汉字字形形式化描述方法研究*)

林 民[1,3] 宋 柔[2]
(北京工业大学计算机学院 北京 100022)[1] (北京语言大学信息科学学院 北京 100083)[2]
(内蒙古师范大学计算机与信息工程学院 呼和浩特 010022)[3]

摘 要 本文分析了目前汉字处理应用中存在的主要问题,归纳出问题的核心是由于缺少能涵盖一切可能汉字的、可计算的字形形式化描述体系,从而造成应用中有一系列障碍。发现了现有字形描述方法共同存在的特征选取缺陷,最后给出了一种可行的汉字网格字形描述方法,该方法不仅能表示一切可能的汉字字形(包括错字),而且为字形特征异同的自动计算奠定了可靠的基础。

关键词 汉字字形,形式化描述,网格字形,特征计算

Formal Description of Chinese Character Glyph

LIN Min[1,3] SONG Rou[2]
(College of Computer Science & Technology, Beijing University of Technology, Beijing 100022)[1]
(College of Information Sciences, Beijing Language and Culture University, Beijing 100083)[2]
(College of Computer & Information Engineering, Inner Mongolia Normal University, Huhhot 010022)[3]

Abstract This paper analyzes the main problems existing in the Chinese character information processing applications, and concludes that the core of the problems is due to the lack of a formal description method of Chinese character glyphs which is computable and can cover all Chinese characters at the same time, resulting in a series of obstacles in the applications. Secondly, finds existing feature selection defects in current formal glyph description systems. Finally, proposes a grid description method of Chinese characters, which can not only describe all Chinese characters, including typos, but also support reliable automatic computation of the differences and similarities of Chinese character glyph features.

Keywords Chinese character glyph, Formal description, Grid glyph, Feature calculation

1 引言

从1974年"国家748工程"至今,汉字处理领域取得了长足发展,涌现出多项重大成果,如汉字激光照排、汉字标准字符集、汉字的编码输入、整句输入、印刷体识别、手写识别等。汉字字形处理技术的成功带来了我国出版业和办公事务处理的革命,极大地促进了我国信息产业的发展和社会信息化水平的提高。但是,汉字字形处理方面仍存在许多问题没有很好解决。

·国际标准字符集越做越大,但仍然不能满足人名、地名、方言字、古籍字输入的需要。目前正在全国各地发放的第二代身份证,仅北京就有数万个身份证因人名、地名用字是字库中没有的冷僻字而无法发放。

·各种汉字输入法都需要使用者对汉字的字音或部件(字根)、笔画有一定的认知才能掌握,但是没有接触过汉字文化的人(如外国人)不具备这种认知,无法使用这些方法输入汉字。

·各种汉字识别输入(包括手写识别输入)软件,受到识别原理的限制,只能识别训练集内的汉字[10],无法输入集外的汉字。

·国内语文教学和汉语国际推广都需要对错字进行定量分析,如错字的描述、校对、分类、界定、计算机辅助汉字书写学习、书写水平标准化自动评测等,但目前计算机没有错字输入和比较的方法,严重制约了对错字进行深入定量分析研究的水平。

·汉字文本校对、汉字识别的后处理、涉及汉字的历史文化研究(如古籍字、异体字的比对、界定)等应用都需要分析汉字字形的相似性,但目前也没有支持这种字形分析计算的有效方法。

·独立建立的集外字表因缺少有效比对工具而难以共享和归并。许多出版部门利用各种造字工具建立了集外字表,但因缺少有效的字形比对工具,各部门间的集外字表无法归并和共享,以致数字化图书只能各自使用独立的字库,造成资源和空间的浪费。

·各种电子出版物以至网络出版物中有许多集外字,这些字可以输出,但读者无法输入,从而也无法查询、检索包含这些字的内容。

·目前汉字字模的生成需要书法家写出字,再通过TrueType方法来描述,工作量极大,以致计算机上虽然带有很多字体,但多数只限于GB2312的字模范围,亟需字模快速生成工具的支持来建立各种字体的大字符集字模。

