

水族文字笔形编码方法研究

杨撼岳, 陈笑蓉, 郑高山

(贵州大学计算机科学与信息学院, 贵阳 550025)

摘 要: 针对水族文字不易编码的问题, 提出一种基于笔形特征的编码方法。分析《水书常用字典》收录的467个水族文字的字形特征, 总结出7种基本笔画。通过统计比较, 将7种基本笔画按几何形状细分重组为21类笔形。依据编码规则, 取水族文字3个角的笔形组成有序序列, 实现水族文字的编码。

关键词: 中文信息处理; 水族文字; 笔形; 编码; 输入法

Study on Coding Method of Shui Nationality Characters According to Stroke Shape

YANG Han-yue, CHEN Xiao-rong, ZHEN Gao-shan

(College of Computer Science & Information, Guizhou University, Guiyang 550025, China)

【Abstract】 In order to work out the problems that the method coding Shui nationality characters is difficult, a coding method based on stroke shape for Shui nationality characters is proposed. The shape of 467 Shui nationality words of the Common Shui Script Dictionary is analyzed, and seven kinds of basic strokes consists mainly of Shui characters. Through the statistical comparison, 21 kinds of stroke shape are subdivided. Shui nationality word is code by an orderly sequence composed by three strokes on its corners according to the coding rules.

【Key words】 Chinese information processing; Shui nationality character; stroke shape; coding; input method

DOI: 10.3969/j.issn.1000-3428.2011.14.097

1 概述

水族是中华民族大家庭中的一员, 其90%以上的人口聚居在贵州省, 少量散居于广西壮族自治区、云南省等地^[1]。水族有着自己的语言和文字, 水族文字被水族人民称为“水书”, 其字形特征与甲骨文、金文相像, 主要用于记录天文历法、宗教信仰和民间知识等。目前研究发现的水字有500多个单字(不含异体字, 包括异体字共有1700多个字)^[2]。

已有的水族文字键盘输入法分为2类: (1)形码输入法: 文献[3]提出水族古文字计算机输入法, 选取133个基本字根, 并分布于25个字母键位, 最大码长为2; 文献[4-5]设计了水书文字编码方案, 采用不等长编码, 最大码长为4, 由6种基本笔画与14类部件组合成400多个水书文字。(2)水-汉字映射编码输入法: 韦双周、韦双波开发的中国水族文字输入系统, 采用水族文字对应简体汉字的五笔编码和拼音码。例如: 输入“树枝”汉字的五笔编码“SCSF”或拼音“SHUZH”后, 在应用程序中显示为𐊉𐊐字。

上述输入法存在2个关键问题: (1)形码输入法没有对图形文字编码, 如𐊉、𐊐、𐊑等。(2)水-汉字映射编码输入法虽然能做到一字一编码, 但在不知道某水字的意译时给输入带来了不便。因此, 本文分析《水书常用字典》收录的467个水字的特征信息, 提出一种基于笔形的输入编码方案。

2 水族文字的特征信息和熵

水族文字(下文简称水字)所负载的特征信息可分为3类: 字形类特征信息, 字音类特征信息, 字义类特征信息^[6]。本文仅讨论水字字形类特征信息。

2.1 水字字形类特征信息

由于水族文字没有形成统一、标准、规范的字体, 本文以《水书常用字典》收录的水字为蓝本, 对水字字形进行分

析和讨论。

2.1.1 水字笔画特征信息

下笔后按一定方向连续画成的每一笔, 就叫做笔画^[7]。本文分析了《水书常用字典》收录的水字, 去粗取精, 得到水字7种基本笔画: 横, 竖, 撇, 捺, 点(实心圆点), 折(两笔相接成角), 曲(一笔连续写成的、复杂的曲线), 具体如表1所示。

表1 水字的7种基本笔画

名称	横	竖	撇	捺	点	折	曲
基本形式	—		/	\	·	┐	〰

水字笔画之间的组合关系大致可分离散、连接、交叉、包围结构, 如表2所示。

表2 水字笔画间的组合关系

笔画间组合	示例
离散	𐊉 𐊐 𐊑 𐊒
连接	𐊓 𐊔 𐊕
交叉	𐊖 𐊗 𐊘 𐊙
包围	𐊚 𐊛 𐊜 𐊝

2.1.2 水字字根特征信息

字根(部件)是组配水字的构字单元, 通过笔画相互间“单、散、连、交”式的关系组成。字根可以根据设计需要,

基金项目: 贵州省科学技术基金资助项目“水族文字键盘输入系统研究”(黔科合J字[2010]2093号)

