符识别)技术的研究，以便代替人工键盘输入。经过40多年的不断改进和完善，并随着计算机技术的飞速发展，现已将OCR技术广泛应用于各个领域，使大量的文字资料能快速、方便、省时省力和及时地自动输入计算机，实现信息处理的“电子化”。

汉语计算机处理是关系到我国信息事业发展的头等重要问题，从汉字操作系统、汉字字库到汉卡等有关汉字的显示、汉字的排版输出、汉字的存储、检索等等都围绕着计算机的汉化而进行，而汉字的计算机输入问题则是计算机汉化首先要解决的问题。数百种汉字编码方案的提出和实现，解决了用小键盘输入成千上万汉字的人工键入的问题，但这是相当费人费时且只适用于专业输入人员的。汉字输入的困难，无疑已成为计算机普及的“拦路虎”。因此，深入研究中文OCR技术，解决汉字的计算机自动识别输入，是解决大量汉字输入的极为关键、极具战略地位的问题。换句话说，不解决好汉字自动输入的问题，即使修建好四通八达的“信息高速公路”，也难有满载文字信息的车辆在此“信息高速公路”上奔驰。

1.2 中文OCR技术简介

1.2.1 什么是中文OCR

汉字是形、声、义有机组合的方块文字，其特点是数量浩大(常用简体汉字在

4000—7000个以上，繁体汉字超过10000个)、结构繁杂、字体字形变化多端。将汉字输入计算机，无论是人工键盘输入，还是利用OCR技术自动识别输入，都是十分困难的。常用的汉字计算机输入方法见表1.1。

表 1.1 汉字计算机输入方法

<table>
<tr><td rowspan="2">人工键入</td><td colspan="3">大键盘输入</td></tr>
<tr><td colspan="3">计算机小键盘编码方案(如拼音、五笔等200余种)</td></tr>
<tr><td rowspan="4">自动输入</td><td rowspan="3">汉字识别(中文OCR)</td><td colspan="2">印刷汉字识别</td></tr>
<tr><td rowspan="2">手写汉字识别</td><td>联机手写汉字识别</td></tr>
<tr><td>脱机手写汉字识别</td></tr>
<tr><td colspan="3">语音识别</td></tr>
</table>

我国政府对汉字自动识别输入的研究从80年代开始就给予充分的重视和支持，经过科研人员十多年的辛勤努力，印刷体汉字识别和联机手写汉字识别(手写板)技术的发展和应用，已受到世人瞩目。这一成就，是对中华文化宝贵遗产的继承和发扬，在世界电脑发展史上，必将留下光辉的一页，这也是造福子孙千秋万代的大事。国家高技术研究发展“863”计划、国家重点科技攻关计划、国家自然科学基金和军事基础研究基金都对这一研究课题予以极大的重视和大力的支持，本书主要介绍有关印刷体汉字识别方面的内容。中文OCR系统的基本流程如图1.1所示。

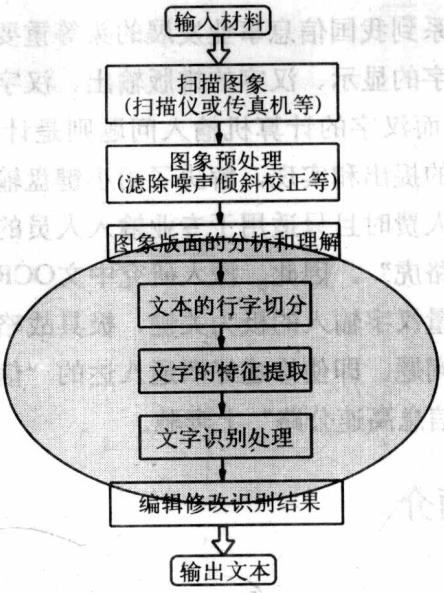


图 1.1 中文OCR系统的基本流程

由此可见，中文OCR技术是可以用来代替人的大脑和双手，完成汉字自动输入的一种模式识别技术。汉字识别的过程简单来说，是先将文本经过扫描仪扫描，进行光电转换得到图象信息；然后利用汉字识别技术，将文本文字的图象信息转化为计算机可以直接处理的文字代码形式，完成文本的计算机自动输入。

