文字识别原理与策略

黄襄念 程萍 著

西南交通大学出版社

前 言

汉字识别是一个超多类复杂模式分类问题，研究难点在于字符集庞大、字形结构复杂、构形变化大。近年来，汉字联机识别字体已经历了规范手写体和限制性手写体的发展阶段，研究热点集中在自然手写体识别，这是发挥它巨大市场潜力、推广和普及的关键之一。系统结构方面，多分类器综合集成系统的研究是发展趋势且起步不久，有望实现重大突破。

书中介绍了一种适用于联机识别自然手写汉字，将统计分类和结构分类、联机识别和脱机识别有机融合，带一个预分类器的三级分类器串联集成模型。针对前两级分类器要求识别率较高和误识率尽可能低的系统特性，作者设计了识别较准确、误识率极低的两级结构分类器，从多个识别角度完成汉字识别。前两级分类器发生的拒识，系统第三级分类器提取脱机统计特征作补充识别。

书中还介绍了一种适用于自然手写拉丁字母和数字等拼音文字识别的集成模型和特征编码方案。通过将输入字符图像模式矢量化处理和定义有效的方向笔段相互位置关系来描述识别对象，利用多算法集成的策略，采用匹配法完成识别。该模型具有高效实用的优点。

对于汉字识别特征编码，书中介绍 01 码方案及其全码、简码两种码制。该特征码具有抗模式变形能力强和覆盖度广的优点，特别适合于集成分类系统预分类器。在分析中文键盘输入法的几种形码编码方案基础上，结合了联机识别技术特点，对五笔、笔形和笔顺键盘码进行适应性改造，介绍了 WB 码和 SO 码两种识别码，从

信息利用的不同角度出发，完成多角度和多层次模式识别。

针对联机识别特点，介绍了一种实用高效的预处理方法，包括数据采集、字符分割、归一化、删孤点和重合点、插值加密、滤噪、断笔连接，对有关算法进行改进。将字符书写过程中存在的随机断笔定义为平直型和转向型两类，介绍相应的夹矢量法和点距法实现断笔连接。

在基元定义和提取方面，定义折点分割的非等长方向笔段特征，并介绍反映笔道全局和局部走向的α-β 动射线法，折点分割正确率为 100%。针对自然手写体汉字大量存在的习惯连笔、倒画笔和勾笔现象，介绍无效笔段概念及实现技术，滤除识别模式的冗余信息，有效地提高特征模式的稳定性，一定程度上对模式畸变起规整作用，即降低字形畸变程度，减轻后续算法的研制难度。

采取树型模式描述和分层识别技术路线，介绍对模式结构层进行层间分级判定的逐级过滤策略，集成分类过程中引入人脑智能，实现人工监控，解决系统错误累积问题。

在观察分析自然手写拉丁字母样本集及编制大量细分类函数的经验总结基础上，介绍自然手写拉丁字母字元对象概念及字元对象判别细分类算法。该算法能压缩细分类函数的代码量数十倍以上，减轻细分类函数的编制难度和结构的复杂性，极大地减轻系统开发、调试、测试、完善和维护工作量。重要的是，它能对系统训练不足作有效的补偿，降低主观不利因素的影响，大大增强系统识别率的稳定性。

最后，给出一个实验系统及测试结果，可观察相关算法和策略的有效性。

黄襄念

2002 年 9 月

目　录

第1章 绪 论

1.1 识别系统应用简介

科学技术是第一生产力。计算机科学技术是我国高技术领域的一个重要方面，随着计算机硬件的迅速发展，计算机应用领域不断开拓，急切要求计算机能够有效地感知诸如声音、文字、图形、图像、温度、振动等人类赖以生存和发展的信息资料。这些引起了计算机科学工作者的密切关注，投入了大量的人力、物力从事计算机模式识别理论与应用研究，在许多领域已经取得了令人鼓舞的成果，开发出了众多的识别系统。下面列举一些不同的识别对象或不同应用领域的识别系统：

（1）数字、拉丁字母、希腊字母、汉字等文字字符识别系统；
（2）特殊字符、数学符号、速记符号等图形符号识别系统；
（3）语音识别系统；
（4）工程图纸识别系统；
（5）三维物体识别系统；
（6）目标形状识别系统；
（7）航空图片识别系统；
（8）CT 图像识别系统；
（9）汽车牌照识别系统；
（10）手写签名识别系统；