·汉字排序没有一以贯之的原则。按音序排列无法解决

*)本项研究得到国家自然科学基金项目(60272055,60572159)的资助。林 民 博士生,副教授,主要研究方向:人工智能、自然语言处理;宋 柔 教授,博士生导师,主要研究方向、人工智能、语言信息处理。

同音字和多音字问题;按部首、笔画排列无法解决同部首、同笔画字的问题。

这些问题和困难严重地阻碍了社会管理信息化、数字图书馆工程、汉语教学与国际推广等多项事业的发展。以上问题的核心是目前汉字缺少一致有效的字形形式化描述方法,从而无法对字形特征进行完整细致的刻画和自动分析处理。前人对汉字字形描述做过很多工作,一般方法是把汉字的构形方式按照人的认知分类,并使用人认知的部件、笔画来描述。这些描述确实对相当多的汉字有效,但也存在着大量的歧义和描述缺失,无法自动计算字形特征,因而无法解决上述问题。

因此,建立一种统一有效的汉字字形形式化描述体系和基于该体系的字形特征计算方法,能涵盖一切可能的字形(包括正字和错字),能表示各种字形骨架的异同,是一项非常重要的基础性工作,在应用上既能作为一种大众生僻字输入工具,又能支持字形特征的自动计算,从而解决汉字处理面临的一系列障碍,具有重要的意义。

2 汉字字形描述方法分析

2.1 主要汉字字形描述方法简介

许多学者看到,目前汉字字形处理存在的问题是把一个汉字整体作为编码单位来处理,这样就无法分析、计算其内部成分。实际上,汉字字形是可以分解来划分结构类型的,并且以部件、笔画作为基本的构形单位[3,13]。例如:"落"第一层可分解为由"艹"和"洛"组成的上下结构,其中"艹"可再分为构成它的三个笔画"一丶丿",而"洛"又进一步可分解为由"氵"和"各"组成的左右结构。如此逐层递归分解下去,直到所有成分都分解为笔画为止。从这样的观点出发,有一批研究成果。代表性成果介绍如下。

(1)汉字信息字典[1]

上海交通大学汉字编码组编,科学出版社 1988 年出版。主要特点是将汉字递归地分解成部件和笔画的组合,组合的结构类型有左右、上下、包容、嵌入 4 种,描述了 7785 个汉字。

(2)汉字部件规范(GF3001-1997)[2]

国家语言文字工作委员会于 1997 年 12 月 1 日发布的《信息处理用 GB13000.1 字符集汉字部件规范》,主要特点是穷尽式地列出了国家标准通用多八位编码字符集中 20902 个汉字的部件表,这 20902 个汉字已经根据这些部件进行了逐个拆分。

(3)表意文字描述序列 IDS(Ideographic Description Characters Sequence)[4]

Unicode 联盟于 2000 年提出的表意文字描述符系统,作为 Unicode 4.0 标准。主要特点是将汉字递归地分解为部件的组合,组合的结构类型有 12 种。将结构类型符作为操作符,汉字或部件作为操作数,组成前缀表达式,可以表现 Unicode 集内的绝大部分汉字和一些集外汉字的字形。从实现的效率考虑,对表达式的长度和其中连续排列的部件个数有限制。部件集合不固定,同一汉字的描述方法也不固定。

IDS 的出现,反映了 ISO 也认识到单纯用扩充编码的方法支持更多汉字是行不通的,只有从汉字的构形出发,才能真正解决汉字的计算机表示问题。

(4)汉语文档处理语言 CPL(Chinese Document Processing Language)[5]

台湾中央研究院信息技术研究所文献处理实验室在 1990 年代开发,为古籍整理服务。主要特点是将汉字递归地分解成部件和字根的组合,组合的结构类型有直连、横连、包含三种,还有几种重叠形式。确定出 1 千多字根表现 4 千多部件,涵盖了 5 万多字形,并使用 CPL 作为其研发的汉字构形数据库的字形描述语言。

(5)汉字数学表达式[6,7]

国防科技大学孙星明、殷建平、陈火旺等于 2002 年提出,主要特点是将汉字递归地分解为部件的组合,组合类型有 6 种(左右、上下、左下包、左上包、右上包、全包含),固定出 505 个部件,并给出了关于结构类型的结合律和传递律,使得字内任意两个部件的结构关系能通过逐层的推导而得以确定。

