中 文 信 息 学 报
第6卷 第1期 JOURNAL OF CHINESE INFORMATION PROCESSING vol.6 No.1

印刷体汉字识别技术在我国的发展和应用

张炘中 沈兰生

（北京信息工程学院）

一、印刷体汉字识别技术的历史和现状

摘要 本文论述了我国印刷体汉字识别技术研制的三个阶段，指出了我国印刷体汉字识别系统的特色，并提出了今后发展的动向。

我国印刷体汉字识别技术的研究自七十年代末起步，至今已有十几年了，回顾这段历史，大致可以分为三个阶段。

1. 探索阶段(1979—1985)

在对数字、英文、符号识别研究的基础上，自70年代末，国内就有少数单位研究人员对汉字识别方法进行了探索，发表了一些论文，研制了少量模拟识别软件或系统。这个阶段漫长，成果不多，但孕育了下一阶段的丰硕果实。

2. 研制阶段(1986—1988)

1986年初到1988年底，这三年是汉字识别技术研究的高潮期，也是印刷体汉字识别技术研究的丰收期。有11个单位进行了14次印刷体字识别的成果鉴定，这些系统对样张识别能达到高指标：可识别宋、仿宋、黑、楷体，识别字数最多可达6763个，字号从3号到5号，识别率高达99.9%，识别速度在用286微机条件下达到10～14字／秒。如果对实际使用的大量文本也能达到以上指标，那末，我国印刷体汉字文体识别的难题就解决了。事实上，用以上识别系统识别实际文本，好的对部分文本还可以达到95%左右的识别率，差的识别率下降到50%以下。这是由于以上系统对印刷文字形状变化的适应性和抗实用文本种种干扰（如文字模糊，笔划粘连、断笔，黑白不均，纸张质量，油墨反透等）性能差。这三年研制的识别系统为印刷体汉字识别系统实用化打下了基础，是识别系统从研制到实用化必经的过程。

3. 初步实用阶段(1989年以后)

从1989年到目前不到三年的时间内，已有5、6个系统脱颖而出，初步达到实用，已在市场销售，迄今约销售300套左右。它们主要指标为：①识别字数：3755。②识别率：对中等质量印刷文本达到95%～99%。③识别速度：10～30字／秒以上。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。

我国已鉴定的印刷体汉字识别系统示于表1。

表1. 我国已鉴定的印刷体汉字识别系统

单　位	字体	字数	字　号	输入要求	拒识识别率	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 (8×8mm²)	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 (7.4×7.4mm²)	传真机 8线/mm	95%	/	0.5字/秒 LS－83	单	1985.5
浙江大学 计算机系	宋	3755	5号 (3.7×3.7mm²)	摄像机 12字输入	98.3%	/	2字/秒 mv－400+PC－XT	单	1986.6
沈　阳 自动化所	仿宋	3755	3号 (5.6×5.6mm²)	传真机 3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－XT	单	1986.11
郑州炮兵军 电子技术学院	宋 黑	3755 3755	4号 (4.8×4.8mm²)	传真机 8线/mm	98.6%	/	<3字/秒 CROMBMCO系统 3	双	1987.4
河北大学	宋 老宋 仿宋 黑	0763 0763 6079 4074	2号－4号	摄像机 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	3756	小3号	图文扫描器 (12线/mm)	98.8%－ 99.9%	/	4.1字/秒 PC－XT	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，还宋大5号，大6号)	图文扫描器 300dpi	98.85－ 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－小5号	图文扫描器 300dpi	99.65% 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
郑州炮兵军 电子技术学院	宋 仿宋	3756 3755	3号	图文扫描器 300dpi		文件97.5%	5字/秒 386机	双	1988.6
清华大学 无线电系	宋 仿宋 楷 黑	各3755	3－5号	图文扫描器 300dpi	97%－ 98.7%	95%－98%	2.6字/秒 386机	多	1989.6
中科院 计算所	仿宋	3755	3－4号	图文扫描器 300dpi	/	97.65%	1.5字/秒 PC－XT(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
沈　阳 自动化所	宋 仿宋 楷 黑	3755	3－小5号	图文扫描器 300dpi		95%	30字/秒 (386机 33MHz)	多	1990.12

