双码三笔汉字输入法的编码技术

严 喻 华泽玺
(西南交通大学 成都 610031)

摘 要 针对现有输入法的易学、快速和音形切换问题,提出了一种新的汉字编码——双码三笔汉字输入法编码。其设计思想是音形结合,字音采用汉字的读音首字母,字形采用汉字常用的五种笔画(横竖撇点折),在码元设计和键元映射时降低重码率,并采用频度索引来减少选字翻页频率,最终让会写不会读、会写又会读和会读不会写的用户无需切换输入方式就能轻松、快速地实现输入。

关键词 输入法,编码,音形码,双码三笔

Coding Technology of the ShuangMaSanBi Chinese Input Method

YAN Yu HUA Ze-xi
(Southwest Jiaotong University, Chengdu 610031, China)

Abstract To resolve the problem, that is the balance between easy to learn and fast to input and the switch between input methods, a new coding technology of the Chinese character was proposed. That is the code of the ShuangMaSanBi input method. The design idea is as follows. Sound and shape of the Chinese character have been combined together. The sound adopts the pronunciation of its first letters. The shape adopts five universal strokes of Chinese characters(HengShuPieDianZhe). The coding re-rate is reduced in the coding design and its key mapping. The flip frequency of the selected Chinese character is reduced by the frequency index. Finally, the user can input easily and quickly.

Keywords Input method, Coding, Yinxing coding, ShuangMaSanBi

1 引言

伴随着信息时代的来临和计算机的普及,中文输入法的发展可以用"百花齐放"和"万马奔腾"来形容[1],在 20 多年间,共出现过成百上千种编码方法。目前,汉字输入法编码技术根据编码规则的不同大体可分为:汉字字音编码(音码)、汉字字形编码(形码)、音形混合码(混码)和汉字数字编码(数码)。双码三笔汉字输入法编码是一种新的音形结合的汉字编码。

2 双码三笔汉字输入法编码的设计

2.1 原理

汉字作为一种图形文字,是由汉字的音、形、义来共同表达的。在汉字输入过程中必不可少的是编码过程,其编码原理基本上都是采用音、形、义与特定键位相联系,再根据不同汉字进行整合编制而完成的。双码三笔汉字输入法编码也是如此,其输入流程如图 1 所示,其中"内码"即"内部逻辑码"[2]是汉字在计算机系统内部的存在形式,用户按照某种编码规则输入外码即"汉字输入码",通过查找编码字典,经过软件处理后,完成从汉字外码到内码的转换。

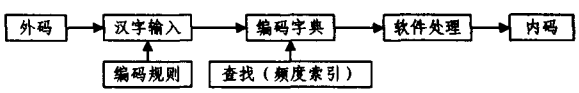


图 1 双码三笔汉字输入法的输入流程

2.2 编码的设计

双码三笔输入法编码设计的标准:①编码方案易于辨认;②编码规则尽可能少;③码元只用字母键;④码长最好为 3;⑤码元键盘分布合理均匀;⑥科学选择和安排简码字;⑦词组数量大,且具备实时造词功能。

该编码采用音形结合的模式,其编码设计如下:

(1)字音和字形的选取

以王力德统计和分析的汉字属性评估表[3](如表 1 所列)为参考标准,其中必修系数为某汉字属性对"中小学语文课所必修"的隶属度:必修为 1、非必修为 0;规范系数为某汉字属性对规范性的隶属度:公认无歧义为 1、基本公认模糊性较大为 0.5、无公认模糊性严重为 0;简单系数为某汉字属性对简单性的隶属度:简单为 1、一定难度为 0.5、难学为 0;记忆系数为某汉字属性对记忆性的隶属度:好记为 1、难记为 0、介于之间为 0.5。

