中 文 信 息 学 报
第1卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.1 No.3

我国汉字识别研究的进展

张 炘 中

(北京信息工程学院)

一、前 言

作为模式识别和人工智能一个分支的汉字识别技术，自七十年代末，我国一些大学、研究所开展研究以来，至今已有十年历史。十年来，从无到有，从几个单位少数人探讨进展到有一定规模的研制队伍在认真探索，从纯原理、方法研究发展到理论、方法、模拟实验、识别系统齐头并进。特别是最近两、三年，汉字识别的研究在印刷体汉字、联机手写汉字、手写印刷体汉字等领域全面开展起来，取得了一些可喜的初步成果，已有一些指标赶上甚至超过世界水平。可以说，我国汉字识别的研究就要摆脱摸索阶段，开始向实用系统研制的道路前进了。

本文着重论述了逐渐形成我国特色的汉字特征选择方法，介绍了在印刷汉字识别、联机手写汉字识别、手写印刷体汉字识别方面我国研究的近况，并且对汉字识别的进一步研究提出了建议。

二、逐渐形成我国特色的汉字特征选择方法

特征及其抽取是模式识别的核心，汉字识别也是如此。抽取什么特征，用什么方法抽取特征，多数情况下就确定了识别方法，也决定了系统可能达到的性能指标。对汉字字形分析研究，得出表达汉字结构本质的特征及抽取方法，是汉字识别中的一个十分重要的问题。

以前发表的粗分、细分所用的汉字特征及抽取方法，大部分是日本人研究出来的，从最早采用的时域到频域变换（Fourier 变换）、轮廓投影发展到转动惯量、复杂指数、四边码一直到当前常用的粗外围、粗网格、笔划密度、笔划方向、网格单元、从背景分析文字特征等。这些特征对两、三千日本汉字可能是有效的，但对七千多中国汉字，有效性如何？很值得商榷。要在微机上实现众多中国汉字识别，原封不动地使用以上日本人提出来的各种特征恐怕是难以实现的，必需使用新的特征。

新的特征怎样选择呢？分析一下日本统计法识别汉字所采用特征的演变可以得到启发。日本最早采用的特征，如轮廓投影Fourier 变换，是把汉字看成一般的两维图形来抽取特征，

本文1987年10月15日收到

丝毫没有考虑到汉字结构本身的特点。发展到四边码、粗外围、笔划密度、网格单元就已经注意到汉字结构特点。我们认为，汉字特征的选取原则上可以从三条思路来考虑：①把汉字看为一般的两维图形，用一般图形选取特征的方法来选取。②汉字是有汉字结构特点的特殊图形，只考虑几千或几万个汉字的区别来选择结构特征。③在汉字结构信息中再选取关键的稳定的部分作为特征，鉴于汉字结构信息的冗余度，这个想法是有根据的。几年来的研究表明，国内愈来愈多的人认识到，从第二条思路考虑选取特征比第一条有效，而从第三条思路来考虑更有效。目前，我国研究者提出的一些粗、细分用的特征，已经比日本人更多地更准确地考虑到汉字结构特点，而且力求从这些结构信息中找出稳定的关键的特征来，在特征提取方面，有不少人使用了人工智能的方法，使特征的提取和运用更有针对性。这就是说，我国汉字识别的特征选择和提取已经开始摆脱日本的束缚，逐渐形成有自己特色的能解决七千汉字分类、识别的特征选择方法。不对文字进行细化处理，直接从二值化点阵文字中提取特征，也是国内外近年来的一种趋势。

国内提出的这类特征是多种多样的，例如：①文字特征点：汉字笔划上端、折、歧、交点和汉字背景上的关键背景点称为汉字特征点。和以往不同的是把笔划特征点和关键背景点两者结合起来，并且直接根据特征点本身的信息（类型、数目、位置等）自上而下或自下而上来识别。汉字特征点反映了汉字结构的本质特征，集中了主要的结构信息。文字特征点是粗、细分类都可以采用的特征，如用文字四周的端点数或文字四个象限的端点数可以进行良好的粗分类。②用“脱壳透视”法计算文字四周的笔划线长度作为分类特征，再将每个汉字的复杂结构抽象成一个具有典型特征的稳定框架模型来抽取笔划向量、特征点、四边外轮廓笔划长度等特征进行细分识别。③用数学形态学的方法提取汉字结构特征，第一级采用汉字边框特征，第二级是和字根相似的局部结构（由横、竖结构线段组成）特征，最后用端点、结点特征来细分识别。④汉字四角或三角上笔划、线段特征。⑤文字的长横、长竖在字中不同区域分布的特征等。它们的共同点是：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、文字局部稳定结构以及长横长竖等特征；在抽取的方位上，着重于文字上、下、左、右四边或者文字四角。