中文OCR技术主要包括：

1　扫描输入图象。

2　图象的预处理。

3　图象版面的分析和理解。

4　图象的行切分和字切分。

5　基于单字图象的特征选择和提取。

6　基于单字图象特征的模式分类。

7　将被分类的模式赋予识别结果。

8　识别后处理。

其中4，5，6，也就是图1.1中的阴影部分，是中文OCR的核心技术。近几年来，中文OCR系统的单字识别正确率已经超过95%，为了进一步提高系统的总体识别率，扫描图象、图象的预处理以及识别后处理等方面的技术，也都得到了深入的研究，并取得了长足的进展，有效地提高了中文OCR系统的总体性能。

中文OCR的研究工作，是十分困难的理论和技术问题，主要包括两个方面的要求：一是识别方案本身，对数量繁多的中文文字是否具有足够识别能力，也就是信息量的问题；二是识别系统是否具有很好的稳定性和可靠性，以满足实际应用中千变万化的不同需求，也就是应用的问题。

1.2.2　为什么要使用中文OCR系统

在当今信息化的社会里，大量的信息加工和处理，如文件资料的存贮、整理和利用，各种中文资料库、档案库的建立，书籍再版，情报检索，通讯以及文字的自动翻译等等都需要计算机来完成。而这一切，首先需要将汉字输入计算机。特别是计算机应用的日益普及，计算机信息处理的速度逐年提高，具有海量存贮量(高达数百兆)的光盘和高速输出汉字的激光打印机等的出现，都使得汉字输入的问题变得格外突出。

中文OCR系统的用途，主要包括以下八个方面：

第一，建立中文文字资料库。建立中文文字资料库时，需要大量输入各种各样的中文书籍、杂志和报刊等等。使用中文OCR系统，可以自动实现文字的输入，缩短建立资料库的时间，大大节约人力物力。

第二，出版社、杂志社和报社等的书刊重印。使用中文OCR系统，自动输入已经出版的文章和印刷文件，稍加修改即可编辑成新文章或成为再版书刊。中文OCR系统输出的标准中文文本文件，可以用于各种排版系统，重新印刷输出。

第三，办公自动化和情报检索。国家机关、工矿企业和公司等各个单位日常办公用的文件资料，都可以使用中文OCR系统，输入计算机，以备存储和检索。

第四，图书馆的资料输入。使用中文OCR系统输入图书的相关资料，可以建立书目数据库或输入经典图书全文，将纸上的文字转换为计算机的文本，查询检索都十分方便，并可以让专家用计算机处理研究书籍内容。

第五，智能汉英翻译系统中的中文输入。智能机器翻译的第一步是将全文输入计算机，使用中文OCR系统进行文字的自动输入，然后再应用翻译系统，可以实现高速的自动翻译。

第六，个人资料的输入。作家及其他人员写文章，可以使用中文OCR系统自动输入自己或别人已经发表的文章，参考其部分或全部内容。例如，教师出考题，可以用中文OCR系统输入参考试题等等；此外，个人的信件、传真等，也可以用中文OCR系统，事先输入大量的参考文件，再经过计算机的编辑修改，打印输出或者直接在计算机上通过网络通讯传输。

第七，自动阅读机和盲人阅读机。将中文OCR系统和语音合成系统结合起来，能够形成自动阅读机或盲人阅读机。

第八，中文文字材料的压缩存储和传输。使用中文OCR系统输入中文文字所形成的文本文件，比扫描得到的图象压缩100倍以上，这是目前任何别的压缩工具所无法达到的，大大节省了存储容量，加快了传输速度。

第2章　清华TH-OCR®系统简介

2.1　清华TH-OCR系统的研究与开发

2.1.1　汉字识别的主要困难

汉字是人类最古老的文字之一，也是使用人口最多的文字。汉字字数众多，字型变化复杂，是任何其他民族的文字所不能比拟的。高性能实用汉字识别系统的研究是基础理论研究和应用技术研究的结合。它涉及数字信号处理、图象处理、模式识别、人工智能、自然语言理解、编码理论、数据结构、算法理论与计算机编程等许多学科。传统的模式识别分为结构(句法)模式识别与统计模式识别两大类，这两种方法应用于汉字识别时各有优缺点。单纯的结构方法以语义文法为基础来描述汉字的结构关系，难以适应汉字结构的变化，易受噪声干扰；单纯的统计方法完全将汉字简单地看成二维图象，没有考虑到汉字结构的分布规律，不能很好地反映汉字复杂的拓扑结构。由于汉字是具有笔划结构的二维图形，目前对汉字识别的研究，常常采用结构与统计两种方法的结合，即采用基于汉字结构的统计决策方法。

