中文信息处理丛书

汉 字 识 别 技 术

张　中　著

清 华 大 学 出 版 社
广西科学技术出版社

内 容 简 介

本书全面地论述了汉字识别技术的原理、方法和系统。分别讨论了汉字识别的概念、汉字字形、汉字识别的原理和一般方法、汉字识别的预处理技术、联机手写汉字识别、印刷体汉字识别、手写印刷体汉字识别、汉字系统的构成，并介绍了五个具体系统。

这是中国第一本汉字识别技术的专门书籍，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，也可作为从事文字识别和中文信息处理研制、生产、开发、使用部门研究与技术人员的参考。对国内外研究汉字识别技术有促进作用。

(京) 登字 158 号

汉 字 识 别 技 术

张 中 著

*

清华大学出版社出版

北京 清华园

北京航空航天大学印刷厂排版

???? 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: 11 字数: 260 千字

1992 年 月第一版 1992 年 月第一次印刷

印数:

ISBN 7-302-01088-9/TP・407

定价:

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　(按姓氏名划排列)

王鼎兴　杨芙清

李三立　施伯乐

徐家福　夏培肃

董韫美　张兴强

徐培忠

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
广西科学技术出版社
计算机学术著作出版基金办公室

1992 年 4 月

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，越来越起着更为重要的作用，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，总数多达五万余。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,需要我们自己去解决。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都属于中文信息处理领域。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科学的分类，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，有些技术已经产业化，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

近年来，模式识别技术随着计算机的迅速发展不断取得新的进展，改善了人机之间的信息交互能力。文字识别是模式识别的一个重要分支，汉字识别是文字识别中最困难的部分。通俗地说，汉字识别是用计算机自动辩识印刷在纸上和人写在纸（或介质）上的汉字。学科上属于模式识别和人工智能的范畴，应用上是中文信息处理系统高速、自动输入的手段。汉字识别还是新一代计算机智能接口的一个重要组成部分。汉字识别涉及到模式识别、图像处理、人工智能、形式语言和自动机、模糊数学、组合数学、信息论、计算机、中文信息处理等学科，也涉及到语言文字学、心理学、仿生学等，是一门综合性技术。近十年来，国内外对英文、日文字母和汉字等文字的识别技术进行了认真、细致地研究，学术上百花齐放，完成了一批实用的识别装置和应用软件。

本书是作者在新加坡大学讲学的讲稿基础上加工整理、充实完善而成。书中内容包括作者和作者所在科研组研究工作的总结。还包括几年来作者收集同行的科研成果。

本书前四章主要讲述汉字识别的基本原理、方法和预处理技术，中间三章分别讨论了三种类型的汉字识别：联机手写汉字识别，印刷体汉字识别和手写印刷体汉字识别，最后一章介绍了汉字识别系统的构成（包括版面分析和识别后处理），并列举出五个不同类型的具体汉字识别系统，这些系统都已有了产品，有了用户，使用情况良好。

本书可作为和模式识别、人工智能、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，也可供从事文字识别和中文信息处理技术工作的研究、技术人员参考。

中国中文信息学会理事长陈力为教授在百忙中仔细审阅了全书，提出不少宝贵的改进意见，新加坡大学赖金定博士在本书写作、出版过程中给予了全力支持和帮助。中国中文信息学会基础理论专业委员会委员廖寿琪研究员对本书提出了宝贵建议并做了大量编审工作。本研究组的沈兰生、刘秀英、李燕、金田坤、嵇文汶、谢森、程公先等提出了宝贵意见并热情帮助做了许多工作。在出版过程中，得到国内外有关专家、同行的热情支持。在此，对以上提到的各位先生及对本书的工作提供过帮助的所有人员一并表示感谢。

由于汉字识别技术正在日新月异地发展，目前中国尚无专著，加上作者水平所限，因而本书无论从内容选择，还是学术观点，一定存在不少缺点或错误，敬请广大读者批评指正。

目　录

第一章　汉字识别概述

汉字识别(Chinese Character Recognition,简称 CCR)通俗地说,是用电子计算机自动辨识印刷在纸上和人写在纸(或介质)上的汉字。学科上属于模式识别和人工智能的范畴,是文字识别技术的最高蜂;应用上是一种汉字信息处理系统中高速自动输入方式。汉字识别也是新一代计算机智能接口的一个重要组成部分。

汉字识别技术,涉及到模式识别和图象处理,人工智能,形式语言和自动机,统计决策理论,模糊数学,组合数学,信息论,计算机,汉字信息处理(Chinese character information processing)等学科;也涉及到语言文字学、心理学、生物学等,是一门综合性的技术。

