数字键汉字编码技术的研究和应用

王永民

(中国王码集团 北京 100089)

摘 要 按照国家标准(GB/T18031—2000)的《通用要求》,对数字键汉字输入的键位设计和编码设计进行了理论探讨,并以数字王码为例,提出了键位、码元和取码规则的匹配策略,介绍了基础、初级、中级和高级等 4 套方案实例,以期作为我国数字键汉字输入技术研究应用及其实现标准化的参考.

关键词 汉字编码;数字王码;王码;汉字输入;国家标准

中图法分类号 TP317

Research and Application in Chinese Input Technology for Numerical Keyboard

WANG Yong-Min

(*China WangMa Group*, *Beijing* 100089)

Abstract According to the "General requirements for Chinese input on numerical keyboard" in the national standard (GB/T18031—2000), the paper theoretically investigates the keyboard and code design regarding the numerical method of Chinese input. Taking the case of numerical WangMa, the paper provides matching strategy among numerical keyboard, Chinese character components and encoding rules, and also presents four solutions: Basic, enhanced, intermediate and advanced. Hopefully the study makes contribution on research, application and standardization of the numerical technology of Chinese input.

Keywords Chinese coding; digital WangMa; WangMa; Chinese input; national standard

1 引 言

自 20 世纪 80 年代开始,数字产品在我国大量出现.汉字的"数字化输入",已成为举世公认的难题.

当前,国内的数字产品大都采用不符合我国汉字规范、技术落后的"进口"汉字输入法,不但使汉字文化受到污染,每年我国还要向"外商"交付巨额的"专利费".尽管国内的汉字输入方案有上千种,但未形成一种公众趋势和标准,因此,亟待我国的科学工作者从国家标准出发,尽快实现符合我国语言文字规范,科技体系完备、普及型的输入法.

2 国家标准及技术参数

2.1 数字键汉字输入技术的国家标准

国家质量技术监督局于 2000 年 3 月 17 日发布了一项信息技术领域代号为 GB/T18031—2000 的国家标准,名称为《数字键盘汉字输入通用要求》(以下简称《通用要求》),现摘引其中涉及"形码"的内容如下:

4.1、使用键位在 0~9 数字键范围内;

4.3、编码规范:

数字编码涉及的汉字笔画、笔顺应遵从《现代汉语通用字笔顺规范》

收稿日期:2008-01-08.**王永民**,1943 年生,教授级高级工程师,30 年来从事字母键汉字输入技术与数字键汉字输入的理论研究和应用工作. E-mail: wym@wangma.com.cn.

4.4、基本笔画：

汉字的基本笔画分为 5 种，其键位（数字代码）如下表：

数字键	1	2	3	4	5
笔画名称	横	竖	撇	捺	折
基本笔画	一丿	丨	丿	丶乀	乙

（注：《通用要求》规定：提笔丿归于横；竖左钩亅归于竖，乀归于捺）

4.5、易学性

学会使用汉字数字编码输入的时间应尽量短.

4.6、汉字输入平均码长（击键次数）

- 单字输入平均码长应小于 6；
- 字词混合输入平均码长小于 4；

4.7、- 单字笔画码输入重码率小于 8%；

- 字词混合输入重码率小于 10%.

2.2　汉字的基本笔画

根据许慎在《说文解字》中“独体为文，合体为字”的观点，汉字可分为独体字与合体字两大类别. 统计表明，在 10000 个常用汉字中，独体字只占 5% 左右，而由独体字“组合”而成的“合体字”，则多达 95%. 这里，我们把作为“零部件”组字时“组字频度”高的独体字，如口、人、日、廿、氵、纟等，叫做“字根”或“部件”[1].

汉字由字根组成. 字根则是由“笔画”组成. 作者对于“笔画”的定义如下：

书写汉字时，一次写成的连续不断的笔迹.

由此可导出 3 个与《通用要求》完全吻合的推论：

（1）凡是一笔写成的，无论什么方向，无论如何弯折，都是“一个笔画”. 如“马”的前两笔，“飞”的第一笔，“凸”的右上角一笔，都不能切断成几个笔画；

（2）凡是抬了笔，经两次或两次以上写成的一个笔画结构，如十、八、勹、口等，都不是“笔画”（可称为部件或字根）；

（3）笔画的类别，只和运笔方向和书写次数有关，而与其长短大小无关. 如丶和乀只能是一种笔画（right falling），丿和丿也是同一种笔画（left falling）.

就类别和数量而论，汉字结构中的笔画、字根、整字和词汇，作者认为可以和物质的构成类比如表 1 所示.