清华TH-OCR系统将结构与统计两种方法结合起来，较全面地提出并综合利用各种处理技术和算法，解决了具有较高性能的印刷汉字识别问题。

最早对印刷体汉字识别进行研究的是IBM公司的Casey和Nagy。1966年他们发表了第一篇关于汉字识别的文章，用模板匹配法识别1000个印刷汉字。70年代以来，日本学者做了许多工作，其中有代表性的系统有1977年东芝综合研究所研制的可以识别2000汉字的单体印刷汉字识别系统；80年代初期，日本武藏野电气研究所研制的可以识别2300个多体汉字的印刷体汉字识别系统，代表了当时汉字识别的最高水平。此外，日本的三洋、松下、理光和富士等公司也有其研制的印刷汉字识别系统。这些系统在方法上，大都采用基于K-L数字变换的匹配方案，使用了大量专用硬件，其设备有的相当于小型机甚至大型机，价格极其昂贵，没有得到广泛应用。

我国对印刷汉字识别的研究始于70年代末、80年代初。考虑到国情，我国的汉字识别系统均在微机上实现。近年来，我国汉字识别的研究发展很快，手写印刷体汉字识别的研究正在逐步深入；联机手写汉字识别已初步实用化，并将在新的条件下掀起新的高潮；最引人注目的则是印刷汉字识别的研究，目前已初步推广、应用，特别是

近三四年以来，已有五六个系统脱颖而出，并占领了一定的市场，进入实际应用。这些系统的综合指标为:

1 识别字数: 4000左右。

2 识 别 率: 对中等印刷质量的文本96%—99%。

3 识别速度: 10—30字/s(486微机)。

4 识别字体: 宋、仿宋、黑、楷及相应繁体。

5 识别字号: 6号以上。

6 具有一定的版面分析和后处理能力。

同所有的模式识别系统一样，汉字识别的主要性能指标是正确率和识别速度，其中正确率尤为重要。汉字识别的研究目前被认为是字符识别中最为困难和复杂的问题之一。其主要困难在于:

其一，汉字字数多。我国公布的国家标准信息交换汉字编码GB2312-80中规定的汉字共6763个，其中一级汉字3755个，使用频度为99.7%; 二级汉字有3008个，两级汉字累计的使用频度为99.99%。这给汉字识别带来巨大的困难。从模式识别的角度来看，这是一个超多类模式识别的问题，理论上和技术上难度都很大。

其二，汉字结构复杂，字体多变。汉字图形结构非常复杂，其平均笔划数是英文字母的十倍以上，约有半数的汉字在13划以上，而对于同一汉字，同一笔划，由于字体不同，也存在很大差异。笔划多、结构复杂、字体多变，这也给汉字识别带来很大困难。

其三，汉字中存在较多的形态上相似的字，易受干扰影响。统计表明，大约有3%的汉字存在类似“己－已－巳”的相似字情况，对噪声干扰极为敏感。如“大”字的下面如果有一个噪声黑点，就会成为“太”。由于这种较差的抗干扰性，给汉字的正确辨识带来困难。

作为高性能的实用印刷汉字识别系统，除了遇到上述的字数多、结构复杂和字形相似三方面的困难外，还面临着以下五个方面的难点:

一是系统的广泛适应性问题，包括对于不同字体，不同印刷方式和不同印刷质量的适应性。目前常见的印刷文本材料，有铅印、激光照排、激光打印机打印、电子打字机打印、普通针式打印机打印及其复印件和胶印件，此外，还有经打印蜡纸后的油印件，等等。由于印刷方式不同，不仅字模存在差异，印刷质量也千差万别，如何对不同印刷质量，不同印刷方式(尤其是打印件、复印件)保持较高的识别率，是高性能实用印刷汉字识别系统的首要问题之一。

二是各种复杂版面的印刷文本的识别问题，包括横、竖排文本和各种汉字、字符