手写汉字的机器识别

周昌乐 著

科学出版社

序　言

手写汉字识别的研究是人机接口自然化和智能化进程中的一个重要课题，尤其是对于我国计算机高新技术的发展更是如此。随着手写汉字识别研究的深入发展，实验系统渐趋成熟，手写汉字实用识别系统的研究就成为一个迫切需要解决的问题。在这种情况下，系统总结该领域中已有的研究成果、思想方法、理论技术和前景展望，出版一部内容充实，思想新颖，既能反映这一研究领域的前沿动态，又可为一般工程技术人员了解这一研究领域的概貌提供一部有益的参考书，无疑对于促进实用化手写汉字识别研究进程，普及手写汉字识别新技术，都有着重要的学术价值。鉴于目前介绍汉字识别技术方面的书籍极少，因此我们撰写了这部专著。

这部专著是作者在博士论文的基础上，经5年潜心研究的结果。原本想基于协动计算思想，构造一种新的视觉计算理论方法，以用于手写汉字的机器识别，但经过8年来的深入研究发现，手写汉字的机器识别问题远非开始所想象的那样简单，许多表面上看起来轻而易举的事情，实际做起来往往会遇到种种意想不到的困难。所以，作者8年来基本上只是围绕研究中遇到的各种具体问题，来展开协动计算原理的研究。本书的内容除了广泛性介绍外，也可以看作是作者所做研究工作的一个阶段性总结，充分利用作者8年来的研究成果，在作者等提出的协动计算思想的框架下组织内容，给出了一种识别手写汉字的协动计算原理及其手写汉字识别各阶段实现方法的系统描述。

从大处着眼，从小处着手，是作者所进行手写汉字识别研究的主导思想。一方面，要对人类认字活动的现象和规律作全面的考察，找出认字活动最基本的原理，我们称之为协动原理；另一方面，要对人类认字活动过程，通过数学和计算的方法，进行拆解重整，

完成整个手写汉字识别系统的构建，而这其中又自始自终围绕着最基本的协动原理。

协动一词，英文为 Coaction，是协作行动的简称。把它作为我们研究思想的核心概念，其内涵已远远超过字面上的解释，整体性、同时性、关联性、过程性等等都可以从中窥见端倪，所以也就最难理解。好在我们不必去探究其中哲学上的意义，把协动简单解释为计算系统各部分相互作用和共同行动所表现出来的行为活动规律也无不可，或许这样更容易为人们所接受。

最后，研究工作之所以取得初步结果，首先要感谢许多相识的与不相识的专家、学者所给予的启示和帮助，感谢我的导师马希文教授对我早期工作的悉心指导和教诲，感谢中科院院士石青云教授所领导的北京大学视觉与听觉信息处理国家实验室 5 年来对我研究课题的赞助和支持，以及感谢浙江省科委、自然科学基金委、杭州大学校系领导对我研究工作所提供的研究经费和便利条件。此外，对作者历届研究生骆建华、苗兰芳、胡剑峰、金雄伟、张雄伟、钟晓、古红英、吴劲松等所具体从事汉字识别系统的实验工作表示感谢，本书的最终形成与他们的辛勤劳动是分不开的。

目　录

第一章　研究综述

语言是信息交流的主要工具，因而社会越发达，处理的信息量越大，我国每天待识认的手写汉字就越多，无疑手写汉字自动识别的需求也就越迫切。这就要求人们对手写汉字的机器识别问题进行广泛而深入的研究，找出实际有效的解决方法。为此，首先必须对手写的汉字识别研究的现状以及问题的各个方面，作全面分析，以便一开始就能明确这项研究工作的已有成就和困难，以及我们应该持有的观点。