二、印刷体汉字识别系统

印刷体汉字识别系统通常由扫描、微型计算机和相应的识别软体构成（见图 1）。

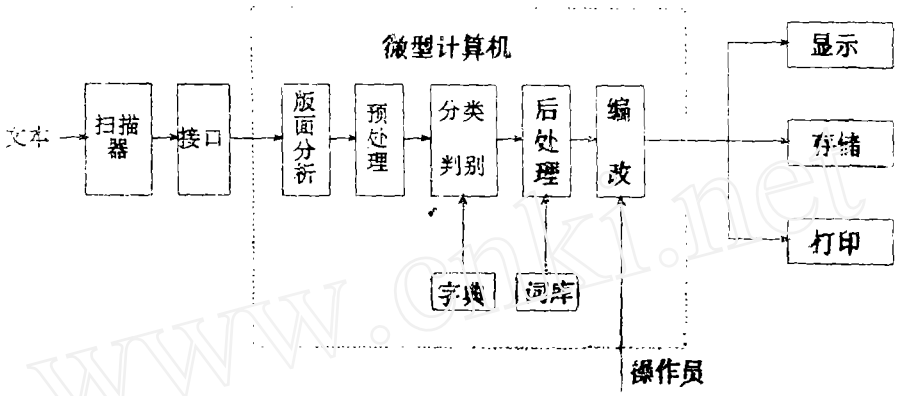

图 1. 印刷体汉字识别系统组成图

从原理上看，一个识别系统可分为扫描图象输入，包括版面分析的预处理，特征抽取、分类判别和后处理四个部分。其中特征抽取、分类判别是核心。

1. 印刷体汉字识别系统的组成

1). 扫描图象输入

通常采用平板型图文扫描器作为汉字文本图象扫描输入设备。它以线型 CCD 器件作水平扫描，用步进电机移动扫描头或摆镜作垂直扫描，文本原稿不动。这种扫描器如惠普 9195 型，其主要指标为：①分辨率：300dpi(12 线／毫米)。②扫描速度：20.4秒。③灰度级：256 级(一般扫描为 16 级)。④扫描幅面：A4。⑤接口：双向 8 位并行。手持式扫描器(如HS-3000PLUS)也可作为扫描输入设备，其分辨率可达 400dpi，缺点是因手持扫描劳动强度大，且扫出图象有畸变和干扰。传真机也可以扫描输入，分辨率只有240dpi。

2). 版面分析和预处理

印刷文本有的版面简单（如书、文件），有的复杂（如刊、报），经扫描输入的文本图象虽先要对版面进行分析，分割出一个个文字才能识别。版面分析的任务就是利用印刷文本的先验知识，自动实现对文本正文域、标题域、图象域、图形域、表格域等切分和标识，并把关连的正文块根据上下关系连接起来，按标题组织文本。

版面分析后连接的正文块是带有随机干扰、噪声的二值数字信号图象，要对它进行行切分、字切分、去噪声、规范化等预处理，再进行单字识别。

3) 特征抽取、分类判别

一种汉字识别的方法就是指特征抽取、分类判别的方法。特征抽取、分类判别是汉字识别的核心，决定了一个印刷体汉字识别系统的品质。

尽管不同的识别系统具体的识别方法各不相同，但就本质而言，可以归纳为统计和结构

两类基本方法。早期国内外的汉字识别方法，往往不是属于统计法就是属于结构法，不善于把两者结合起来。当前汉字识别的结构方法不是纯结构方法，统计方法也不是纯统计方法。在结构方法中应用了统计方法的模式分布性质，而在统计方法中，模式的表示也体现了模式结构特征。

汉字识别基本方法的沿革不是偶然的，是由这两种理论方法的基本不同所决定的。统计法一般采用多维特征向量叠加的办法，把局部噪声和微小畸变“湮没”在最后的叠加和里，但是，可以用来区分结构敏感部位的差别也随之湮没；结构法由于采用结构分析，这些差别不但不被湮没，相反可以得到加强。结构方法对结构特性的敏感也导致了它的不稳定性。统计法常用距离或类似度来判别，即把与标准向量最接近的模式作为识别结果；而结构法则依据一个串形决策过程来判别，其中任何一个误判都可能把整个判别引入歧途。尽管可以采用误差校正分析等手段，但校正过来的可能性不大，且时间开销大。结构方法采用分而治之的方法，把一个复杂的任务分解为若干复杂程度较低的子任务；而统计法由于缺乏相应措施，只好增加特征向量的维数来区分细微差别，对相似字，则用特殊算法进行处理，显得笨拙。

总之，统计法宜识别有噪声的文字，特征抽取容易且稳定，但不能很好地利用结构信息；而结构法可利用汉字字形的结构关系来识别，对文字变体、变形适应性好，但抗干扰能力低。所以，对汉字识别来说，着重汉字字形结构特点，把统计和结构方法结合起来，存优去劣，是当前基本识别方法的主要发展方向。

4） 后处理

识别后处理是实用的识别系统不可缺少的一个重要环节。识别实际文本时，文中大多数字和它相邻的文字受到词、句法、语义的约束因而是相关的，距离该字愈近，相关性就愈强。识别系统可以利用这些相关性来改善孤立字识别时的性能，增加系统识别率。