1.1　汉字识别研究范围和应用领域

1.1.1　汉字识别研究范围

汉字识别研究范围可以用图 1.1 形象地表示出来,它构成了三维空间。按识别文字类型包括联机手写体汉字识别(On-line handwritten Chinese character recognition,简称 OLCCR),单体印刷体汉字识别(Single-font printed Chinese character recognition,简称 SPCCR),多体印刷体汉字识别(Multi-font printed Chinese character recognition,简称 MPCCR),手写印刷体汉字识别(Handprinted Chinese character recognition,简称 HCCR),手写行书汉字识别(Handrun Chinese character recognition)等;按识别文字和版面质量包括高、中、差三种;按文字数量和版面复杂程度包括一级、二级汉字、10000 汉字和简单、中等、复杂三种版面。显然,离开原点愈远,研究的难度愈大。

图 1.1　汉字识别的研究范围

目前汉字识别包括以下类型。

1. 联机(或实时或在线)手写汉字识别(OLCCR)用笔在图形输入板上写字,人一面写,机器一面认,是一种方便的汉字输入手段,也是汉字识别中最简单的一种类型。

2. 单一印刷体汉字识别(SPCCR)识别印刷在纸上的某种印刷体(一般为宋体)或某种打印机、照排机输出的汉字。

3. 多种印刷体汉字识别(MPCCR,印刷体汉字识别简称为 PCCR)。同时能识别印刷在纸上的宋、仿宋、黑、楷等体以及若干典型的打印机、照排机输出的汉字。

4. 手写印刷体汉字识别(HCCR)识别人写在纸上的规整汉字,要求楷书、笔划数正确、书写在方格中等。书写用纸和笔也不能任意。

5. 确定人手写汉字识别(Dependent handwritten Chinese character recognition, DHCCR)是手写体汉字识别的一个特例。但从书写汉字的规整性而言,限制要比手写印刷体识别松,允许书写部分行书。笔迹鉴别也属于此类。

以上五类,除第一类是联机识别外,其余都是脱机识别。

1.1.2 汉字识别应用领域

汉字识别研究介于基础研究和应用研究的边缘,它的研究对加速建立汉字信息库,对汉字信息处理系统(包括办公自动化)全自动化,对开拓新一代计算机的智能输入都有重要意义。汉字识别应用范围有:

1. 使汉字高速自动输入计算机,解决汉字信息处理系统中手动输入效率低这个关键问题。随着计算机技术的发展,汉字信息处理系统处理和输出汉字的高速度(如用激光印字机输出汉字,一秒钟可高达 1000 个),越来越和用手工操作的低速汉字输入(平均每秒输入一个汉字)产生矛盾,使得汉字输入计算机成为解决整个系统效率的“瓶颈”。代替手动输入汉字的自动输入方法,虽然有汉字(字形)识别和汉语语音识别两种,但是,使汉字高速输入计算机,在原理上能与汉字输出速度相匹配,从目前看,汉字识别是唯一的方法。

2. 是办公自动化和建汉语语料库中不可缺少的文字自动输入设备的基础和便于输入(联机识别)的手段,也是建立在自然语言理解基础上的自动翻译的理想输入方法。

3. 是智能计算机智能接口的组成部分。智能计算机是在更高的程度上,更完善地模拟和取代人类部分脑力劳动的全新一代的计算机。智能计算机能认识文字、图形和景物,能听懂语言,能理解文章……。视觉是智能计算机接受外界信息的重要手段。随着文献、资料、统计报表等逐年增加,对文字信息识别的智能接口日渐重要。

4. 汉字图象经识别后形成代码,信息量压缩了 100 倍以上,对汉字信息压缩、传输有重要意义。

5. 联机手写汉字识别是一种很方便的汉字输入方法。是在各种自动识别输入的方法中,能够代替或部分代替人工编码输入的唯一可能的方法。笔迹鉴别 Script identification 以及利用汉字识别技术制成的自动阅读机(或盲文阅读机)等,对扩大计算机在国民经济各部门的应用有实际意义。

1.2 汉字识别研究的历史回顾和现状

随着模式识别和人工智能研究的进展,在英文、数字识别的基础上,六十年代开始对汉字识别进行研究,七十年代出现了初步成果[1-1],[1-2],[1-3]。最近十多年,各国都进行了大量的研究工作,并已经取得了不少成果[1-4]。