把统计法和结构法结合起来，选取包含丰富结构信息的特征，并且在抽取特征时，更多地应用人工智能方法，进行知识化程序设计，是汉字识别的一条捷径。当然，提取笔划或线段作为基元的纯结构分析法，把模式的结构信息和统计信息结合起来描述的属性文法在汉字识别（特别对手写印刷体汉字识别）中也有重要价值。

日本的汉字识别系统，一般是由小型机或更大的机器构成的，制作专门的硬件，甚至做成专用机，系统价格昂贵，通用性差。例如日本东芝综研2000单体印刷汉字识别装置，类似度计算全部采用硬件，相当一台专用大型机；代表当代世界最高水平的日本武藏野电气通信研究所多字体印刷汉字识别装置是一台带专用硬件的小型机。这样做虽然是由于系统指标高（例如识别速度≥100字/秒），但也和所用特征包含结构信息少，使得特征维数高、存贮量大、算法复杂有关。

在我国目前条件下，广泛使用类似日本价格昂贵的专用汉字识别装置是不现实的，只能推广用微机实现的低价系统，识别率能达到95%（甚至实际识别率能在90%以上），识别速度能在2-10字/秒就可以使用。利用有自己特色的关键的稳定的结构特征来识别汉字，才有可能使这样的系统实现。

总之，分析汉字结构特点和内在规律，选择富含结构信息的关键的稳定的文字特征来识别汉字，构成一个价格低廉的采用普及型图文扫描仪作为输入设备的微机汉字识别装置，是不同于日本的有我国特色的一条研制汉字识别装置的道路。

三、汉字识别各个领域的进展

近年来，随着研究队伍的不断扩大，特别是有一批年青科研人员充实进来，各个领域的汉字识别的研究都生气勃勃的开展起来了，取得了一些初步成果。

在当前汉字识别研究的主流——印刷汉字识别方面，国内已经鉴定了九、十个识别软件和模拟系统（见表1），都用硬设备扫描输入，识别的字体有宋、仿宋、黑体，字数能到

表1 我国已经鉴定的印刷汉字识别软件和模拟系统

单位	字体	字数	文字大小	输入设备及方式	识别率	*识别速度	鉴定时间
南通市电子技术应用研究所	宋 仿宋 黑	各 1200	近似一号字 ($9\times9mm^2$)	专用CCD扫描，每次输入一个字。	95.9%	10秒/字	1985年12月
哈尔滨工业大学	宋	3755	二号字 ($7.4\times7.4mm^2$)	传真机 (4线/mm) 页式	≥95%	1.36秒/字	1986年5月
清华大学计算机系	宋	3755	五号字 ($3.7\times3.7mm^2$)	摄象机 $2\times6=12$字 成块输入	98.3%	预处理：4字/秒 识别：4.8字/秒	1986年6月
中科院沈阳自动化所	仿宋	3755	三号字 ($5.6\times5.6mm^2$)	传真机 (8线/mm) A4页面，每页658字	99%	1—2字/秒	1986年10月
清华大学无线电系	宋	6763	三号字 ($5.6\times5.6mm^2$)	传真机 (8线/mm) 页式	≥98%	3—4秒/字 (不计切分)	1986年11月
郑州解放军电子技术学院	宋 黑	各 3755	四号字 ($4.8\times4.8mm^2$)	传真机 (8线/mm) 每次输入四行94字	98.57%	3.24字/秒 (不计预处理)	1987年4月
河北大学	宋 老宋 扁宋 黑	6763 6763 6079 4274	二号～四号字 ($7.4\times7.4mm^2$ ～ $4.8\times4.8mm^2$)	摄象机 每次输入一个字	宋：98% 黑：95%	5.71秒/字	1987年7月
广州电子技术研究所	宋	3755	小三号	图文扫描仪 (EITPS) A4页面输入 每页600多字	>95%	≥4字/秒	1987年10月
哈尔滨工业大学	宋	6763	二号字	传真机 (8线/mm) 每次输入94字	>99.5%	2.6秒/字	1987年12月