表 1　汉字结构与物质构成的类比

	汉字构成	数量特征	物质构成
第一层	基本笔画	几种	基本粒子
第二层	字根	上百种	原子
第三层	汉字	成千上万种	分子
第四层	词语	无数种	物质

一般来说，在标准键盘的 26 个字母键上设计汉字输入技术，只有以“字根”为单位才切实可行. 比如五笔字型，采用了 125 种字根，分布在 25 个键位上，平均一个键位有 5 种，只要分组布局合理，取码规则得当，设计出一个兼顾相容性（重码少）、规律性（易学习）、协调性（指法顺手）的“拼形组字键盘”，是完全可能的[2].

然而，对于 0～9 共 10 个数字键来说，要把 125 种精减到已近极限的字根，安排在数字键位上“拼形组字”，每个键上平均要安置 10～15 个字根，不要说记忆难度很大，就算是能够科学分组，编码输入时引起的重码，也必将是多得惊人！

为此，正如字母键盘上设计输入技术必须力避“汉字之多”、抓住“字根之少”一样，在数字键上设计汉字输入技术，则必须力避“字根之多”而抓住“笔画之少”.

这也就是在数字键上设计汉字输入技术，必须以汉字的“细节”信息——笔画作为“编码元素”（码元）的原因.

作者对国标 GB2312—80 字集的 6763 个汉字的 5 种笔画作过统计，如表 2 所示.

表 2　汉字笔画统计

笔画种类		出现次数	约占比例/%
横	一丿	21870	30
竖	丨	13728	20
撇	丿	11495	16
捺	丶乀	12013	16
折	乙	12720	18

信息处理的一个重要技术指标是编码效率. 即便是有了笔画分类，却规定把汉字拆分成一个个单笔画，全部按笔顺输入也是不可取的. 正如“T9”输入法，虽然其 5 种笔画的分类也符合《通用要求》，可是该输入法等于在数字键上“写字”，“码长”不定（等同于笔画数），显然不可能有高效率.

可见，要解决汉字的数字键输入，必须首先解决 3 个问题：

（1）汉字的基本笔画有几种（码元）；

（2）编码输入汉字，用几个数字键（键位）；

（3）一个汉字要打几下键（码长）.

这 3 个基础性的问题，在国家标准 GB/T18031—2000 中都已经做了明确规定. 国家标准的《通用要求》，是我国数字键汉字输入的“一定之规”，对此设计者不能“不予理会”而闭门造车. 否则即使有些“新颖性”，也不可能具备实用性.

2.3 数字键汉字编码的技术参数

编码的技术参数,可以用来定量地评定方案是否符合《通用要求》.其中最重要的几个参数如下.

① 码元 M

指被设置在数字键上,用来参加编码的汉字元素——笔画、字根或部件,以一个键位上设置 1～3 个为宜.

② 键位数 K

指设置码元的键位数.

③ 编码长度 L(码长)

指为单字和词语编码时的编码位数(用整数表示),也即输入一个汉字时所需的最多按键次数,其最大码长用 $L_{\max}$ 表示;

④ 汉字字集 H

指有待编码的汉字集合的汉字总数(国标一级字 3755 个,一、二级字 6763 个,GBK 字集 21003 个,国标 18030 字集 27533 个等).

⑤ 编码空间 Ω

指某一编码方案的最大编码容量.编码容量与待处理的汉字数相比,必须有较大的冗余度,否则其重码的概率就会很高.为此,我们事先限定:$\Omega \geq H$;例如:$K=6$ 时,$L=5$ 和 $L=6$ 的编码空间分别是

$$\Omega = \sum_{i=1}^{5} 6^i = 9330;\ \Omega = \sum_{i=1}^{6} 6^i = 55986.$$

再如 $K=9$,也即用 9 个数字键输入,码长 $L=6$ 时,编码空间便大得多:

$$\Omega = \sum_{i=1}^{6} 9^i = 597870.$$

这个编码空间 Ω,比字母键上五笔字型的编码空间 $\Omega = \sum_{i=1}^{4} 25^i = 406900$ 还要大.这就是说,对于同一字集,在数字键上要想达到"五笔字型"那样少的重码,就应当按 $K=9$,$L=6$ 设计方案.

⑥ 重码率与重码字数

重码率是输入编码唯一性的定量指标,重码率高,输入时必然增加翻屏选字的次数(键选率上升).