(6)字符描述语言 CDL(Character Description Language)[8]

美国加州大学伯克利分校研究人员创办的文林研究所于 2003 年提出了基于笔画和汉字部件的字形描述系统,并采用 XML 作为元语言。主要特点是将汉字递归地分解为部件的组合,最底层的部件是笔画。CDL 没有结构组合类型的概念。它处理部件间位置关系的核心思想是:每个部件有一个隐藏的外包矩形轮廓,可以通过改变外包矩形斜对角顶点的坐标来达到移动和缩放对应部件的目的。小部件(可能是笔画)的外包矩形移动和缩放后成为大部件或整字。CDL 笔画集合是固定的,笔画的形状用它的起点、终点、拐点的横、纵坐标,以及两点间笔段的走向和弯曲方向表示。CDL 没有固定的部件集合,所以它描述字形有极大的灵活性,可以描述各种可以想见的汉字,可以表现异体字的特异性。另一方面,它对笔画的描述十分细致,不仅表示出了形状,而且表示出了走向和弯曲方向,所以可以用于汉字书写方法的教学。

2.2 字形描述方法特点评述

(1)字形特征描述能力方面

以上成果的各种字形描述方法共同点是都把汉字看作大部件到小部件的递归组合,因而有很强的字形能产性,一定程度上克服了大字符集方案的封闭性弊病。

汉字部件规范给出了现有标准字符集内汉字的所有部件,但难以保证对集外汉字的支持。

IDS 划分结构类型不固定部件,有很大灵活性,但缺少规范性,其中"覆盖类型"的构形描述很模糊,难以据此构建确定的字形。例如"幽"是覆盖结构,还是下三包结构见仁见智,难以把握。

CPL 和汉字数学表达式都不仅划分结构类型而且固定了部件,比较规范,但也在一定程度上丧失了灵活性,字形产生能力有限。

这几种描述方案共同存在的问题是:都是面向人的字形描述体系。字形拆分的主要原则是汉字的字理,这些原则对于一般大众和没有汉字文化背景的外国人而言仍很难使用;另一方面,描述中都采用了汉字教学中引入的结构类型思想[11],这一思想对于面向人的教学确实很有效,但是不适合进行计算机处理。因为相当多汉字的结构类型是有歧义的,有些字到底是上下结构还是包围结构或是独体部件,依赖于人的认知。比如"着"是左上包围结构,而"眷"不看成左上右三包围结构却看成上下结构,就很费解。"乘"看成"北"包围"禾","裹"看成"衣"包围"果",则需要专业知识。"卡"的中间一横应归在上半部还是归在下半部,则是见仁见智。一个字由于结构类型的认知不同,从而描述不同,会被计算机误识成两个字。而拆分标准不统一、不规范,也难以被机器实现。

相对而言，CDL只固定笔画不固定部件，直接列出位置坐标而不划分结构类型，一定程度上避免了前几种方案的缺点。同一个汉字可以用不同的结构表达式表示，只要最终的笔画相同，仍可以被认同。它的另一个特点是信息量大，不仅表现了静态的字形，还表现了动态的笔向，可用于汉字书写方法的教学。

但是，CDL固定笔画，仍然会有问题。一个问题是：不同的人对笔画的认知不同，会造成歧义。比如“果”。若从字源看，应看成“田”加“木”，此时有两个竖笔；若从一般人的书写习惯上看，往往是“木”的竖笔向上伸到“日”中，只有一个竖笔。又如“卑”，有些人认为中间是一个“田”，下面是“十”，“十”旁有一撇；有些人则认为“田”的一竖应该是一撇，延展到“十”的旁边。笔画分解不同，一个字表现成两个字。

CDL固定笔画的另一个问题是影响了对汉字的表示能力。因为大量的异体字，特别是汉字教学中遇到的错字，其笔画形状是难以列举的。

CDL的再一个缺点是字形描述方式太复杂，只能为专业人士所使用，难以被大众使用。但是，电子出版物的检索者往往是大众，对于需要检索的生僻字，他们需要用一种字形描述方式来告诉机器，CDL难以担当此任务。