综合分析表 1 的数据可知,声母和笔顺与其他汉字属性相比其规范性和易学性是最好的,由此确定读音为其字音,笔

到稿日期:2008-12-16 返修日期:2009-06-22 本文受西南交通大学学校科学技术研究基金项目(2007A09)资助。

严 喻(1981—),女,硕士生,主要研究方向为计算机软件等,E-mail:yanyu.swjtu@yahoo.com.cn;华泽玺(1968—),男,副教授,主要研究方向为检测技术、传感器及其应用等。

顺为其字形。

表1 汉字属性评估表

汉字属性		必修系数	规范系数	简单系数	记忆系数	合计
字形	部首	1	0.5	0.5	0	2
	字根	1	0	0	0	1
	角形	0	0.5	0.5	1	2
	字型结构	1	0.5	0.5	1	3
	笔顺	1	1	1	1	4
	笔画数	1	1	0	0	2
字音	声母	1	1	1	1	4
	韵母	1	1	0.5	1	3.5
	声调	1	1	0.5	1	3.5

(2)字音和字形的确定

字音是以《汉语拼音方案》为基础确定为汉字的拼音或其首字母。其中汉字拼音为26个英文字母的组合，首字母为拼音字符集={a、b、c、d、e、f、g、h、j、k、l、m、n、o、p、q、r、s、t、w、x、y、z}中的一个字符；字形是以GF3002《GB13000.1字符集汉字笔顺规范》为基础确定横、竖、撇、点、折5种基本笔画，其对应的笔形如表2所列。

表2 笔画-笔形对照表

笔画	笔形
横	
竖	
撇	
点	
折	

(3)码元与键元的映射

音元与键元的映射以GF3006-2001《汉语拼音方案的通用键盘表示规范》为基础，考虑到双码三笔输入法编码设计的理想标准将其确定为一一对应映射关系，即音元字母与键元字母相同；形元与键元的映射以音托和位托为基础，结合键位当量[4]属性，考虑到双码三笔输入法编码设计的理想标准及其重码率问题，将其映射关系确定为{横-H、竖-U、撇-P、点-D、折-V}。

2.3 编码的方案

双码三笔输入法编码方案包括两种编码方式，并且每种方式最多只用3个码元来实现，即任何一个汉字最多只需3个编码就能实现其输入，为了叙述方便，以下将编码方案的两种编码方式简称为编码1和编码2。其编码方案及其实施案例如表3和表4所列。

表3 字编码方案与案例

方案	案例
编码1=①+②+④ 编码2=②+③+④ 其中：①=汉字拼音首字母；②=汉字首笔划；③=汉字第二个笔划；④=汉字末笔划。	编码1：严=y+横+撇=Y+H+P=YHP 编码2：严=横+竖+撇=H+U+P=HUP

表4 词编码方案与案例

方案		案例
两字词组方案	编码1=①+②+③ 编码2=①'+②'+③ 其中：①=第一个字拼音首字母；①'=第一个字首笔划；②=第二个字首笔划；②'=第二个字次笔划；③=第二个字末笔划。	编码1： 社会=s+点+点=S+D+D=SDD 编码2： 社会=点+折+点=D+V+D=DVD
三字词组及以上方案	编码1=①+②+③ 编码2=①'+②'+③ 其中：①=第一个字读音首字母；①'=第一个字首笔划；②=第二个字次笔划；③=最后一个字末笔划。	编码1： 大中国=d+竖+横=D+U+H=DUH 社会主义=s+撇+点=S+P+D=SPD 编码2： 大中国=横+竖+横=H+U+H=HUH 社会主义=点+撇+点=D+P+D=DPD

3 双码三笔汉字输入法编码的关键技术

3.1 降低重码率和翻页率

该输入法形码设计中以音托为原则，键元映射关系应为{横-H、竖-S、撇-P、点-D、折-Z}，经过对GB字集的统计分析，S和Z其重码和键选率均较高，因此笔者以位托为原则，结合键位当量属性，将其竖、折键元确定为音码设计没有涉及的U、V字母，从而降低重码率。在此基础上，笔者还采用了频度索引来减少重码翻页次数，其原理为：在编码字典设计时，采用双码三笔输入的频度算法，对常用和简易字实行频度计数和智能处理，并生成一个频度索引库，让用户在输入汉字时先查找索引库再到编码字库查找。

