文章编号:1000-582x(2000)03-0087-04

24

87-90

自然手写汉字五笔码识别法*

TP391.41

黄襄念,程 萍,杨 波,黄 敏,龙辉敏
(重庆大学 人工视觉实验室,重庆 400044)

摘 要:在深入剖析五笔字型键盘输入法编码原则和字根结构基础上,结合联机识别技术特点对五笔字根作出适应性改造后,提出了一种联机识别自然手写汉字新方法:五笔码识别法。在构建的识别体系中提出和采用了层间分级技术,并提出将键盘输入技术与联机识别技术有机融合,为联机识别自然手写汉字探索新途径。

关键词:联机识别;汉字识别;识别方法;笔画分类
中图分类号:TP 391.4 **文献标识码**:A

自然手写汉字
五笔码识别法

随着计算机日益社会化、家庭化,工作、生活中有大量的汉字要输入计算机。目前,基本上都是采用键盘输入。但是无论何种输入法,要想获得一定的键入速度,使用者必须具备最基本的两方面能力:1)熟悉键盘;2)熟悉输入法。大多数人对计算机键盘并不十分熟悉,不仅要求记忆键盘上每个键位,而且要求良好的指法训练;选择并熟悉某种键盘输入法决不轻松。以应用最为普遍最典型的拼音输入法和五笔字型输入法为例,拼音输入法不仅重码多,而且要求拼音准确,对于方言较重、普通话不标准的人来讲就更难了。五笔字型输入法多用于专业录入员,尽管录入速度快,但其难学难记的特点令人望而却步,即使软件工程人员,不熟悉者也不在少数。可见,要高效地使用好键盘输入法决非易事。因此,自然手写汉字联机识别系统的研制日益显露出重要性。可以预见,它必将有着广阔的应用前景。

迄今,提出的键盘编码有500余种,在机器上实现并已商品化的就有好几十种。键盘输入技术的研制和开发时期早,使用时间长,特别是对汉字结构特征和编码方案进行了长期研究、积累了丰富的使用和研制经验,有着众多研究成果。为此,在充分分析前人研究成果基础上,笔者结合联机识别技术特征,将两者有机融合,并采用以结构分析为基础的模式识别方法和多层多级分类方案。

1 技术难点

联机识别手写汉字是字符联机识别中难度较高的问题,这也是为何这么重要的问题至今尚未完善解决的原因。其复杂性可归纳为以下三个主要方面:

1)字符集庞大

GB2312-80汉字集一级3 755个、二级3 008个,两级共6 763个(不包括繁体字),它覆盖了常用汉字的99.99%。所以,识别系统应该以两级国标汉字作为识别字符基本集,视具体应用还可能需要加装扩展字符集,以实现生僻字和繁体字的输入要求。它比拉丁文字的识别字符集要庞大得多,大大增加了识别系统复杂度和研制难度。

2)字形畸变

汉字源远流长,字体种类较多;常用的有楷体,行楷体,行体以及草体等。自然手写体字形更是千变万化,由于个人书写习惯、风格、认真程度、受教育程度、情绪、书写条件等因素作用,使每个人的书写风格都不尽相同,十分杂乱,单字构形存在无限多种可能的形状。

3)结构复杂

汉字是图形文字,最终由笔画组成,而笔画间相互位置关系十分灵活。特别是自然手写汉字,某些类别的笔画位置关系很不稳定。不象英文单词是由字母按从左到右顺序排列,只要正确识别出52个大小写字母

* 收稿日期:1999-08-30
作者简介:黄襄念(1964-),男,四川自贡人,重庆大学博士生,主要从事计算机图形文字识别系统领域研究工作。

即可识别单词。

上述因素的组合作用所造成的困难可以想象。尽管目前手写汉字识别有一定的成果，但距自然手写识别的系统目标，还有一段相当长的路要走，寻找识别特征和匹配算法是汉字识别的两大支柱问题。

2 五笔字型与联机识别技术

五笔字型属形码编码类型，把汉字拆分为字根，字根拆分为笔画，按字根编码输入汉字，其策略与汉字联机识别技术的多层分类识别十分接近，由于汉字的复杂性，整字级识别汉字的难度很大。因此无论是联机还是脱机识别，研究较多的是多层分类策略[1-7]，化复杂为简单，逐步求精。但是，怎样分层、定义和选取字根、笔画；数目多少；怎样定义字根之间的位置关系。描述方式和匹配文法等问题的处理，不同的识别系统其技术特点各不相同。

2.1 五笔字型的字根编码

经过优选，它定义了 199 个字根和横、竖、撇、捺、折五种笔画，故称“五笔字型”[8]，用字根进行不同方位组合可以形成全部汉字。它把 199 个字根按其首笔画类别分为 5 个大类（称为区），又将每个大类按字根的第二笔类型分为 5 个小类（称为位），每一个小类的几个或十几个字根都安排在一个键位上，具有相同的字根编码。这样就得到了每个五笔字根区位码 11 ~ 15、21 ~ 25、31 ~ 35、41 ~ 45、51 ~ 55，第一位数字是区号，第二位数字是位号，见文献[8]“五笔字型键盘字根总表”。