（11）人脸自动识别系统；
（12）医学图像识别系统；
（13）指纹鉴别系统；
（14）掌纹鉴别系统；
（15）足迹检验系统；
（16）手腕骨图片识别系统；
（17）印鉴鉴别系统。

1.2 联机汉字识别的理论和实用价值

1.2.1 应用前景

（1）研究背景

随着计算机日益社会化、家庭化，工作、生活中有大量的汉字需要输入计算机。目前基本上都只得采用中文键盘编码输入法。但无论何种键盘输入法，要想获得一定的键入速度，使用者必须具备三方面的基本能力：① 熟悉输入法；② 熟悉键位；③ 良好击键指法。可见，要高效地使用键盘输入法，并非一件很容易的事，这限制了广大用户的录入速度。考虑我国的国情，目前广大的计算机普通用户，包括不少在读大学生对键盘并不熟悉，更不要说有良好的指法训练。再者，选择并熟悉某种输入法也决不轻松，必须记忆许许多多的规则。不妨以应用最普遍、最典型的拼音输入法（音码）和五笔字型输入法（形码）为例来说明。

拼音输入法：其一是重码多（同音字），挑选汉字严重影响了录入速度；其二，要求拼音准确。对那些方言较重、发音不准的用户来讲，常常为输入某一个字，反复试探输入多次才能成功；其三，遇到不认识或知其意但不知其发音的字时，就要临时换用

其他形码输入法或查阅字典或请教同事才能完成。

五笔输入法：不仅要熟记199个字根及所在的键位（不亚于记忆日文假名），而且还要熟记汉字五笔码拆分原则和方法，才能达到下意识的心手合一程度。它多用于专业录入员，速度快，重码极少，基本上可做到盲打输入，但其固有的难学难记的特点令人望而却步。不仅如此，还需要经常的、反复的、大量的录入实践才能熟记，一旦搁下，又会生疏。即便计算机相关专业人员，不会或不熟悉当不在少数。

现阶段，中文信息处理技术已作为国家优先发展的高技术重点之一。随着计算机应用技术迅猛推广、普及，越来越巨大的汉字录入量和键入速度慢的矛盾就日趋突出。在此背景下，以“不用学、不用记、方便自然、实用高效”为特征的新一代普及型汉字智能录入系统的研制愈显迫切。它与各种人工编码击键输入方法相比，其优点明显：①输入简便，不要求书写人学习记忆；②可在构思文章的同时进行输入，不打断思路，是一种“想打”型输入装置；③利用图形输入板（写字板）可方便地编辑、修改和图形输入。可以预见，它必将有着广阔的应用前景。

（2）笔式计算机

微软推出Windows CE后，不少家大公司相继推出基于Windows CE的笔式计算机（HPC）。第一代HPC虽然提供了输入笔，但不支持手写识别，仅担当鼠标的作用。Microsoft当时解释：手写识别技术尚不成熟。随着中文笔式计算机的兴起，大大促进了手写汉字联机识别技术的研究和发展。笔式计算机这种笔输入个人计算机是微机浪潮后的又一个浪潮。它取消了键盘，缩小了机器体积，笔输入、笔控制计算机。正如张忻中教授所讲“联机识别技术在今后推动笔式计算机在我国的发展和应用方面将起到决定性作用”。由于笔式计算机的巨大市场潜力和笔输入技术的不断突破，为手写识

别提供了一个进入计算机主流市场的机会。谁能在手写识别技术上有重大突破，谁将在新一代笔式计算机市场占有主导地位。

（3）Internet 和电子商务

随着 Internet 和电子商务的迅速崛起和发展，世界软件巨人十分清楚地认识到手写识别系统在未来发展中的重要作用。比尔·盖次在《未来之路》一书中将手写识别作为 Internet 未来发展的战略技术之一。