* 由于各系统所用机器及软件的类型不同，识别速度仅供参考。

6763个，文字大小可以到五号字，识别率达到98～99%。改善输入设备，增强系统对不同印刷样张的适应能力，完善系统硬、软件配置，在一定范围内实际试用，是印刷体汉字识别的当务之急。

在联机手写汉字识别方面，已经研制了几个初步实用的装置，并已在试用。其指标大致为：识别字数：6763个；识别率：一般在90%以上，字写的愈规范，识别率愈高；识别速度：基本能跟上人书写的要求。书写时要求文字笔划数目和类型正确，笔顺可不要求，属于联机手写楷书汉字识别范畴。低限制的联机行书手写汉字识别正在研究。联机识别装置要直接和键盘人工输入汉字相竞争，要在价格、使用方便性、稳定可靠性和输入速度等方面有优越指标才能占领市场。为此，今后仍要在研制价格便宜、性能稳定可靠、书写方便的输入板和笔；配置众多的方便用户的实用软件；研究低书写限制的识别方法等方面努力开拓，为最终实现能识别像人作笔记时所用的字体和速度来书写的文字而努力。

在手写印刷体汉字识别方面，十年来，一直在坚持研究，由于研究人员较少，课题又不集中，到目前尚无明显的突破。虽然手写单字识别比印刷体要困难，但是：①手写体识别在输入页张格式上比印刷页张简单且规格化，基本上不存在版面处理问题。②对扫描输入分辨率的要求较低。③限制严格程度可以控制识别难度。所以，只要有合适的识别方法，对写在方格中的限制较严格的手写规整汉字，可以较快地研制出初步实用装置来。

总之，经过十年的努力，我国汉字识别研究就要从探索阶段走出来，开始向研制实用的汉字OCR进军了。

四、几点建议

1. 抓紧当前良好时机，大力深入开展汉字识别的研究。尽快研制出一批不同风格的能初步实际使用的印刷体汉字识别和联机手写汉字识别装置，在实用中发现问题，改进方法，不断完善。为此，除研究识别方法本身外，还要加强和实用相关连的问题的研究。例如，在阅读实际书刊、文章、文件时存在的干扰（不同字模、纸张、油墨、印刷质量、污点、断笔、输入设备的鉴别率和精度）条件下，如何保持一定的识别率？从正文、标题、图象、图形、表格、公式等混排的页面上如何切分出汉字？如何做成一个开放性系统，让用户自己可以制作、扩大、修改识别字典，进行自学习、提高系统识别率？如何使识别后的文件便于编辑修改，方便用户使用等。

2. 加强汉字识别基本方法的研究。

1）汉字特征选择和抽取的研究。虽然我国不少学者已经提出了众多能反映汉字本质结构信息的特征，形成了自己的特色，但是，离开解决中国汉字识别的要求还差距很大。今后仍需加强特征抽取尤其是手写汉字特征抽取的研究和计算机自动抽取特征的准确度和速度的研究。

2）分类方法的研究。在传统的模式识别分类理论的基础上，开展新的适合汉字模式识别的新分类方法的研究，对提高识别速度和识别率能起到重要作用。例如国内提出的全局训练式大型树分类器对提高单体印刷汉字识别的速度和识别率，起了决定性的作用。

3）机器学习的研究。汉字量多，每个字的写法，即使对于印刷体也有十多种，加上还要反映文字中实际存在的各种干扰，一个实用的汉字识别系统的标准子样就有几十万。用这

样庞大的子样进行实验，从中总结规律，提高识别率，就要求机器能够学习。机器学习就是计算机根据事先提供的一定数量的子样，经过人的教授训练，归纳出识别字典，使系统不断增加和改善识别能力。

机器学习是人工智能中一个难度较大的理论问题。用语言结构方法来描述，所谓学习就是文法推断过程。在人提供的正、反训练子样集的条件下，由机器归纳出文法规则。

如果提供的子样集不完整，就要求计算机有更高的学习能力——概念学习能力。要能像人那样从有限的正、反子样中推断出文法规则。要求机器归纳出正子样的共同特点，而排除所有反子样，从而形成一个共同的概念。当新的正子样出现时，学习向更一般的概念移动，以复盖所有正子样；当新的反子样出现时，要缩小概念范围，以便使反子样不包含在该概念中。