对编码设计与翻页率进行统计分析，在输入给定测试样本过程中，给定样本中常用字不得低于50%，样本空间为4000个汉字(即有8000个汉字编码)，翻i页编码种数为n_i，汉字编码种数为N，则索引前的翻页率计算公式是

$$p_i=\frac{n_i}{N}\times 100\%$$

其中，$i=0,1,2,3,4,5$。频度算法由概率统计分析得出，索引后翻页率计算公式是

$$Q_i=\sum_{i}^{5}p_i(1-A_i)^iA_i$$

其中，P_i值由上式计算；

$i=0,1,2,3,4,5$；

$A_0=1, A_1=0.5, A_2=0.3, A_3=0.25, A_4=0.2, A_5=0.1$。

最终得出的统计数据如表5所列。

表5 翻页统计数据

	翻0页率	翻1页率	翻2页率	翻3页率	翻4页率	翻5页率及以上
索引Ⅰ	62.18%	20.56%	5.13%	2.96%	1.78%	7.39%
前Ⅱ	64.06%	21.89%	5.69%	2.14%	1.42%	4.8%
索引Ⅲ	78.02%	13.19%	3.66%	1.4%	0.9%	2.83%
后Ⅳ	86.53%	8.02%	2.34%	0.68%	0.37%	1.86%

表5中，Ⅰ表示形码设计为{h,s,p,d,z}的编码方案，Ⅱ表示形码设计为{h,u,p,d,v}的编码方案，Ⅲ表示输入汉字平均键选15次后的Ⅱ编码方案，Ⅳ表示输入汉字平均键选30次后的Ⅱ编码方案；Ⅲ、Ⅳ均是双码三笔输入法的编码，仅是使用索引的频度不同而已。因此，由表5可知，双码三笔输入法通过编码设计和频度算法设计，可以降低重码率和翻页率，提高汉字检出速度，若能进一步开发，其重码率和翻页率还可降低。

3.2 实施效果

(1)由于拼音和字形合理地融合在一起，编码方式间相互独立，系统会自动匹配其编码方式，因此用户无需做任何切换

就能实现其输入。(2)字元提取规范,码元设计简易,其易学易用性好。(3)编码规整简短,键位设计科学合理,分布均匀,其操作简便输入快捷。

结束语 综上所述,双码三笔汉字输入法编码技术易学易用、快速规范,很好地解决了"易学的打不快,快速的太难学"问题。它比五笔字型输入法易学,比拼音输入法快捷,同时也很好地将音码输入和形码输入融合,让会写不会读、会写又会读和会读不会写的用户无需切换输入方式就能轻松、快速地实现输入,很好地满足了用户需求。

参 考 文 献

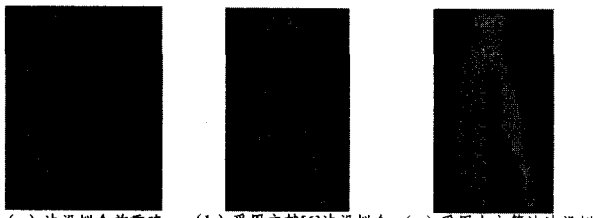
[1] 段曙东. 中文输入法软件产品质量监督抽查结果[N]. 中国质量报,2007-10

[2] 张寿萱,徐建毅,张建生. 中文信息的计算机处理[M]. 北京:宇航出版社,1984

[3] 王力德. 汉字编码的普及目标体系与编码实例[J]. 中文信息学报,1993,8(4)

[4] 陈一凡,张鹿. 键位相关速度当量的研究[J]. 中文信息学报,1990(9)