重码字数:是在汉字集合 H 中建立编码体系之后,编码完全相同的汉字总数 $H_{重}$,可用作者推导出的以下公式计算出来:

$$H_{重} \approx \frac{H^2}{\Omega}.$$

公式说明:汉字编码在编码空间 Ω 中是随机分布的.字集 H 在 Ω 中的平均密度为 $\eta_1 = \frac{H}{\Omega}$,$\eta_1$ 是任将一个编码"投掷"到 Ω 中,与 Ω 中已知的 H 个编码发生"碰撞"的事件概率,即"重码率";另外,若已知在字集 H 中的重码字数为 $H_{重}$,则实际的重码率为

$$\eta_2 = \frac{H_{重}}{H};$$

理论上应当有 $\eta_1 = \eta_2$.但由于 η_2 是实测数据,不会有理想值 η_1 来得"均匀",所以 $\eta_1 \approx \eta_2$,即

$$\frac{H_{重}}{H} \approx \frac{H}{\Omega},\ \text{故}\ H_{重} \approx \frac{H^2}{\Omega}.$$

这是一个根据键位 K、码长 L 和字集 H,便可以计算"平均重码字数"的近似公式,代入前面的 Ω 值,于是我们有

$$H_{重} = \frac{H^2}{\Omega} = \frac{H^2}{\sum_{i=1}^{L} K^i},\quad L = \text{码长}.$$

例如,在字集 $H=6763$ 个字时,键位 $K=25$、码长 $L=4$,其编码空间和重码字数如下:

$$\Omega = \sum_{i=1}^{4} 25^i = 406900,$$

$$H_{重} = \frac{H^2}{\Omega} = \frac{6763^2}{\sum_{i=1}^{4} 25^i} = 112(\text{字}).$$

这便是五笔字型的"25 键 4 码"输入法可以达到的极值.这个极值是当所有的字根"组字能力"都完全相同,每个键位上的字根数也平均分配的情况下才能达到的.然而,由于 100 多个字根的组字能力差异很大,如"口"的组字能力为 7%,而"宀"的组字能力只有 1%,"车"的组字能力只有 0.36%,以致于实际的编码方案,有效编码在整个编码空间 Ω 中的分布并不均匀.所以,五笔字型的实际重码字数,达不到 112 字,而是 260 字.对于杂乱无羁的汉字来说,五笔字型的编码分布已经相当接近理想值了.

需要说明的是,这个公式对于数字键汉字编码的设计也同样适用.

⑦ 取码难度 D

指按某种方法和顺序提取汉字的编码元素时,根据其复杂程度,每一个"操作"所付出的劳动量.例如"建"字按单笔画编码时,取第一笔画、第二笔画的"劳动量"是不相同的.一般来说,即使对于会写汉字的人,越往后面的笔画越不好认定,越容易出错,这说明"取码难度"是随着笔画的"深入"而递增的.这里,我们设定依笔顺取码时的取码难度依次递增:

顺序	取第 1 笔画	取第 2 笔画	取第 3 笔画	…	取第 n 个笔画
取码难度	1	2	3	…	n

这样,"建"字"逐笔取码"的"取码难度"便是

$$D=\sum_{i=1}^{8}D_i=36.$$

⑧ 难度系数

取码的“难度系数”，直接关系编码法的易学性.

当依照习惯的规范笔顺取码时，取码难度可设为按 1 递增，增量为 1；但是，当逆向取码时，例如要取倒数第 1 笔、倒数第 2 笔，…，显然其难度比顺向取码要大得多.“逆向取码”因其操作违背了既有的书写习惯，应当增加一个难度系数 d，这里设定 $d=2^n$，n 为倒数的笔画数：倒数第 1 笔，$d=2^1$；倒数第 2 笔，$d=2^2$；…；倒数第 n 笔，$d=2^n$.

同一个键位上的“码元”越多，累计的取码难度会很大. 可以通过试验证明，一个键位上若有 5 个码元，其取码过程中的实际难度，和在 5 个键上取码是等同的，只是手指打键的操作得到简化而已. 所以应当尽量减少每个键上的码元数.

⑨ 部件的实用频度 Ps

汉字的基本笔画和部件除了“组字频度”外，在设计输入法中显得更为重要的是实用频度. 例如“的”在现代汉语文中的出现频度高达 4%，所以，构成“的”的笔画和部件立即“身价百倍”. 数字编码的技术指标算法，常采用“实用频度”而只参考“组字频度”，目的是要考量方案在实际应用中的效率，而不特别关心能涵盖多少个字.

作者依据海量汉字文本，统计出来的部件的实用频度结果如表 3 所示(前 15 个)[3-4].