3. 在识别软件成熟的条件下，进行识别装置硬件的研究。研制输入设备，改革硬件体系结构，提高识别速度。

目前汉字识别输入装置大都采用CCD元件制成的普及型平板式图文扫描器，分辨率为300线/吋，扫描A4页面约为30秒，灰度级为16。基本上满足了识别五号或五号以上汉字的需要。对书刊常用的五号宋体字，分辨率稍嫌不足。配置或研制分辨率和精度都能达到14线/mm的图文扫描仪，能够提高系统的识别率，扩展系统的应用范围，将会有力地推动汉字识别装置的实用化。

用一般微机识别汉字，识别程序全部采用软件，识别速度很难超出15字/秒。所以，在识别软件经过实践的检验后，应当研制相应的并行处理、阵列处理硬件卡插入微机，使识别速度达到20～50字/秒，缩短汉字信息处理系统输入速度低、处理和输出速度高的差距。

古老的汉字是我们民族的瑰宝，是亿万人表达感情、交流思想的工具。历史遗留下来的浩如烟海书刊、资料、典籍是我们巨大的财富，在当今广泛使用计算机的信息社会里，不解决汉字自动、高速输入计算机，历史资料的综合利用，办公自动化，中文资料库的建立，情报的检索传递，文字的自动翻译以及新一代计算机智能输入的研制都是不可想像的。

国外的字符识别机，早在五十年代就投入实用，到七十年代，技术已经成熟，输入速度达到每秒数千字符，识别率达到99.99%，其速度和质量已远远超出人力所为，我们相信，随着我国实际应用的需求和研制队伍的不断扩大，目前尚不成熟的汉字识别技术，在不久的将来，会全面开花结果，各种类型的汉字识别装置将要在我国四化建设中发挥出作用。

参 考 文 献

[1] 第二届全国汉字及汉语语音识别学术会议论文集，1987年8月，大连。

[2] K. Mori and I. Masuda, Advances in Recognition of Chinese characters, Proc. of 5th Intern. conf. on Pattern Recognition (1980) PP692—720.