(2)字形同一性认定及特征计算能力方面

CDL用横、纵坐标表示部件的绝对位置，比起用上下、左右、包围等相对位置关系要确切、灵活，但是同一个字中同一个部件的位置可能被不同人描述得不一样。虽然可以用软件来做同一性认定，但计算工作量会相当大。而且，这种同一性认定能力也是有局限性的。比如，“土”和“士”很相似，但它们是两个字，差别在于两横长度的对比。但是“青”上面两横无论哪个长哪个短，都是同一个字。CDL无法区分这两种情况。

这个例子说明，字形的同一性认定必须是人机结合来实现的，机器提供相似字，人来判别是否同一。基于特征计算的字形相似性比较功能是字形同一性认定的基础，但CDL和其他几种方案以结构类型、部件或笔画作为基础特征，基本特征的颗粒度过大，影响了特征抽取和比较的效果。“果”的两种描述方法结构类型不同，但表现的是同一个字；“我”和“找”特征很相似，但是由于字中间是一横还是两横造成两个字的描述(结构类型、部件)有极大差别。又如“单”同“草”、“卓”相似度很高，但中间一竖穿进了“日”，造成描述结果都相差非常大。对于这类差别，这几种方案都能表现出来，使得不同字形不被混淆，但不能很好表现出相似性的程度，因而也不能真正解决字形同一性认定问题。

从实际应用来看，基于特征计算的字形相似性比较功能有着大量的应用需求，应当是字形描述系统具备的重要能力。但是，由于CDL等字形描述方案在字形特征计算能力方面存在的局限，难以支持这些应用。

(3)字形描述、特征计算能力与汉字输入、输出处理能力的关系

手写识别方法[10,12]依靠字形特征来识别汉字，描述了训练集内不同字形的差别特征，对训练集内字的字形区分处理能力很强，但忽略了字形间相似性特征的描述，特征相似性计算能力不强，又缺少训练集外字的特征，限制了对集外汉字的识别输入能力。

由此可见，只有具有强大特征描述能力和特征计算能力的汉字字形处理系统，才能彻底解决汉字的输入、输出处理问题。

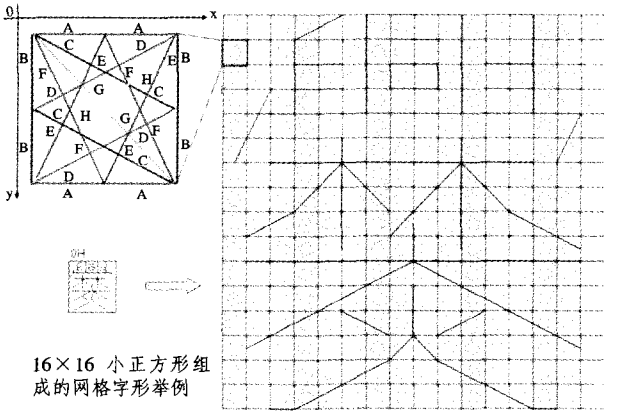
3 汉字网格字形描述

汉字网格字形是一种采用平面网格结构来描述汉字字形本质特征的形式化方法。网格结构描述定义如下：

- 汉字字形网格是 $n*n$ 个小正方形合成的大正方形。
- 每个小正方形区域内含有28个笔段：
  - 各顶点到它所在边中点的连线，共8段；
  - 各顶点到它对边中点的连线，以连线中点为界分为2个笔段，共 $2*8=16$ 段；
  - 对角线以中点为界分为2个笔段，共 $2*2=4$ 段。
- 整个网格有 $24n*n+4n$ 个笔段。
- 网格中每个笔段可取有线或无线两种状态，所有有线的笔段分布构成了汉字的网格字形描述。