以当前在汉字识别方面居于世界前列的日本为例,约在七十年代开始对印刷体汉字识别进行研究。饭岛[1-5],[1-6]、中野[1-7]、山本[1-8]、坂井[1-9]、河田[1-10]等在 1973 年前后发表了一批汉字识别的论文,到 1977 年,完成了日本通产省制定的"图像信息处理系统"中印刷体汉字的识别装置,并于 1980 年 10 月进行了公开表演[1-11]。该装置可识别 2000 个汉字,识别速度为 100 字/秒,识别率达到 98.4%。1984 年,日本研制成识别 2300 汉字的多体印刷汉字识别装置,识别率为 99.88%(专用 OCR 纸),识别速度大于 100 字/秒,代表了当时印刷汉字识别的水平[1-12],[1-13]。机器采用中小型机,价格昂贵。从七十年代中期始,手写印刷体汉字识别,在日本也开展起来,进入八十年代,研究工作日趋活跃[1-14]～[1-23],最近几年,有少数使用高档微机的印刷、手写印刷日本汉字装置出现,识别字数为 2000～3000 汉字,识别速度为 5～40 字/秒,识别率对印刷体可达 98%～99%,对手写体> 90%,见表 1.1。日本联机手写汉字识别装置对不允许笔划数改变的楷书已有产品出售,对允许笔划数改变的手写行书联机识别正在研究[1-24],见表 1.2。两表中识别率都是在研制时良好样张条件下取得的,比实际识别率高。

中国从七十年代开始进行主要用于邮政信函分检的数字识别和计算机输入用的英文、数字、符号识别,七十年代末,一些大学和研究所开始对印刷体和手写印刷体汉字的识别进行原理性研究,至今已有 12 年历史。前 6 年,少数单位少数人进行识别方法的探索;后 6 年,是中国汉字识别研究的丰收期。其中,从 1986 年初到 1988 年是印刷体汉字识别和联机手写汉字识别研究的丰收期;从 1988 年到目前是印刷体和联机手写汉字识别实用系统的研制和初步实用期,也是手写印刷体汉字识别研究的高潮期[1-25]～[1-30]。联机手写汉字识别已经研制出了几个初步实用的装置,其指标为:识别字数 6763～12000 字,识别率初次使用为 80%左右,经常使用可以达到 95%以上,但也有三分之一的人书写很难达到高识别率。识别速度基本能跟上人书写的速度。书写时要求笔划数目和类型基本正确,不要求笔顺,最常用的少数字可以连笔书写,是属于联机手写楷书识别的范围。低限制的联机手写行书汉字识别正在研究。和击键编码人工输入汉字相比,联机识别装置虽然输入速度较慢,但有不需记忆规则,人人会操作的好处。它是"想打"型输入设备,一边构思文章,一边输入,输入不妨碍思考,这是编码输入很难解决的问题。印刷体汉字识别是汉字识别研究的主力军,自 1986 年以来,各种识别软件与系统像雨后春笋般涌现,其中少数识别装置可以初步实用,它们的主机全部采用微机。中国已鉴定的印刷体汉字识别软件和系统示于表 1.3。都采用硬设备扫描输入,对样张识别不少系统都达到高指标,可识别宋、仿宋、黑、楷体,识别字数最多为 6763,字号从 3 号到 6 号,识别率高达 99.9%,识别速度在用 286 微机时达到 9～14 字/秒。最近两三年中,已有五、六个系统脱颖而出,初步达到实