1.1　汉字识别的研究现状

在我们的社会里，每天都有数以亿计的手写汉字要识认和处理，并且随着现代信息交流的不断强化和计算机应用的不断普及，这种需求逐年猛增。邮电通信、新闻出版、办公事务，甚至日常工作，都需要将手写的汉字，转变为机器内部可以保存的形式，以便能方便地进行变换、传输和输出，加快语言信息的交流。目前，这项工作只能依靠人工输入的方法来完成，采用五笔字型或拼音等各种编码方式，逐字逐码击键输入。很明显，这不但同自动化的机器方式相脱节，而且劳动强度大，工作效率低，远不能满足高速度、高质量自动化汉语信息处理的需要。人们自然会想到，如果能让机器来自动识别手写的汉字，那该多么理想！这便是手写汉字机器识别研究的缘由。

1.1.1　研究领域界说

从严格意义上讲，手写汉字的机器识别研究，就是要使用科学和技术的人工手段，通过计算装置，在某种程度和范围上来实现手

写汉字文稿的自动识读，其核心问题是要识别出一个个书写正确的汉字，给出其对应的字类归属。过去近30年的研究业已表明，这并非是一件轻而易举的事情，许多问题有待于人们去探索、研究和解决。

手写汉字的机器识别，在学科上可以看作是模式识别领域中文字识别的研究范围，其中汉字识别是文字识别中最困难的部分，而手写汉字识别又是汉字识别中最困难的部分。因此可以说，手写汉字的机器识别研究是文字识别研究的最高难度问题。从应用角度看，手写汉字的机器识别，也是人机接口自然化和智能化的一个重要研究分支，特别是在我国，要实现人与机器的直接交流，那怕是部分实现，都离不开汉语文字的读写。这其中手写汉字的识别便是一个不可回避的基本问题。

也许有人会认为，与机器直接交流，我们也可以采用口语的形式，让机器具备听说能力。但由于汉语同音字现象十分严重(平均同音字在17个字左右[1])，加上语音学特性本身的不确定性，即事实上说同个音节时，不同的人(甚至同一人在不同的时候)往往产生特征不同的声波；而在说不同的音节时，又完全有可能产生相似的声波[2,3]。因此，依靠语音识别来实现人机接口自然化和智能化，存在着更大的难度。

比如，赵元任先生创制的一则《施氏食狮史》[4]，就是只能看不能讲说的，其文如下：

“石室诗士施氏嗜狮，誓食十狮，氏时时适市视狮。十时，氏适市，适十狮适市。是时，氏视是十狮。恃十石矢势，使是十狮逝世，氏拾是十狮尸适石室。石室湿，使侍试拭石室。石室拭。氏始试食是十狮尸。食时，始识是十狮尸实石十狮尸。是时，氏始识是实事实。试释是事。”

可见，人机接口自然化和智能化更多地只能依靠于书写形式之上，于是手写汉字的机器识别也就成为必由之路。

另外，归纳起来看，手写汉字机器识别的实现，不但可以普及计算机的使用，提高使用效率，而且能够解决人机通信的矛盾，为

机器理解和翻译自然语言、实现邮电通信自动化、完成新闻出版程式化等提供了理想的输入手段，并有利于信息的压缩、传输和交流。特别是将图像汉字转变为符号汉字，大大缩小了(100 倍左右)数据量和提高了传输速度，比起采用传真技术实现的手稿文本传输来说要优越地多。

所以，手写汉字的机器识别研究，有着广泛的应用前景，比如像邮政报文和信封地址的识读，手写文稿的自动输入和编辑，书刊的出版印刷等等，其自动实现，都离不开手写汉字本身的自动识别。

正因为这样，于是围绕着手写汉字的机器识别这一目标，在印刷体汉字识别技术的基础上，利用图像处理和分析、模式识别、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、拓扑几何学以及语言学、汉字学、神经科学、认知科学和系统科学等相关学科的理论和技术，近 30 年来，主要是在日本和中国逐渐展开了手写汉字识别的研究工作。