按书写顺序将汉字字根“敲”入计算机，有时存在重码汉字。分为两类：1）字根分解式相同：例如字根“日”“九”和“口”“八”，按同样的分解式，可以构成“旯”“旭”和“叭”“只”；2）字根分解式不同；由于 199 个字根安排在 25 个键上，每个键上有几个到十几个字根（它们的区位码相同），当这些编码相同的字根与某些字根组合时会产生重码；例如，在同一个键位上“木”“丁”“西”左边与“夕”组合可以构成“沐”“汀”“洒”。所以，要加入所谓的“末笔交叉识别码”，即字根间位置关系（字型码）和末笔笔画类型（末笔码）加以区分。

五笔字型认为上述交叉识别码仅仅在由字根数目较少（少于 4 个，但很多是常用字）组成的汉字中才起作用，信息量较少而产生重码；如果汉字拆分的字根数目较多（大于 4 个），信息量已经足够，不用加入该码，取前 3 个字根码和最后一个字根码即可。

2.2 字根独立编码

五笔字型键盘输入法，要把 199 个字根安排在 25 个键位上，每个键上就有几个到十几个字根，不可能把每个字根分别安排于一个键位上，用 199 个键。同键字根必然编码相同，不可避免地造成重码而加入附加识别码来区分。联机识别则不受键位限制，可以采用独立编码，即每一类字根编码不同，避免五笔字类的第 2 类重码，这样的字根特征码序列无疑提高了对汉字的分类能力。

2.3 缩合形似字根

五笔字型字根是汉字印刷体字根，没有变形和简化等现象，非常规范。联机识别时情况有些变化：如五笔字型认为形似字根“刂　、“子　”和“日曰”等是不同字根（尽管它们是同键字根），但手写汉字难以区分它们的差别，如果要强制区分，会导致两种结果：1）限制使用者的书写习惯和书写风格，显然令人难以满意；2）误识或拒识；系统认为书写者写错了。所以，应该把它们缩合为一个字根，但要注意，某些形似字根则需要区分编码，如“主王”“大犬”等。如果它们是单独的成字字根，重码字可以交由第 3 层细分类提取局部特征予以识别；否则，不用区分它们，对汉字识别没有影响。

3 技术优越性

该系统采用适应性分类改造后的五笔字根作为识别字根，主要优越性体现在：

1）字根组字能力强

经过长期使用实践证明，五笔字根组字能力极强，能很好地满足识别字符集要求，即可以组成识别字符集（两级国标字库）的所有汉字。本系统采用五笔字型字根作为识别字根，结合联机识别特点，对部分字根作了适应性分类改造，不影响组字能力，所以具有同五笔字根一致的组字能力。

2）识别重码极少

众所周知，五笔字型在众多输入软件中被首选为专业打字软件，正是因为它几乎无重码，基本上可以作到盲打输入。在这一点上，本系统具有五笔字根组字的优点。另外，如前所述本系统改造后的五笔字根，每类字根编码独立，使其分类能力更强大，大大降低编码重码率。但识别对象不是印刷体而是自然手写体，所以存在由于字形畸变带来的重码，大多是形似汉字，数目很少。对此可以采用细分类函数方法，抽取重码汉字的特殊局部特征加以区分，实现无重码输出。

3）需字根相互位置关系特征

字根相互位置关系大致可以分为左右、上下、内外和杂合等类型，有些系统中认为这是一种重要特征。但怎样定义字根间位置关系、描述方式和匹配文法等

问题，不同的识别系统采用不同的技术，增加了系统时空复杂度。五笔字型不用字根相互位置关系码就达到了极高的区分度，实现无重码输出。所以，本系统不必提取这个特征。由此产生的少量重码汉字，交细分类阶段处理。

4　识别体系与策略

由于采用多层多级识别策略，把复杂问题分解简化、逐步求精，系统体系见图 1. 可以看到，系统在三个层次上进行分类，即笔画层，字根层和单字层，层间设立粗细两级分类，建有层特征码库。技术策略思想主要在于三个方面：

1）减少错误积累，降低组合复杂度

笔者提出的层间两级判定（粗细分类）策略，图 1，体现逐级过滤思想。有些系统也设置有粗细两层分类，其细分类过程位于对整个汉字的匹配或模糊查找识别之后，相当于系统中第 3 层的单字细分过程，专门对重码汉字进行。在识别存在字形畸变的自然手写汉字时，这样的重码汉字会较多，因为存在多种影响因素，如字形畸变大或特征码序列较短（选取的特征数目不多）或难以找到更有效特征或粗分有误等等，导致该层粗分类输出中产生误识，如果把它们遗留到最后第 3 层去处理，会大大增加分类难度和重码数目，出现一些意想不到的奇怪现象，最后才进行细分类，就有可能难以找到区分度大、稳定性好的特征以识别变形校大的汉字，甚至因细分类复杂度太大而导致难解或组合爆炸。考虑到笔画层和字根层的类别数目很少（如笔画层只识别几类基本笔画），此时进行细分类比较容易，所以笔者提出层间分级技术，采用细分函数人工代码来减少错误积累问题，降低系统组合复杂度，为下一层处理提供良好输入，最终提高汉字的正确识别率。