实验表明，$n=12$ 时可以表示绝大多数汉字的字形，$n=16$ 时可以表示所有汉字的字形。

在网格字形描述体系下，笔段为基元，古今中外所有的汉字(包括异体字、错字)，都对应网格中有线笔段的一种分布(但并非任意的有线笔段分布都是一个可能的汉字)。有线笔段的不同分布可能对应具有相同结构或相似结构的汉字。例如“爨”采用 $16*16$ 的网格字形描述，如下图。



16×16 小正方形组成的网格字形举例

结束语 网格字形描述方法，用定义好的有限方向的直线段——笔段作为描述单位，特征规范、颗粒度适当，克服了现有汉字字形描述体系存在的以下问题：

1)以结构类型、部件、笔画作为描述单位，特征颗粒度过大，特征值集合固定，无法表现开放的汉字集合，无法比对一些相似字的差别。

2)点阵字形或TrueType曲线轮廓字形的描述基元颗粒度太小，规范性差，包含许多因书写工具不同和美学观念不同带来的书法上的差别，这些差别不是汉字字形的本质差别。

在易用性方面，没有任何汉字文化背景的外国人和文化水平不高的各类录入人员，只要对于图形能正常认知，就能使用支持网格字形描述的系统来画出他所想到看到的汉字。

在支持字形特征自动计算方面，网格字形描述只保留汉字的骨架而剔除字形中非本质的书法特征，一方面可降低字形特征计算的复杂度，另一方面可使计算结果更加可靠。也容易与现有字形描述体系建立对应关系。

因此，该方法是一种既可作为大众生僻字、错字输入工具，又能涵盖一切可能字形并支持特征自动计算的有效字形形式化描述方法。

参 考 文 献

1 上海交通大学汉字编码组. 汉字信息字典[M]. 北京：科学出版

社，1988
2 国家语言文字工作委员会. GF3001-1997 信息处理用GB13000.1字符集汉字部件规范. 北京：语文出版社，1997. 12. 1 发布，1998. 5. 1 实施
3 王宁. 汉字构形理据与现代汉字部件拆分[J]. 语文建设，1997(3)：4～9
4 Ideographic Description. http://www. unicode. org/versions/Unicode4. 0. 0/ch11. pdf. 307～309
5 http://www. sinica. edu. tw/～cdp/(台湾中央研究院 信息科学研究所 文献处理实验室网站)
6 孙星明，殷建平，陈火旺，等. 汉字的数学表达式研究[J]. 计算机研究与发展，2002，39(6)：707～711
7 张问银，孙星明，曾振柄，等. 汉字数学表达式的自动生成[J]. 计算机研究与发展，2004，41(5)：848～852
8 Cook R. A Specification for CDL(Character Description Language)：an extract of：[PhD Dissertation]. UC Berkeley，Dept of Linguistics，2003
9 http://www. wenlin. com/cdl/(美国加州大学 伯克利分校 文林研究所网站)
10 http://www. xiaoyaobi. com/(北京逍遥笔模式识别工作站网站)
11 梁彦民. 汉字部件区别特征与对外汉字教学[J]. 语言教学与研究，2004(4)：76～80
12 陈良育，曾振柄，张问银. 汉字构形分析与识别[J]. 上海电力学院学报，2005，21(1)：63～65
13 冯志伟. 用上下文无关语法来描述汉字结构. 语言科学[J]，2006，5(3)：14～23

(上接第 169 页)

他文献中的算法在使用相同数据库条件下的分类结果。表 2 给出了本文的方法和文[13]所列出的三种方法在 0°视角下的识别结果。从表 1 和表 2 看出，我们提出的方法相比之下可以取得较为满意的结果。此外，图 5 使用国际上在人脸识别算法中通用的分类性能度量方法 ROS(rank order statistic)[14]，给出了在 90° 视角下的 CMS(cumulative match scores)即累积匹配分值图。

表 1 三个不同视角下算法正确分类率的比较

分类器	算 法	CCR (%)		
		0°视角	45°视角	90°视角
NN	文[4]的方法 (STC 度量)	65.00	63.75	77.50
	文[4]的方法(NED 度量)	65.00	66.25	85.00
	文[5]的方法	65.00	60.00	40.00
	本文的方法	73.75	66.25	80.00