表 1.1

表 1.2

表 1.3　中国已经鉴定的印刷体汉字识别软件和系统

单　　位	字体	字数	字　　号	输入设备	样张识别率	实际识别率	识　别　速　度	单/多体	鉴　定时　间
南通电子技术应用研究所	宋 仿宋 黑	1200 1200 1200	1 号 (9× 9mm^2)	专用 CCD 单字输入	95.9%	/	0.1 字/秒 CJ-709, 2.5MHz 48 位	多	1985.12
哈尔滨工业大学	宋	3755	2 号 (7.4× 7.4mm^2)	传真机 8 线/mm	95%	/	0.5 字/秒, LS-83	单	1986.5
清华大学计算机系	宋	3755	5 号 (3.7× 3.7mm^2)	摄象机 12 字输入	98.3%	/	2 字/秒 mv-4000+ PC- XT	单	1986.6
沈阳自动化所	仿宋	3755	3 号 (5.6× 5.6mm^2)	传真机 8 线/mm	99%	95%	1-2 字/秒 PC- XT	单	1986.10
清华大学无线电系	宋	6763	3 号	传真机 8 线/mm	98%	/	0.3 字/秒, PC- AT	单	1986.11
郑州解放军电子技术学院	宋 黑	3755 3755	4 号 (4.8× 4.8mm^2)	传真机 8 线/mm	98.6%	/	< 3 字/秒 CROMEMCO 系统 3	双	1987.4
河北大学	宋 老宋 扁宋 黑	6763 6763 6079 4074	2 号—4 号	摄象机 单字输入	宋 98% 黑 95%	/	0.18 字/秒 HP-9000	双	1987.7
广州电子技术研究所	宋	3755	小 3 号	图文扫描器 (12 线/mm)	97%	/	4.8 字/秒 PC- AT	单	1987.10
哈尔滨工业大学	宋	6763	2 号	传真机(8 线/mm)	99.5%	/	0.35 字/秒 PC- XT	单	1987.12
西安交通大学	宋	6763	3 号	传真机(8 线/mm)	98%	/	0.9 字/秒 PC- XT	单	1987.12
南开大学	宋 仿宋 黑 楷	3755 3755 3755 3755	3- 4 号(对宋体, 还有大 5 号大 6 号)	图文扫描器 300dpi	98.8% ～ 99.9%	/	9- 14 字/秒, 加强 PC/XT, 80286 8MHz	单	1988.1
北京信息工程学院	宋 仿宋 黑 楷	6763 3755 3755 3755	3- 5 号	图文扫描器 300dpi	99.65% 99.4% 99.4% 99.4%	书刊 95.2% 文件 97.9%	5- 11.5 字/秒 (PC- AT, 6MHz) > 20 字/秒(386 机)	单	1988.5
郑州解放军电子技术学院	宋 仿宋	3775 3775	3 号	图文扫描器 300dpi		文件 97.5%	3 字/秒, 386 机	双	1988.6
清华大学无线电系	宋 仿宋 楷 黑	各 3775	3- 5 号	图文扫描器 300dpi	97% ～ 98.7%	95% ～ 98%	2.6 字/秒 (386 机)	多	1989.4
中科院计算所	仿宋	3755	3- 4 号	图文扫描器 300dpi	/	97.7%	1.5 字/秒 (PC/AT 6MHz)	单	1989.11
河北大学	宋 仿宋 楷 黑	3755		图文扫描器 300dpi			20 字/秒 386 机 20MHz	单	1990.11
邮电部数据通讯技术研究所	宋 仿宋 楷 黑	4500		图文扫描器 300dpi	≥99%	95- 98%	37 字/秒	单	1990.12

用,已在市场销售。它们主要指标为:①识别字数:3755;②识别率:对中等质量印刷文体达到95%～99%;③识别速度:10～40字/秒;④识别字体、字号:宋、仿宋、楷、黑体,3号～6号字;⑤有一定的版面分析和后处理功能。这些系统都配备了方便的用户界面,从版面分析、文本识别到识别结果后处理,形成了一个完整的识别输入系统。中国印刷体汉字识别系统主要产品概况示于表1.4。手写印刷汉字识别的研究,也在认真地研究中,自1989年底以来,已有六个软件与系统研制成功并鉴定,参见表1.5。

当前,为适应中文笔式计算机的兴起,联机手写汉字识别正在新起点兴起新高潮。它将向两个方向发展。一是研究不严格依赖于笔划和笔顺的手写行书文字(或自然书写文字)识别;二是研制价格便宜、性能稳定可靠、特别是书写方便的板和笔。印刷体汉字识别要提高识别系统的品质和效率,增强系统对不同文本的适应性,扩大使用面。要加强版面分析、识别结果上下文匹配后处理和各种实用化技术的研制。手写印刷体汉字识别要着重于能实用的识别方法研究,促使初步实用的产品出现。

表1.4　中国印刷体汉字识别系统主要产品概况

单位	系统名称、型号	字数	字体	字号	识别率	识别速度	实用时间	备　　注
北京信息工程学院	BI-Ⅰ PCCR	3755～4000	宋仿宋楷黑	3号～6号	95%～98%	20字/秒	1989.6	
沈　阳自动化所	SY-OCR	3755	宋仿宋楷黑	3号～6号	95%～98%	20～30字/秒	1990.12	①多体 ②楷体识别率 ＜80% 黑体识别率 ＜90% ③可识别部分繁体字
邮电部数据通信技术研究所	ZXOCR	4500	宋仿宋楷黑		95%～98%	30～40字/秒	1990.12	
河北大学	HQ-1	3755～4000	宋仿宋楷黑		95%～98%	20～30字/秒	1991.1	可识别部分繁体字
清华大学电子系	清华OCR	3755～4000	宋仿宋楷黑	1号～5号	95%～98%	3-6字/秒(纯软件) ＞20字/秒(硬件)	1991.3	①多体 ②有版面分析、切分功能 ③楷体识别率:84-86%
北京信息工程学院	BI-Ⅱ PCCR	3755～4000	宋仿宋楷黑	3号～6号	97%～99%	12～20字/秒	1991.6	①可识别部分繁体字 ②可识别多体多字号混排英文 ③有版面分析、切分功能