1.2.2 社会效益和经济效益

（1）社会效益

① 汉字高速录入，突破大多数人在汉字信息处理时键入低效这个瓶颈；

② 代替或部分代替键盘输入，减轻劳动强度、心理负担和精神压力，提高工作热情和效率，保护劳动者身心健康；

③ 将成为计算机标准配置和重要人机智能接口。随着计算机的迅速渗透，各行业人士希望使用简单快速、高效实用的汉字智能输入装置来完成有关文献、资料、报表等的快速处理；

④ Internet 发展迅猛，在线交流日趋平凡。提高汉字输入速度可加快信息交流，显著降低网络使用费用，增加用户上网交流信心和上网交流用户的数目和次数。

（2）潜在的经济效益

调查表明，各行业不少人士对使用该系统有浓厚兴趣。但对系统识别指标期望值较高，这也正是识别系统目前还没有真正做到实用化和大量普及的原因，需要不断研究和完善。

1.2.3 研究的理论价值

文字识别技术即计算机自动高效地识别数字、拉丁字母、希腊字母、汉字、数学符号、特殊字符、日文假名等，学科上属于模式

识别和人工智能范畴。涉及模式识别、人工智能、数字图像处理、计算机图形学、人工神经网络、专家系统、模糊数学、组合数学、数学形态学、小波理论、分形理论、系统集成、信息论、自然语言理解、认知心理学等众多科学研究领域。一个实用技术系统的完整实现，还将会涉及到硬件技术、软件工程、数据结构、接口技术、网络技术、数据库技术等技术领域。文字识别是一个综合性强、跨多学科领域的边缘学科，研究成果蕴涵理论基础和应用两部分。

目前，字符识别技术发展迅猛，研究工作非常活跃。不能或不能高效实现的算法，具有理论指导价值，但它不具有现实应用意义。如果没有基础理论研究的创新和应用，想要取得突破性的进展是非常困难的。所以字符识别技术所取得的每一步进展，都是相关理论基础和应用研究的进步。对于像自然手写体汉字识别这样一个超多类复杂模式识别问题，公开报道的实验系统种类很多，研究立足点不尽相同，反映在采用的方法和技术路线上各有特色。但是可得出这样的结论，字符识别技术已从最初的模板匹配单一技术进展到用大系统识别汉字；从统计分类、结构模式识别进展到统计—结构分析技术相结合。在传统技术基础上，不断地探索、研究、引入、拓展相关学科领域的新思想、新方法、新技术。技术创新不仅表现在局部技术上，如各种神经网络分类器、小波神经网络分类器、模糊分类器、智能分类器等，且表现在将这些局部技术作有机融合，构成综合集成分类系统，诸如各种不同集成方案的集成分类器研究。字符识别技术不断创新和突破，必然将推动该学科领域的研究发展，也必将丰富相关学科的研究。

1.3 字符联机识别的一般原理

字符联机识别是模式识别领域的一个重要分支，是迄今为止在模式识别中研究得相对较充分的一个领域。自然手写汉字识别

是联机字符识别中具有挑战性的课题和研究热点，图 1.1 是字符联机识别典型的原理框图。

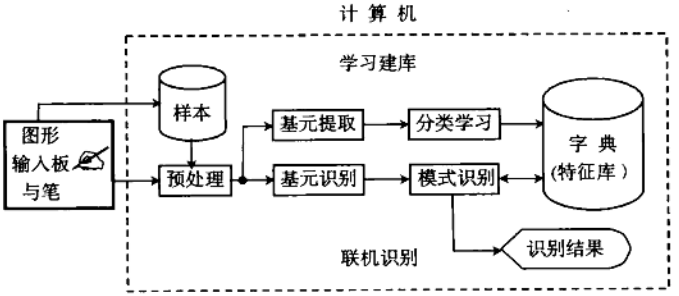

图 1.1 手写字符联机识别原理框图

一台计算机上配置图形输入板（又称手写板）即可以进行字符联机识别软件系统的研究。研制出识别软件后，将它安装到计算机内，加上手写板就构成了一套联机识别装置。所以手写板和识别算法是联机识别的关键，手写板性能如何直接影响系统识别率和书写速度。

手写板任务是实时检测写字时笔相对于板的绝对坐标，将笔画上各点坐标对(x, y)通过 RS-232 或 PS2 不断输入计算机，完成信号 A/D 转换。

识别算法的任务就是要解决文字的模式分类问题。一般通过特征提取和特征匹配来实现，这一直是文字识别的两大支柱。文字识别的历史，是特征抽取和匹配技术的发展史。采用不同的识别策略和系统拓扑结构，抽取特征的类型和数目会有所不同，可选取已经研制出的数目众多的特征或开发新特征。特征抽取的三原则：① 稳定性好。同一字符构形变化时，特征一般不变或变化小，即容忍度强或敏感性弱。特征稳定性好一方面有利于建立较完备且模板数量较少的特征库，加快查找速度；再者对提高系统识别率有重要作用。② 区分度高。即对字符的分类力强，提高特征的有效性。在不同