表 2 0°视角下各算法识别结果的比较

算 法		Top1 (%)	Top5 (%)	Top10 (%)
Wang 等的方法[13]	STC, NN, No validation	65.00	—	—
	STC, NN, With validation	68.75	—	—
	NED, NN, No validation	65.00	—	—
	NED, NN, With validation	70.00	—	—
BenAbdelkader 等的方法[2,13]		72.50	88.75	96.25
Collins 等的方法[3,13]		71.25	78.75	87.50
本文的方法		73.75	91.25	97.50

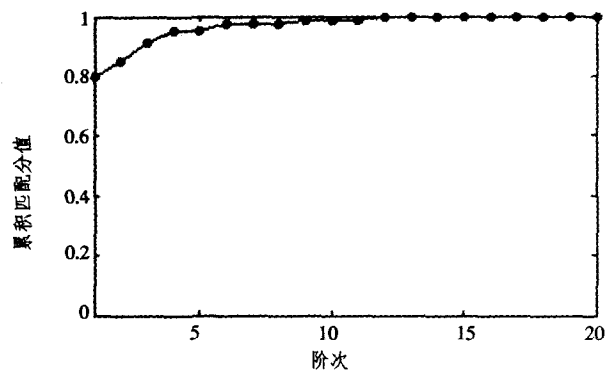


图 5 90°视角下累积匹配分值图

结论 近年来步态识别已经成为生物特征识别领域中的研究热点。本文依据人体解剖学的知识将人体步态侧影图像分为 5 个子区域，在提取每个子区域的不变矩特征的同时，计算不变矩在图像序列中的变化特征，以此获得描述步态序列的特征向量。本文的方法较之一些基于模型的步态识别方法简单、易操作，且实验效果较好。在此基础上，下一步研究工作的重点为提取多区域的其他变化特征，并与不变矩特征相结合，以便进一步提高人体步态的识别效率。

致谢 感谢中国科学院自动化研究所提供了 NLPR 数据库。

参 考 文 献

1 Wang Liang, Tan Tieniu, et al. Automatic gait recognition based on statistical shape analysis. IEEE Transactions on Image Processing, 2003, 12(9): 1120～1129
2 Ben Abdelkader C, Culter R, et al. EigenGait: motion-based recognition of people using image self-similarity. In: Proc. 3rd Int Conf. Audio- and Video-based Biometric Person Authentication, 2001. 284～294
3 Collins R, Gross R, et al. Silhouette-based human identification from body shape and gait. In: Proc. Int Conf. Automatic Face and Gesture Recognition, Washington, DC, 2002. 366～371
4 王亮，胡卫明，等. 基于步态的身份识别. 计算机学报，2003，26(3)：353～359
5 Ekinci M. A new attempt to silhouette-based gait Recognition for human identification. In: Canadian AI 2006, LNAI 4013, 2006. 443～454
6 Kale A, Rajagopalan A N, et al. Gait-based recognition of humans using continuous HMMs. In: Proc. The Fifth IEEE Int Conf. on Automatic Face and Gesture Recognition, 2002
7 Hayfron-Acquah J B, Nixon M S, et al. Automatic gait recognition by symmetry analysis. Pattern Recognition Letters, 2003, 24: 2175～2183
8 Lu Jiwen, Zhang Erhu, et al. Gait recognition using independent component analysis. In: ISNN 2005. LNCS 3497, 2005. 183～188
9 Yoo Jang-Hee, Nixon M S, et al. Extracting human gait signatures by body segment properties. In: Proc. Fifth IEEE Southwest Symposium on Image Analysis and Interpretation, 2002
10 Hu M K. Visual pattern recognition by moment invariants. IRE Transactions on Inform Theory, 1962, 8: 179～187
11 Vezien J M, Tarel J P. A generic approach for planar patches stereo reconstruction. In: Proc. The Scandinavian Conf on Image Analysis, Uppsala, 1995. 1061～1077
12 NLPR 步态数据库. http://www. sinobiometrics. com
13 Wang Liang, Tan Tieniu, et al. Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(12): 1505～1517
14 Philips P J, Moon H, et al. The feret evaluation methodology for face recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(10): 1090～1100