作者简介: 杨撼岳(1975—), 男, 硕士研究生, 主研方向: 中文信息处理; 陈笑蓉, 教授; 郑高山, 硕士研究生

收稿日期: 2011-01-10 E-mail: yhy_75@yeah.net

按照一定标准, 决定选择个数。水字字根集确定后, 在其中挑选有限个字根, 可以拼成某一个水字(如图1所示)。

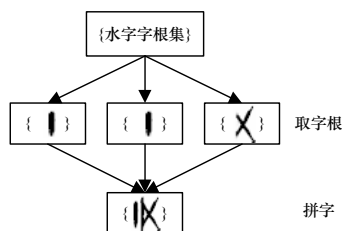


图1 字根拼字的过程

文献[3]采用五笔输入法的部分主要字根作为基础, 选取133个字根用于构造308个水字。文献[4]按照特征相似的选择标准, 同时也考虑组字频率, 将400多个水字拆分为60多个部件(字根)。

2.2 水字的熵

一个信源(本文提到的信源均指离散信源)由输出符号集和控制输出的规则组成。熵是信源的重要属性, 是描述每个信源符号所传送的平均信息量的度量^[8]。其数学定义为:

$$H(A) = -\sum_{i=1}^m p_i \lg(p_i) \quad (1)$$

其中, A 为信源符号集; m 是 A 集中符号的数量; p_i 为第 i 个符号的概率, 而且 $\sum_{i=1}^m p_i = 1$ 。若 $p_1 = p_2 = \dots = p_m = 1/m$ 时, 信源的熵值最大, 称为最大熵。记为:

$$H_{\max} = \lg m \quad (2)$$

水字的最大熵为 $H_{\max} = \lg 467 \approx 8.87$ 。事实上, 在水书中水字不是等概出现, 因此,

$$H(\text{水字集}) = -\sum_{i=1}^{467} p_i \lg(p_i) \leq 8.87$$

即每个水字载荷的信息量不会超过 8.87 bit(前提是水字符集的大小为 467)。

3 水字笔形输入法的设计要素

3.1 特征信息集的构造

本文在大量实验和统计的基础上, 将7种基本笔画按几何图形特征, 细分或重组成21种笔画(见表3)。“曲笔”被细分为圆形、上下凸形、左右凸形、波形等; “折”笔被细分为上下角、左和右角。在笔画归类时, 倾斜的凸形容易引起歧义, 于是规定倾角小于 45° 归于“上下凸”, 大于 45° 归于“左右凸”。通过实验发现, 如果把“左角”和“右角”归为一类, 重码字数提高了19%, 因此, 将它们细分为2类笔形; 而“上下凸”、“左右凸”和“上下角”按上下左右方向细分后, 重码字数减少了2%~4%, 对重码的影响不大, 不再对它们按方向细分。

水字一个最明显的特点是离散结构多, 如等。为了压缩相似的特征信息, 可以把离散结构视为一个笔形类。本文把3笔以上笔形聚集或包围某笔形的离散结构定义为“大聚集”笔形。对所有的重码字分析发现, 有10%左右的重码字含有3笔的离散结构。于是将笔画数等于3笔的离散结构定义为“小离聚”笔形。增加“小离聚”笔形后, 重码字数减少了9%。

水字的另一个主要特点是“复笔”数量大, 而且形态多。水字的“复笔”笔形态除去常见的几何图形(如三角型、四边形等)外, 还有部分不常见的几何图形(如等)。如果把这些笔形一一细分, 势必增加特征集的大小, 对输入法的易学性、易用性产生不利影响。因此, 把这部分不常见的、

出现频率不高的几何曲线定义为“异”形。

构成输入编码的笔画序列方式主要有2种: 笔顺方式, 取角方式^[6]。本编码方案采用“取角”方式, 取一个水字3个角的笔形, 组成一个有序序列表示该水字。水字“取角”最主要的优点是可以避开图画文字中间复杂的图形; 不利是重码略微增加。

3.2 键盘上特征信息元的分布

本输入法采用“形托”为主、“音托”为辅的方式, 将13类笔形分布在其形似的字母键位, 5类笔形对应于其发音相似的字母键位, 3类笔形安排在易于敲击的键位, 键位分布如表3所示。