[3] 桑源启治，大分类の段階ガぼ完成した手书き漢字认识の研究，日经エレクトロニクス，NO. 279 (1981)，PP148～PP167。

[4] 森健一等，2000字種を100字/秒で読む印刷汉字OCRの开発，日经エレクトロニクス，NO. 172 (1977)，PP102～P128。

[5] 饭岛，混合类似度による识别理论。电子通信学会研究会资料，PRL74—20，(1974)。

[6] 桑源启治，印刷、手书き汉字の认识技术を展望すろ，日经エレクトロニクス，NO.154 (1977) PP42—59。
[7] 坂井、森，汉字パターンの大分类，电子通信学会研究会资料，PRL73—14 (1973)。
[8] 河田、平井、森，汉字认识のための汉字大分类手法，昭和48年度情报处理学会全国大会，114 (1973) PP227—228。
[9] 张炘中，汉字自动识别研究综论，中文信息，创刊号 (1984)，PP11—14。
[10] 张炘中，对我国汉字识别研究的建议，中文信息，NO.2 (1986)，PP36—39。
[11] 小高、荒川，增田功，ストロークの点近似による手书き文字のオニライン认识，电子通信学会论文志，Vol. J63—D，NO2 (1980) PP153—160。
[12] M. Yoshida and M. Eden, Handwritten Chinese character recognition by an A-b-S mathod, proc. of 1st Intern. conf. on Pattern Recognition, (1973) PP197—204。
[13] 叶培建，计算机实时手写中文自动识别，自动化学报，No1 (1987)。
[14] 童为群，低限制手书体汉字联机识别，中文信息，4 (1986)，PP5—12。
[15] 小高和已，若原徹，橋木新一郎，オンライン手書き文字认识装置，信学论 (D)，J65—D，NO8，PP51—58 (1982)。
[16] 梅田三千雄，マルチオント印刷汉字の分类，信学论 (D)，Je2—D，NO2 (1979) PP133。
[17] 目黑员一，梅田三千雄，マルチフオント印刷汉字の认识，电子通信学会论文志，Vol. J65—D，No.8(1982)，PP1062—1038。
[18] 北岛宗雄，漏れ书体を含むマルチフオント印刷文字认识，信学论 (D)，J66—D，NO4 (1983)，PP400—406。
[19] 北岛宗雄，射影方向线素による多種字体印刷汉字认识，信学论 (D)，J67—D，NO3 (1984)，PP249—256。
[20] 郭宝兰，汉字识别的“包含配选法”及其应用，通信学报，Vol. 4 (1983) P52—58。
[21] 武裕朴，赵景台，杨力，2500个印刷体汉字识别研究，自动化学报，VO1. 10，NO1 (1984)，PP58—60。
[22] Q.R. Wang, C.Y. Sun, Analysis and Design of a Decision Tree Based on Entropy Reduction and its Application to Large Character Set Recognition, IEEE Trans. Vol. PAMI—6 No.4(1984), PP407—417。
[23] Y.X.GU, Q.R. Wang, C.Y. Sun, Application of A Multilayer Decision Tree in Computer Recognition of Chinese Character, IEEE Trans. Vol. PAMI—5, (1983), PP83—89。
[24] Michio Umedu, Recognition of Multi-Font Printed Chinese Chararter, Proc. 6th IJCPR., (1982), PP793—796。
[25] 朱夏宁，吴佑寿，丁晓青，二级印刷体汉字的识别，清华大学学报，27，No1 (1987)，PP39—49。
[26] 小林等，ストローりマツチングによろ手書汉字认识，电子通信学会技术研究报告，PRL 81—38 (1981)
[27] 山本，Relaxation法によろ手书教育汉字认识，同上，PRL81—31 (1981)，
[28] 荻田、增田，文字線の方向性に着目した手书き汉字の识别，同上，PRL81—13 (1981)
[29] 内藤、小森、淀川，手书汉字认识のためのストローり密度特征，电子通信学会论文志，Vol.81，NO.7 (1981) PP593—600。
[30] 藤井等，多元的な特徴によろ手书汉字认识の检讨，电子通信学会技术研究报告，PRL81—32，(1981)。
[31] 张炘中、夏莹、孙承鉴，用抽取笔划法识别限制性手写汉字的探讨，计算机学报，5:6 (1982) PP455—462。
[32] 冈隆一，セ儿特徴を用いた手書き漢字の认识，信学论 (D) J66—D，NO1 (1983) PP17—24
[33] 山本和彦，弛缓整合法による手書き教育漢字认识，信学论 (D) J65—D，No9 (1982)，PP1167—1174
[34] 张炘中、夏莹、The automatic recognition of handprinted Chinese characters-A method of extracting an ordered sequence of strokes, Pattem Recognition Lettersl (1983) PP259—265。
[35] T. Takahashi, I. Masuda, Handprinted Chinese character recognition using stroke extraction method with referral to peripheral structural information, J. IECE, J67—D, 9. (1984) PP1052—1059。
[36] H. Hase, M. Yoneda, M. Sakai, J. Yoshida, A method of the elastic extraction of subpatterns in handwritten Chinese characters, J. IECE, J67—D, 8. (1984) PP829—836。

[37] S. Mori et al., Line filtering and its application to stroke segmentation of handprinted Chinese characters, proc. 7th ICPR, (1984)。

[38] 夏莹，张炘中，自动识别手写印刷汉字系统中的部件分离问题，计算机学报，Vol.8，No.6 (1985)。

[49] 夏莹，张炘中，用于机器识别和学习的汉字表达式，自动化学报，Vol.12，No.3 (1986)。

[40] 张炘中，夏莹，限制性手写汉字中笔划的抽取、分析和合成，计算机学报，Vol. 10，No.3 (1987)，PP166—174。

[41] 路浩如，基于笔划元分析描述汉字模式的属性文法，Proc. ICCIP (1987)，北京。

简　　讯

中国计算机学会中文信息技术专业委员会于1987年10月成立。该专业委员会的研究方向是中国语言文字处理的计算机系统和以计算机为工具研究语言文字处理技术。它将同其他学会开展广泛的横向合作。

第一届专业委员会目前由31人组成。主任委员由蒋维镛担任。副主任委员为毛德行、孙强南、姚天顺、陶沙、龚滨良、鲁元魁、傅永和、韩承德、雷天模、嘎日迪。秘书长为王之燿。挂靠在电子工业部第十五研究所。

根据本学科的发展及专业委员会的特色，目前设立技术标准、系统及应用、输入输出技术、计算语言学及机器翻译、少数民族语言文字处理等五个学组。