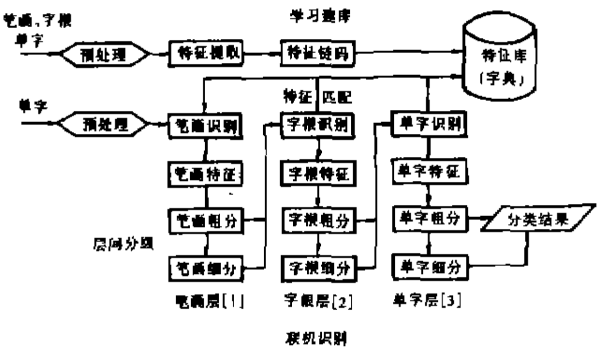

图 1　系统体系结构

2）特征分类性能可视化

各层任务相对独立，本层识别结果的输出作为下一层识别的输入。这样及时发现、单独调试每一层的识别效果，得到各个影响识别率的因素，为衡量和改进特征选取、匹配算法的好坏提供可视化依据。

3）不同的特征码序列和匹配算法

分层独立识别，各层可以采用仅该层有效、有针对性的不同特征及其匹配算法，而不是对整个汉字采用统一的特征码序列和匹配算法。合理利用特征，对系统的时间和空间复杂度特别是识别率来讲都是非常有利的，作到了有的放矢。

5　实验结果

笔者用采集的 100 个大学生书写的 6 763 × 100 个汉字作为测试样本对系统进行测试，实验结果为：系统平均正确识别率 97.42%，平均误识率 0.77%；平均拒识率 1.81%. 实验结果证明了笔者提出的五笔码分类法的有效性。本文提出的层间分级技术，使单字层粗分类结果的重码量由 130 多个减少到 25 个以下，降低了细分函数的分类模糊度，大大简化了后续单字识别细分函数的研制难度和代码工作量。

6　结束语

本文介绍了一种实验系统的体系结构和识别策略，结合笔者在数字、英文字母、数学符号、特殊字符等领域已取得的研制开发经验，构建了以改造后的五笔字根为基础的分类系统。至于其他键盘编码方案在汉字联机识别系统中的实现和五笔编码与其他字根编码的对比论证工作有待进一步研究。

参　考　文　献

[1]　刘迎健．在线手写汉字识别的字形结构排序法[J]．自动化学报，1988，14(3)：207～214．

[2]　赵明．手写印刷体汉字部件的抽取[J]．中文信息学报，1988，2(4)：59～64．

[3]　赵明．手写印刷体汉字识别方法综述[J]．计算机研究与发展，1993，(4)：59～64．

[4]　贾永康．识别联机手写体汉字的多级分类法[J]．信号处理，1995，(12)：32～34．

[5]　周昌乐．一种手写汉字拓扑图表示及其动态获取[J]．计算机科学，1996，23(5)：60～62．

[6]　余楚中．联机手写体汉字识别中的笔画分类及笔画识别[J]．重庆大学学报，1998，21(2)：131～134．

[7]　赵学军．手写数学符号的基元识别方法[J]．重庆大学学报，1998，21(2)：117～124．

[8]　宁爱军．电脑汉字处理速查手册[M]．辽宁科学技术出版社，1996.82～94．

Five-Strokes Encoding Recognition of Unconstrained Handwritten Chinese Character

HUANG Xiang-nian, CHENG Ping, YANG Bo, HUANG Min, LONG Hui-min

(Laboratory of Artificial Vision, Chongqing University, Chongqing 400044, China)

ABSTRACT: Keyboard input techniques are introduced into on-line recognition of Chinese character with the study of chinese character five-strokes encoding keyboard input method, and a new approach for on-line recognition of unconstrained handwritten Chinese character, Five-Strokes Encoding Recognition, is proposed. Besides, the set of character-roots of Five-Strokes have to be altered slightly to meet the on-line recognition technique's features.

KEYWORDS: on-line recognition; Chinese characters recognition; recognition method; stroke classification

(责任编辑 张小强)

(上接 86 页)

Application of Weathered Coal Granular Water Purifier in Printing and Dying Waste Water Treatment (Ⅱ)

WANG Nan, LIANG Zhu, CAO Jian, LIANG Zhong, YAO shu-sen

(College of Environment & Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China)

ABSTRACT: The decoloring rate and the optimal adsorption wavelength of six kinds of dye by humic acid typed granular water purifier is determined. It is proved that the decoloring rate can get to 80%. The adsorption dynamics is further studied, which is a foundation of the industrial application of HA water purifier.

KEYWORDS: weathered coal; humic acid typed waste water purifier; the treatment of printing and dying waste water

(责任编辑 张小强)