1.1.2 历史简要回顾

最早进行手写汉字机器识别研究工作的是日本 NNT(国家电话电报公共公司)的研究人员[5]，接着 Fujitsu 的研究人员构建了第一个用于手写汉字识别的计算机模型[6]，并在日本的第 56 届商展上(1981 年 5 月)公开展出。这不但打破了机器不能识别手写汉字的心理学非难，而且也在日本掀起了更多有关手写汉字识别的研究[7]工作，包括手写汉字样本字库的建立，各种特征选择和匹配方法的研究和实验。这些研究，代表了早期手写汉字的研究水平。各种基本的识别方法都有了雏型，有的甚至已成体系，无疑为后来的研究奠定了基础。就识别率而言，在小范围内(ETL 样本字库等)，最高可达 98.8%[8]，最低也有 74.8%之多[9]。

中国(包括台湾在内)的起步较晚，大约在 80 年代中期，才陆续有这方面的研究报道[10~12]。与日本的研究人员不同，由于面对的汉字数量更大，所以更偏重于结构方法的研究。主要有基于属性

文法方面的研究，基于笔划、部件抽取方面的研究，也有一些特殊方法的探讨研究，以及有关手写汉字识别中前后处理方面的研究。

80 年代中后期到 90 年代初，是手写汉字识别研究最为活跃的时期，各种思想、方法和系统不断涌现；无论是在中国大陆、中国台湾，还是在日本，都呈现了前所未有的好形势，取得了众多的研究成果。特别是在结构匹配、松弛计算以及人工智能的应用方面，有着长足的发展[13~20]。

最近几年，由于神经网络的影响，在手写汉字识别研究中也出现了位数不少的神经网络匹配方法[21~35]。将神经网络方法与其它获得最好结果方法进行比较，可以清楚地看到，神经网络识别方法依然处于起步阶段。或许，要取得令人信服的结果还需要做更多的努力。

目前，随着手写汉字识别研究的深入发展，实验系统渐趋成熟，手写汉字实用化研究成为一个普遍关注的问题，无论是识别范围还是识别性能，都趋向于实用化的研究。比如在最近发表的研究论文中，就有针对大容量汉字字库的识别研究[36]，利用语境信息的文本确认研究[37]，以及性能综合分析的研究[38,39]等等。

1.1.3 一般原理概述

纵观已有的研究，通常手写汉字识别的一般过程可以用如图 1.1.1 所示的原理图来表示。主要包括预处理、单字识别、后处理三个阶段以及各阶段所需的样本字库、识别字库和关联语库三个数据基。视情况不同，单字识别常常又有特征提取和粗细分类二个不同步骤的区分。考虑到识别字库的形成方式，有时相应地增设有关标准模板自学习的功能。

首先，对于输入的手写汉字图像文本，预处理[40~43]包括所有需要将输入汉字图像转换为对于系统特征提取部分可接受形式的步骤。其内容和要求依赖于后续处理中的识别方法，一般有文本切分、数值化(二值化)、平滑、细化以及规整化等几个步骤，往往采用现有的图像处理技术来完成。

经预处理后，得到的数据量是很大的，而每个像素数据所含的有用信息却很少。特征提取就是要将图像中的有用信息集中于一些少量的、经精心选择的特征上。自然特征选择的好坏直接影响到系统性能的好坏，所以选择稳定而有代表性的特征，就成为手写汉字识别研究中的核心问题之一。

通常我们可以将已有研究所采用的特征种类，泛泛地分为统计和结构二大类。其中统计的又可分为局部的、整体的二类。

局部特征的明显特点是只注意汉字的局部区域的统计性质而相对忽略汉字的结构规律和整体统计性质。比较常用的局部特征包括像素本身[44~46]、线元梯度[47]、笔划域[48~51]、方向段强度[52,53]、方向笔划数[54~56]，方向映象[57]、方向密度[58~60]和周边形状描述[61~67]。当然，也有采用多种局部特征综合研究的[68~81]。