表3 键位分布

键位归并方法	键元	笔形名	笔形	键位归并方法	键元	笔形名	笔形
形托法	A	三角		形托法	Y	丫	
	C	左右凸			Z	右角	
	F	插			D	点	
	I	竖		音托法	E	异	
	L	左角			H	横	
	O	圆			N	捺	
	T	箭头			P	撇	
	U	上下凸		易于敲击原则	B	四边形	
	V	上下角			K	大离聚	
	W	波			S	小离聚	
	X	叉					

3.3 码长、键元集和编码空间

方块字的特征信息元所依托键位的集合是编码键元集, 输入一个方块字所需的击键次数称之为码长。以键元数(a)为底, 取码长(l)的乘方(a^l)称之为编码空间, 它的值代表了在限定的键元集和码长下, 可容纳输入编码的数量^[6]。

限定键元数为21, 字符集的数量是467时, 本文令码长为2、3和4, 则它们的编码空间分别是441(21^2)、9261(21^3)和194481(21^4)。码长为2时的编码空间(441)小于字符集的数量(467), 意味该空间不能容纳整个水字符集。码长为3、4的编码空间均大于467, 因此, 理论上最大码长取3或4都是可行的。

本文假设编码是随机的(没有编码规则)。在总数为 n 的编码空间中, 为 m 个水字编码($m < n$)。显然, 每个水字出现的概率相等, 即 $p = 1/m$ 。第1个水字编码时的重码概率为0; 当编第2个水字时, 产生重码概率是 $1/n$; 第3个, 第4个, ..., 第 m 个水字编码的重码概率分别是 $2/n, 3/n, \dots, m/n$ 。 m 个水字编码完毕后, 产生的字符集重码概率为:

$$P'_s = P \frac{1}{n} [1 + 2 + L + (m-1)] = \frac{m-1}{2n} \quad (3)$$

对于非随机性有确定编码规则的输入系统, 字符集重码概率为:

$$P_s = R \frac{m-1}{2n} \quad (4)$$

其中, R 由编码规则决定, 称为人工干预因子^[6]。

当码长等于3和4时, 它们的字符集重码概率分别是:

$$P_{s3} \approx 0.025R, \quad P_{s4} \approx 0.001R$$

从低重码的角度来考虑, 码长取4最理想。但是重码只是输入编码方案设计的一个侧面, 输入效率也是要考虑的重要因素。码长取3时, 字符集重码概率为 $0.02R$, 其重码的

量级也是相当低的, 因此, 本方案设计最大编码码长为3。

4 水族文字笔形输入编码规则

编码键元集: 共用了21个字母键。编码码长: 不等长码, 最长编码为3码。笔形和键元的对应关系如表4所示。

表4 输入法键位分布表

键元	笔形名	笔形	例字	特征
A	三角			封闭包围的三角形
B	四边			封闭包围的四边形
C	左右凸			向左或右凸的曲线
D	点			实心圆点
E	异			非其他20种笔形的曲线或封闭包围结构
F	插			一笔穿过2笔以上或封闭包围结构
H	横			水平线
I	竖			垂直线
K	大离聚			3笔以上笔画聚集或包围某笔形的离散结构
L	左角			2笔相接成角, 其角指向左
N	捺			左高右低的斜线
O	圆			封闭包围的圆形
P	撇			右高左低的斜线
S	小离聚			三笔画聚集或包围某笔形的离散结构
T	箭头			形似箭头图像
U	上下凸			向上或下凸的曲线
V	上下角			2笔相接成角, 其角指向上或下(竖角除外)
W	波			形似波的图像
X	叉			2笔相交之形
Y	丫			形似英文字母Y之形
Z	右角			2笔相接成角, 其角指向右

编码输入方法: 对水字进行编码时, 根据表中的笔形, 按照左上角、右上角、左下角、右下角的取角顺序取3码。

(1)若4个角均有笔形, 则取前3个角的笔形。例如: 𐊑(箭头、箭头、竖——TTI); 𐊒(上角、上角、叉——VXX)。

(2)若4个角中有一个笔形占2个角, 则这2个角的笔形只取其一。例如: 𐊕(异形、下角、下角——EVV); 𐊖(叉、点、竖——XDI)。

(3)若4个角被2个笔形占尽, 除取这2个笔形外, 第3个角取剩下字形左上角的笔形。例如: 𐊗(上凸、下凸、离散形——UUK); 𐊘(右角、横、左角——ZHL)。