(上接第 291 页)



(a) 边沿拟合前重建　(b) 采用文献[6]边沿拟合和插值处理后重建效果　(c) 采用本文算法边沿拟合和插值处理后重建效果

图 4　超声图像三维重建效果

表 1　算法的对比

	逼近度	光滑度	采样控制点数	迭代次数
本文	242.4	21.4	26	$1.8*10^3$
文献[6]	182.8	32.8	40	$1.8*10^3$

(曲线原有点数=391)

实验 2　针对有些图像,如图 2(c)所作的处理,希望尽可能地逼近原始图像,这时设置效用权重向逼近度目标倾斜,$U(f(x))=0.5*E+0.3*SM+0.2*f_c$,输出的结果如图 3 的(c)、(d)所示。如表 2 所列,本文的算法在 3 个方面都比文献[6]算法好。

表 2　算法的对比

	逼近度	光滑度	采样控制点数	迭代次数
本文	268.7	30.4	72	$3.8*10^3$
文献[6]	308.6	38.1	73	$3.8*10^3$

(曲线原有点数=571)

结束语 本文针对分段拟合数字曲线的 NP 难问题,提出了基于粒子群的搜索算法来确定 B 样条曲线插值所需的控制参数。算法中采用了多目标优化的优化策略,从而使得所拟合的曲线既能较好地逼近原始轮廓线又能够达到平滑的效果。同时通过建立辅助存储空间来保存粒子群的历史非劣最优解,并计算最优解之间的距离来保持解的多样性以防止粒子群算法的过早收敛;采用分治与递归的思想调整了粒子的内部结构,提高了计算 B 样条曲线节点参数的灵活度,实现了多分辨率插值。最后还引入效用函数来建立对非劣最优解集中元素的评估机制,使得算法可以灵活应对各类图像的处理需求。实验证明该算法能够在较快实现曲线拟合的同时将目标区域边沿噪声去除,并对后续的三维重建工作起到十分积极的作用。

参 考 文 献

[1] Teh C H, Chin R T. On the detection of dominant points on digital curves[J]. IEEE Pattern Anal. Mach. Intell, 1989(11): 859-872

[2] 茹少峰,周明全,耿国华. 基于遗传算法的多边形逼近 3D 数字曲线[J]. 计算机辅助设计与图形学学报,2004,16(4):503-507

[3] Pei S-C, Horng J-H. Fitting digital curves using circular arcs [J]. Pattern Recognition, 1995(28): 107-116

[4] Horng J H, Li J T. A dynamic programming approach for fitting digital planer curves with line segments and circular arcs[J]. Pattern Recognition Lett, 2001(22): 183-197

[5] Sarkar B, Singh L K, Sarkar D. Approximation of digital curves with line segments and circular arcs using genetic algorithms [J]. Pattern Recognition Lett, 2003(24): 2585-2595

[6] 周明华,汪国昭. 基于遗传算法的 B 样条曲线和 Bézier 曲线的最小二乘拟合[J]. 计算机研究与发展,2005,42(1):133-143

[7] Pal S, et al. Cubic Bézier approximation of a digitized curve[J]. Pattern Recognition, 2007, doi: 10. 1016/j. patcog. 2007. 01. 019

[8] Kennedy J, Eberhart R C. Particle Swarm Optimization[C]// Proc. IEEE Int. Conf. Neural Networks. vol. 4, Dec. 1995: 1942-1948

[9] Coello C A. A Comprehensive survey of evolutionary - based multi-objective optimization [J]. Techniques Knowledge and Information Systems, 1999, 1(3): 269-308

[10] Knowles J D, Corne D W. Approximating the nondominated front using the Pareto archived evolution strategy[J]. Evolutionary Computation, 2000, 8: 149-172

[11] 黄贤英,张丽芳. 基于粒子群优化的模糊聚类算法[J]. 重庆工学院学报:自然科学版,2008,22(11):120-123