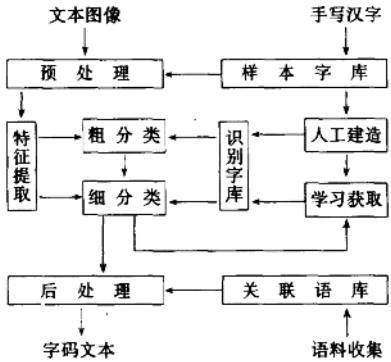

图1.1.1 手写汉字识别的一般原理图

不同于局部特征，整体特征尽管也忽略汉字的结构特性，但在计算统计特征时是把汉字图像作为一个整体考虑的。已用于手写汉字识别的整体特征，包括有一维投影特征和二维变换特征等[82~103]。总的来说，整体特征似乎比局部特征更多地忽略了汉字的内在结构信息，所以往往都仅作为粗分类特征而被运用。

结构特征是那些能反映汉字内在结构性质的特征。这些特征

往往就是构成汉字的基本要素。因此，鉴于对汉字基本要素的不同看法，就形成了多种结构特征类型，主要包括有要点（特征点）[104~107]、笔划[108~125]、笔划序列[126~132]和部件[132,133]这样四种。

特征选择之后就是分类处理，分类通常使用某种策略作为依据，通过将输入汉字的特征取值，与识别字库中每个字类代表的标准取值进行比较来完成，即所谓的匹配。由于在手写汉字识别中，字类总数很大，往往采用多级聚合的策略来进行逐级分类。也即，首先将输入汉字匹配到某个字类范围，然后再对其进行进一步区分，最终获得匹配的单个字类。这便是粗分类[134~140]和细分类[141~143]不同概念的产生原因。

当然，不管是粗分类，还是细分类，也不管区分粗细分类，还是不区分粗细分类，要进行分类匹配，就离不开分类匹配方法的确定。在手写汉字识别研究中，比较常用的匹配方法可以归纳为对比匹配方法[144~148]、结构匹配方法[149~163]和松弛匹配方法[164~174]。

对比匹配方法的主要思想是基于特征空间中某种距离测度的计算。几乎所有统计的局部和整体特征向量都可以采用对比匹配来完成对汉字的分类。如果解决符号串之间的距离计算[175]，也可以将此方法用于结构特征的匹配[176]。

结构匹配方法则主要是针对结构特征的一种匹配技术，大致可以细分为文法匹配、关系图匹配和树搜索匹配等。由于结构匹配方法比较灵活，所以也常常结合模糊数学和人工智能技术来完成字类匹配[177~183]。

既可用于统计特征，又可用于结构特征的一类匹配方法是松弛匹配方法。严格地说，松弛只是一种计算方法，通过使用局部相互制约关系和多次反复迭代来获得一个最优的整体结果。使用松弛匹配的出发点就是同时考虑所有可能的匹配，并通过迭代取舍渐渐收敛到最可能的匹配上。尽管为此付出的时间代价往往很大，但由于识别效果很好，这种方法越来越受到广泛的运用。

在分类匹配中，是要有标准模板的。识别字库存放的便是系统识别汉字范围中全部字类的标准特征数据，也称模板。显然，识别

字库不但与特征选择方法有关，而且也与匹配方法密不可分。所以，识别字库及其构建策略，最能反映识别系统的根本原理。笼统地说，构建识别字库的方法分为二大类，一类是直接通过人工分析汉字书写实例形成每个字类的模板，另一类则是经由系统的自学习，自动获取每个字类的模板。

与识别字库不同，样本字库收集的是众多不同人书写的不同汉字，要求覆盖面尽量宽广，反映汉字书写的实际情况。当然，不同目标的识别系统对样本字库的要求也会有所不同[184~189]，但作为标准，应该尽量满足通用性。样本字库的功能主要有二，一是用于构建识别字库，通过样本字库收取的具体书写汉字的归纳或训练，形成识别字库；二是用于测试识别系统，得到识别系统的性能指标。

最后，在一些系统中，为了增加整个系统的识别正确性，在单字识别完成之后，另再增加相对独立的后处理阶段，通过文本上下文关联，来纠正部分识别错误[190~198]，因为是要利用上下文信息，所以离不开各种关联语库的构建，包括词句关联、语境关联等等。这样的研究，推而广之，就会更多地涉及到汉语机器理解的内容。