(4)若4个角被一个笔形覆盖, 除取该笔形外, 内部字形仍按取角顺序取其前两角的笔形。例如: 𐊙(四方、插、插——BFF); 𐊚(圆、小离聚、横——OSH)。

(5)按实有笔形取码, 最少1码, 最多3码。例如: 𐊛(异形——G); 𐊜(三角、三角——AA)。

取码注意点如下:

(1)含插、叉笔形的字, 如该字不足3码, 用插入的一笔笔形和和被插入的笔形补充(叉形中, 自上而下的单向笔画是插入的一笔), 不得超过最大码长3码。例如: 𐊖(叉、撇、下凸——XPU); 𐊗(插、捺、下凸——FNU)。

(2)含小离聚或大聚集笔形的字, 如该字不足3码, 按取角顺序取构成离聚的笔画补充, 不得超过最大码长3码。若

构成离聚的笔画全部相同, 则只须补其中一笔。例如: 𐊗(上凸、大离聚、捺——UKN); 𐊚(圆、小离聚、撇——OSP); 𐊛(小离聚、波——SW)。

(3)三角形或四边形的某笔端点位于三角形或四边形之外的字, 按照实际书写笔画取笔形, 不取三角形或方形。例如: 𐊘(右角、圆、圆——ZOO)。

(4)独立或平行的笔形, 不论高低, 一概取最左或最右的笔形。例如: 𐊙(横、右角、左角——HZL)。

(5)复笔优先取。例如: 𐊚左上角应取复笔右角(Z), 而不取横(H)。

(6)有2复笔形可取时, 上角较高笔形优先取, 下角较低笔形优先取。例如: 𐊛右上角取复笔叉(X), 而不取右角(Z)。

5 输入编码方法的评测

输入编码方法的评测指标如下^[7]:

(1)水字字符集 S 的大小 $|S|$: 指的是字符集中水字的数量。

(2)键元集 K 的大小 $|K|$: 用于输入特征信息的键元的数量。

(3)键元熵值为 $H(k) = -\sum_{i=1}^k p_{ki} \lg p_{ki}$, 其中, k 为键元数量; p_{ki} 第 i 个键元的使用频率。

(4)平均码长 $\bar{L}$: 平均输入每个水字所用的击键次数(选字键应包括在码长之内)。

(5)重码数 R : 重码数 R = 重码字数 - 重码组数。

通过编制的程序计算得到: 水字字符集的大小 $|S| = 467$; 键元集的大小 $|K| = 21$; 键元熵值 $H(k) = 4.193\ 503\ 698\ 783$; 平均码长 $\bar{L} = 3.88$; 重码字数为100, 重码组数为45, 重码数 $R = 100 - 45 = 55$ 。

6 结束语

本文对水族文字笔形编码方法进行研究, 提出一种基于笔形的输入编码方案。设计的笔画种类远比水字字根的数量少, 因此记忆量小, 而且避开了定义字根的困难, 使输入法易学、易用。尽管本文所编码的对象是水族文字, 但其编码思想也可以用于其他象形文字的输入法设计。

致谢 衷心感谢贵州省三都水族自治县水族文化研究所潘兴文、潘政波和韦世方提供了水字字形资料。

参考文献

- [1] 潘朝霖, 韦宗林. 中国水族文化研究[M]. 贵阳: 贵州人民出版社, 2004.
- [2] 韦世方. 水书常用字典[M]. 贵阳: 贵州民族出版社, 2007.
- [3] 韦宗林. 水族古文字计算机输入法[J]. 贵州民族学院学报, 2000, (4): 13-19.
- [4] 董芳, 周石匀, 王崇刚. “水书”文字编码方法研究[J]. 黔南民族师范学院学报, 2006, 26(6): 1-4.
- [5] 戴丹. 水书文字编码输入系统设计与实现[D]. 贵阳: 贵州大学, 2005.
- [6] 陈一凡, 胡宣华. 汉字键盘输入技术与理论基础[M]. 北京: 清华大学出版社, 1994.
- [7] 郭平欣, 张淞芝. 汉字信息处理技术[M]. 北京: 国防工业出版社, 1985.
- [8] Wells R B. Applied Coding and Information Theory for Engineers[M]. [S. l.]: Pearson Education, Inc., 1998.

编辑 顾姣健