识别层次或级别，有不同要求。特征区分度较高，可显著地降低粗分类输出的候选字符集重码数量。③ 易于提取。若该特征不能或不易于提取或提取速度很慢，即使稳定性好和区分度高的特征也没有实用意义。特征抽取三原则彼此存在冲突，存在矛盾。如区分度高的特征，往往对字形畸变较敏感，稳定性就差一些；而稳定性好的特征，对字形变化的容忍能力较强，但分类能力就受影响。所以，需要研制者的经验和反复测试分析，综合权衡确定。例如，为降低系统误识和拒识率，粗分类时可以降低区分度要求，特征稳定性要求才是第一位的；在细分类阶段，区分度要求成为首要问题，否则，难以有效地识别相似模式。

特征匹配，就是将抽取的待识模式特征向量或特征矩阵，与系统通过样本学习（有教师或无教师）建立的特征库（识别字典）中预存的特征进行各种匹配运算，计算出最相似的模式，给出识别结果。匹配策略可分为精确匹配和非精确匹配（模糊匹配）两大类：精确匹配用相等或完全一致的策略，要求待识模式特征与特征库预存特征完全匹配。采用特征库查找技术，优点是速度快，误识率低，但相对会增加拒识率；非精确匹配应用相似判决策略，采用计算技术计算待识模式与预存模式之间的距离或类似度或隶属度等判决函数值，决定最相似模式作为识别结果。因存在大量迭代计算，减慢了匹配速度，好处是拒识率为零，但误识率会相对增加。特征匹配技术已经从最早的模板匹配发展到现在的多算法集成匹配技术，用系统观点来研究问题。如何选取和开发能反映事物本质的特征及不同算法的综合集成问题，正是有待进一步研究、不断探索和完善的课题。

1.4 结构层次与识别策略

汉字识别属于二维图形模式分类问题。对一个汉字进行机器

识别，就是把一个输入模式（单个汉字）通过识别系统分类后，指定为若干个预定义的模式类（国标两级汉字共 6 763 个模式类）中的一个（无重码输出）或几个（有重码输出）模式，给出分类结果。汉字是二维的图形文字，其分类方法主要有结构分类和统计分类两大类。无论是联机还是脱机识别，由于手写汉字的结构复杂性和模式畸变，其主流是以结构为基础的统计—结构分类法，主要对构成汉字笔画形状、位置及其相互位置关系特征进行分析研究。

对汉字这种结构组合式文字进行识别，采用以结构分析为基础的多层识别策略和体系，要求按结构进行模式划分，即结构分层。逐层用一小组结构较为简单、数目较少的基元和文法规则来描述庞大和复杂的模式。这符合人类处理复杂问题时经常采用的由简到繁，逐步求精策略。刘迎建和戴汝为从汉字识别角度，按机器识别流程，由低到高把汉字分为 5 个结构层：① 笔段；② 笔画，③ 字根；④ 单字；⑤ 词组，见图 1.2 所示。层次愈低，图形构形愈简单，数目愈少（一个量级左右），愈易提取和正确识别。

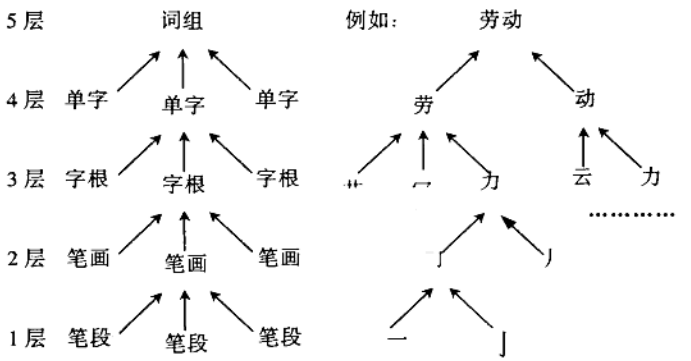

图 1.2 汉字的结构层次

不同的技术系统，分析研究问题的出发点和立场有所不同，

表现在采用识别策略和定义汉字结构分层方面不尽相同：

① 二层识别：笔段—字或笔画—字；

② 三层识别：笔段—笔画—字或者笔段—字根—字或笔画—字根—字；

③ 四层识别：笔段—笔画—字根—字；

④ 五层识别。

至于如何定义和提取笔段、笔画、字根及其相互位置特征？数目多少？如何描述汉字模式？等等。这些问题的不同回答，形成了形形色色的编码方案和技术系统。为了适应自然手写汉字字形较大的变形或模式畸变，具有较高稳定性或变形容忍度，需要对这些问题进行深入细致的研究，才可能构建出较好的识别系统。