1.1.4 研究现状分析

作为具体的手写汉字识别研究，我们更关心的是系统的性能效果。一般而言，无论是实验系统还是实际系统，其性能的好坏在很大程度上取决于所采用的识别策略，包括选择的特征、采用匹配的方法以及系统构建方式。实际上，为数众多的研究工作主要也是在选择特征和匹配方法上有所不同。

表 1.1.1 给出了主要已有研究实验系统的性能分析汇总。从中可以看出，识别率的变化范围是很宽的，从最低的 65.5%到最高的 99.1%；而识别字集的规模也从 4 类到 5401 类不等。如果我们仅考虑识别率在 98%以上，识别速度不低 0.5 字/秒的，那么只有五个系统。而这其中，高识别率的产生往往也是所使用样本局限性的结果，很难适用于实际情况。从已商品化或鉴定的实际系统来

看(表 1.1.2),无论从识别率,还是从识别字集的规模和速度上,都离实用化要求还有很大的一段距离。

表 1.1.1 已有主要实验系统性能分析一览表

索引	分类特征	粗分类特征	分类方法	样本来源	类数	字样本数	速度	识别率
6	笔划数、方向、背景	无	对比匹配	自己	1000	20/20	未知	92.0
8	多边形逼近	无	松弛方法	ETL-8	881	50	未知	98.8
23	笔划长度笔划间距离	无	Hopfield 神经网	未知	1	1/19	未知	85
24	笔划方向	无	神经网	未知	100	14/7	未知	81.5
28	坐标投影	无	神经网	自己	48	10/6	1.6	65.6
30	笔划数、长度	局部方向贡献密度、交叉点	Hamming 神经网	未知	50	30	3.3	85
31	像素	无	神经网	自己	50	1	未知	100
32	像素	无	神经网	ETL-9B	20	40/160	0.01	86.9
34	像素、构件分析	无	神经网	ETL-8B	956	40	未知	89.1
35	方向密度晶格	子网	神经网	ETL-9	270	100/100	未知	74
45	KL 展开	无	对比匹配	ETL-8	952	80/80	未知	74.8
48	象素、方向	无	加权距离	未知	970	20/20	未知	91
49	方向密度晶格	无	平方判别公式	ETL-8	271	130/5	1.36	98.5
50	边界多边形	局部特征	松弛方法	ETL-8	952	80/1	未知	99.0
52	方向密度	无	简单距离	ETL-8	952	10/10	未知	96.1
57	方向密度晶格	粗略晶格	对比匹配	ETL-8	881	40/20	未知	94.8
58	轮廓方向背景密度	无	对比匹配	未知	未知	未知	未知	90 88

续表 1.1.1

索引	分类特征	粗分类特征	分类方法	样本来源	类数	字样本数	速度	识别率
62	边界多边形	类似度	松弛方法	ETL-9	3036	10/1	0.6	98.55
63	边界多边形	方向段强度	动态匹配	ETL-8	952	1/1	0.14	99.6
74	局部特征	密度、背景、轮廓	对比匹配	CCL-HCCR	512	50/50	1	96.7
75	方向密度、笔划域	无	加权距离	自己	128	200/100	未知	97.0
76	方向密度	无	对比匹配	未知	200	未知	未知	90
77	像素	局部匹配 (4×4)	非线性匹配	自己	512	1/1	未知	97
78	方向密度晶格	未知	非线性匹配	ETL-8 ETL-9B	956 2956	80/80 100/1000	未知 未知	97.14 94.42
81	循环层次直方图	未知	马尔科夫链匹配	自己	5	20	未知	69
83	Hough 变换	无	动态程序设计	ETL-8	351	30/50	未知	94.5
91	像素、方向域投影	投影均值	对比匹配	ETL-8 未知 未知	881 2074 2253	80/80 25/25 100/100	未知	99.1 97.8 95.4
92	周边方向	整体、局部方向	对比匹配	未知 未知 ETL-8	2074 2253 881	5/5	未知	97.7 95.4 99.1
94	构造特征	方向密度	对比匹配	自己	2253	200/100	未知	94.9
102	Hermit 变换局部	轮廓、线条数	多级类似度	自己	1008	400/10	50	99.1
103	链码变换	无	对比匹配	未知	18	未知	未知	95.4
111	笔划方向、长度	方向投影	最小距离	ETL-8	881	80/20	未知	92.1