利用实际汉字文本的相关信息，对识别结果代码文件进一步加工，提高系统识别率称为识别后处理，简称后处理。在这些相关信息中，对微机识别系统来说，用的最多的是汉语词。后处理方法大致有三种：①简单上下文匹配：在拒、误识字前后一定范围内匹配，用词库和后补字信息或文本特征来判别。②词切分上下文匹配：对识别文本的句子自动切分，用词库和后补字信息来纠正误识字。③自然语言理解上下文匹配：用词、语法、语义等知识，逐句对识别文本进行分析、理解，由此选择正确的代替字。

2．我国印刷体汉字识别系统的特色

我国印刷体汉字识别系统在发展和逐步成熟中，形成了自己的两个特色。

1） 特征选择和抽取方法

我国研究人员在汉字特征选择和抽取上，提出了一些不同于国外常用特征的有特色的方法，初步解决了我国印刷体汉字识别的问题。

这些特征选择和抽取的方法都是在对汉字字形结构深入研究的基础上提出来的，它们共同的特点是：认为汉字是有汉字结构特点的特殊图形，考虑到几千或几万个汉字的区别来选择汉字结构中关键的、稳定的、富有信息量的特征。具体来说，在汉字结构特征中，着重选择特征点、特征点组合的横竖结构线段、基于小笔段的汉字层次结构、文字局部稳定结构以及长横、长竖等特征。在抽取方位上，着重文字上、下、左、右四边或文字四角。

2） 识别系统配置

我国印刷体汉字识别系统一开始就在微机环境下进行研制，是一个用普及型3000dpi图文扫描器扫描输入，以386、286微机为主机的通用微机识别系统。和国外相比，有价廉通用，主要用软件识别，性能价格比高，利于推广应用等特点。利用有自己特色的特征来识别汉字，才有可能使这样的系统实现。

三、当前印刷体汉字识别技术的动向和应用前景

虽然有几个印刷体汉字识别的产品在市场销售，并且全国已有300个左右的用户在试用，但相距我国浩瀚的印刷汉字书刊、资料自动输入计算机的实际要求，相差仍远。要使识别系统在我国汉字输入计算机的广阔领域中真正发挥作用，关键在于提高系统品质，即要：①提高系统对实际文本的识别率。②增强系统在各种文本字体变化、印刷质量差别时的适应能力。③提高系统效率。

今后，印刷体汉字识别系统发展动向有以下几个方面：

1．扩大中、高质量印刷文本（如省市大、中印刷厂印刷的书、刊、报纸、中央、部委、省市的文件）识别的使用面。

为此：①加强系统版面分析功能，自动分析和人工干预互结合的交互式功能。②友好的高效率的人机交互界面，简化操作，增加系统的自动功能和“傻瓜”性。③人机交互式后编改灵活、多样、方便，某些专用系统可引入自动校对。④提高系统对汉字、字符混合文本的字符切分正确率，希望达到99.5％以上。⑤增强系统自学习能力。⑥降低成本、减少售价。

2．有良好的汉字、数字表格，整版报纸，传真数据等的分析，处理、识别能力。

目前的系统几乎没有这些能力或者很弱。

3．低质量印刷文本（如县城小印刷厂印刷的书刊，参改消息，油印材料，某些计算机打印体或复印件等）的识别。

这是印刷体汉字识别中的难点，也是一个印刷体汉字识别系统向高档实用化发展的重要标志。希望对低质量的印刷文本能达到95～98％的平均识别率。

4．历史古籍、文献、资料的识别

是印刷体汉字识别的最难的领域。它们用到单字数量多，字体字形（繁体）复杂，纸张、印刷质量差，时代的变迁使字形、笔划模糊，干扰严重。这些对识别来说都是一个个的“拦路虎”，要改进识别方法和后处理方法，增强系统抗干扰性能，逐步解决这个问题。

我们相信，随着识别系统性能和品质不断提高，在办公自动化，建语料库，情报检索中印刷书刊、报、文件及各种资料的输入，书刊再版，机器翻译中印刷资料的输入，书刊报自动阅读，汉字图象高倍压缩通讯存贮等领域中，自动识别将逐渐代替人工键入，高速自动使文本进入计算机，使我国在世界上成为第一个广泛使用汉字OCR的国家。

参 考 文 献

[1] 张炘中，我国汉字识别技术的现状与展望，中国中文信息学会成立十周年学术报告会论文集，1991.6。

Development of printed Chinese character Recognition Technigue in China

Zhang Xin Zhong, Shen Lan Sheng
Beijing Information Technology Institute

Abstract

This Paper relates to the technique of the printed Chinese character recognition, and more particularly to the 3 researching stages and the system features of the technique, at the same time, it points out the development direction of the printed Chinese character recognition.