1.5 研究任务与目标

1.5.1 技术背景

作为模式识别和人工智能一个重要分支的汉字识别技术，自20世纪70年代末，我国一些大学和科研所开展研究以来，至今已经历了20年的发展历史。20年来，从几个单位少数人探讨进展到有一定规模科研队伍在认真探索，从纯原理、方法研究进展到理论、方法、模拟实验、识别系统齐头并进。进入90年代后，无论联机识别领域，还是脱机识别领域，都取得了令人鼓舞的初步成果，达到了世界领先水平。

在汉字脱机识别领域，有多套能识别多种字体和字号的印刷体识别系统已经经过国家鉴定，识别率达95%～99.5%，并在广泛应用中不断完善。对于自然手写汉字识别，尚未见有成功的系统出现。这是汉字识别领域具有最高难度的课题，也是汉字脱机识别系统将要达到的最终目标，目前还处于理论探索和实验研究

阶段，近期内难以取得重大突破。研制高识别率、高速度手写印刷体识别系统和限制性手写体识别系统正是汉字脱机识别的现实目标和开发实用系统的立足点。

在汉字联机识别领域，目前市面上已出现了一些商品化装置，都是自然手写体识别系统，这证明了汉字联机识别技术的理论和方法已经开始进入成熟时期，识别系统开始走向实用化。这些实用系统如中国科学院自动化所研制的“汉王笔”，清华大学紫光集团开发的“紫光笔”，其他如“蒙恬笔”等，另外还有附加语音识别（专用话筒）、签名存储和识别、电子词典、绘画板等其他功能系统。可这样认为，这些系统代表了当今汉字联机识别技术的最高成就和最先进水平。但是，核心技术未见报道。

从书写的限制性角度看，汉字联机识别技术从上一阶段研制限制性手写识别已经过渡到研制低限制自然手写体识别。从识别字体来讲，研究热点已从手写楷体、手写行楷体逐渐过渡到手写行书体的研究。在汉字识别领域，常用所谓的“自然手写体”一词表达，它是衡量一个识别系统先进性的重要技术指标，迄今尚没有一个确切定义。在汉字识别领域，指绝大多数人在工作、生活中按自己的书写习惯，自然书写字体。它通常是指手写楷体、手写行楷体、手写行体，甚至包括一小部分手写草体的混合体。尽管个人手写体汉字字形变化万千，但大多数还是与相应的某一种或数种标准字体“形似”，这一点很重要。目前，对手写草书体的研究还不多，因为草体产生极其严重的字形畸变。有道是“楷如立，行如走，草如奔”，草体讲究“神”而非“形”，以目前的按“形”匹配技术，识别草体的难度很大，这是识别方法的研究方向和系统将来要达到的目标。

1.5.2 研究任务

本书主要研究自然手写汉字联机识别方法与系统，但作为一个较完整的联机识别系统，数字、拉丁字母识别也是系统的重要

组成部分。所以，数字和拉丁字母的识别问题也是本书内容。对于一个以实用为目标的识别系统，我们更关心系统性能效果。一般而言，无论是实验系统还是实际系统，识别性能好坏在很大程度上取决于采用的识别策略，如汉字描述、预处理方法、特征抽取、匹配方法、分类特性和系统拓扑结构。实际上，这是大量研究工作的主要不同点所在，本书主要研究任务如下：

① 研究识别理论和系统拓扑结构，构建自然手写汉字联机识别实验系统；

② 研究自然手写数字和拉丁字母联机识别算法，建立高识别率、高可靠性实验系统；

③ 研究系统集成有关问题，提出一个有效的多分类器集成模型；

④ 综合研究汉字构形特点和各种形码编码方案，论证与联机识别技术结合的有效性和适应性改造问题；

⑤ 结合识别算法和系统拓扑结构，研制稳定性好、分类力强、易于提取的识别特征及算法；

⑥ 研究基于多层识别体系的笔画、部件、单字分类／识别算法；

⑦ 综合研究各种精确匹配或非精确（相似）匹配算法。

1.5.3 研究目标

从实用出发，研究特色在于采用目前先进的多分类器集成算法，研究自然手写汉字联机识别问题。综合研究现有描述文法、预处理方法、特征编码和提取、精确/模糊匹配、分类器特性、算法集成，针对自然手写汉字识别难点（字符集庞大、字形结构复杂、模式畸变大），提出和改进相关算法。

1.6 技 术 路 线

如前所述，尽管手写体汉字识别已取得了许多成果，特别是