续表 1.1.1

索引	分类特征	粗分类特征	分类方法	样本来源	类数	字样本数	速度	识别率
114	笔划	无	未知	未知	5401	未知	3.3	81
115	笔划	无	未知	未知	5401	未知	5	90
126	笔划序列	部件	对比匹配	自己	1100	1/1	未知	100
128	笔划序列	无	树搜索	未知	300	未知	未知	90
137	笔划	部首	树搜索	未知	1500	未知	未知	85
141	笔划	笔划方向密度	最小距离	未知	2074	30	未知	99.1
143	笔划	周边链码	松弛方法	未知	240	1/37	0.33	95.3
144	线条方向	无	对比匹配	ETL-8	952	80/80	未知	90.0
153	笔划、要点	未知	模糊匹配	未知	240	未知	未知	91.8
155	外周笔划	无	树搜索	自己	50	5/5	未知	94.0
156	部件	无	文法匹配	未知	80	1200	5	96
157	部件	无	文法匹配	自己	100	0/5	未知	80.2
163	部件	未知	属性文法	自己	3755	未知	0.79	99.5
168	边界样条	未知	松弛方法	ETL-8	952	未知	0.9	98.84
171	笔划	部件	松弛方法	自己	10	100/65	0.033	96.3
172	笔划段	无	竞争性松弛方法	CCL-HCI1	465	1/1	0.28	90.75
176	笔划	无	文法匹配	自己	500	0/5	未知	82
177	笔划段	周边形状	图匹配	自己	1000	未知	0.5	90
179	笔划	无	模糊匹配	未知	881	未知	未知	96
180	笔划	汉字组件结构笔划密度函数	模糊匹配	未知	未知	未知	6.2	100
181	笔划	无	分层文法匹配	自己	72	0/33	未知	81.12
183	笔划、要点	未知	统计匹配	自己	500	6	未知	92.7

表 1.1.2　已有的主要手写汉字识别软件和产品

开发者	名　　称	日期	字数	速度	识别率
日本东芝	V595	1984	2000	50	未知
日本三洋	CLL-2000	1985	3000	5	93
日本东芝	V-3050	1986	2200	150	未知
日本富士通	FACOM6678A	1986	3200	40	未知
日本 NTT	OCR60	1986	3176	20	98
日本三菱	M6560	1987	2416	6.6	99
清华大学	交互式自学习手写汉字识别系统	1989	3755	0.5	<87
北京大学	手写印刷体汉字自动识别系统	1990	3755	0.6	<80
机械电子工业部	手写体汉字识别系统	1990	3755	1.4	<90
中国科学院	手写体汉字识别方法与系统	1991	3755	<0.9	<78
北京信息工程学院	军用手写体汉字识别系统	1991	3755	0.25	<75
武汉工业大学	手写印刷体汉字识别系统	1993	3755	1	<85
浙江大学	手写体汉字识别系统	1992	3755	0.2	71～83

目前，在手写汉字识别研究领域，对于稳定可靠特征的归纳、识别方法的完善以及提高识别速度和通用能力，依然是有待深入研究的探索性课题。不过，我们对手写汉字机器识别的前景不能抱太乐观的希望，这其中有许多实质性的困难，不是单靠研究汉字字形及其特征归纳本身所能解决的。特别是涉及到语境和整体知觉，许多表面上看来显而易见的事情，实际研究往往会遇到种种意想不到的困难。如果我们不能对人类认字能力有透彻的了解，那么手写汉字识别实用系统的机器实现，恐怕永远将是一个悬而不决的问题。

1.2　人类认字的过程分析

了解人类认字过程和规律必然有助于手写汉字机器识别的研究，这是不言而喻的。虽然目前尚不能对人类认字机制有完整透彻