表 3　汉字实用频度统计结果

排序	部件	频度/%	排序	部件	频度/%	排序	部件	频度/%
1	口	7.30	6	亻	2.0	11	月	1.35
2	人	2.28	7	白	1.83	12	又	1.27
3	土	2.22	8	勹	1.73	13	之	1.26
4	日	2.14	9	木	1.68	14	八	1.14
5	火	2.1	10	氵	1.41	15	寸	1.12

由此，还不难计算出 5 种单笔画的实用频度. 方法是把每一个部件先拆成单笔画，再把部件的实用频度作为各个笔画的“权值”，累计便可. 这样得到的笔画的实用频度，对于数字键码元的优选和编码参数的计算都十分有用.

⑩ 取码公式

从一个汉字的结构中，按确定的取码规则取出几个笔画(部件)，以其代码构成输入编码，可以用一个取码公式表示：$W=C_m^n$，W 为输入码，m 为汉字的全笔画序列，n 为被取到的笔画.

⑪ 取码规则

数字键汉字输入技术，在键位、码元确定之后，制订“取码规则”便成为一个事关方案优劣的大问题.

取码规则的制订一般要遵从以下几点：

a) 符合汉字的结构特点，使具备汉字书写知识的人，容易操作，不易出错；

b) 取码顺序总体上要符合汉字的“从上到下、从左到右”的书写习惯，若要人们抛开早已谙熟的汉字笔顺习惯，另外掌握一个新的甚至是“倒行”的拆字取码习惯，其方案必定难以推广；

c) 要拆取汉字结构中信息量较大的部分(或笔画)，以便离散重码提高效率.

3　数字键汉字输入技术的“基础方案”

根据国家标准《通用要求》规定的键位、笔画分类及码长限定，仅仅用 5 种单笔画和 5 个数字键，便可以设计出以下“基础方案”：

(1) 码元：一 丨 丿 乀 乙

代码：1 2 3 4 5

(2) 键位分配

7	8	9
乀 4	乙 5	6
一 1	丨 2	丿 3

(3) 取码规则

① 笔顺全码：依笔顺取完为止；

② 取“前四末一”编码输入.

以上(1)，(2)两项符合国家标准. (3) 则事关码长. (3) 中的①，如 T9 输入法，码长不定，编码空间 $\Omega\to\infty$，难以计量(若一个有 20 个笔画的字，其编码空间 $\Omega=\sum_{i=1}^{20}5^{20}!$)，因而①只是在数字键上“写字”的方法，无码可编，不称其为科学设计的输入法.

(3) 中的②，“前四末一”取码法，是作者 1984 年秋为辅助学习五笔字型而设计实施的“简易输入法”，这项技术出现在作者获得的 ZL85100837.2 号专利的从属权项之中[5].

其取码规则“前四末一”规定：任何汉字，只取“前四个及最后一个”单笔画，如：建，⺕一一一乀(51114)；路，丨⺕一丨一(25121).

根据上一节技术参数的计算公式，容易算出：在处理国标一级汉字(3755 个)时，“基础方案”的技术参数为

$$K=5,\ L=5,$$

$$\Omega=\sum_{i=1}^{5}5^i=3905,$$

重码率：$\eta=\frac{3755}{3905}=96.2\%$，

重码字数：$H_{重}=\frac{H^2}{\Omega}\approx 3610$.

“基础方案”简单易学，但效率很低.编码空间太小，重码字太多，实用价值不高，至多可作为一个“入门”输入法，真正科学实用的数字键汉字输入技术，有待下面研究讨论.

4 数字编码中部件“口”的特殊作用

2006 年 12 月作者获得授权的 ZL03150281.4 号发明专利，就是一个创新的数字键输入方案.

在该发明的《说明书》中，对于汉字编码中的一个特殊部件“口”，作者作过系统的研究[5-6]：若在 5 个单笔画的数字键盘上，增加一个部件“口”，放置在“6”键上，便可以使原来用 5 个基本笔画的“基础方案”的各项技术指标大幅度提升，取得质的进步.

如果说多加一个“键”，使编码空间 $\Omega_1=\sum_{i=1}^{5}5^i$ 变为 $\Omega_2=\sum_{i=1}^{6}5^i$，使编码变得“稀疏”一些降低“重码率”，这样的方法容易想到的话，那么，对于选择“哪一个部件”作为新增的编码码元，便不是很轻易可以确定的了.因为这个被特别选中的“部件”，必须符合以下条件：

① 频度很高.在全部汉字部件中是出现频度最高的，只有这样，这个新增的键位和“码元”，才能被充分利用；

② 构字能力强.被选“码元”的部件，应当在汉字结构的“各个位置上”都经常出现.只有这样的部件，才能充分地发挥“离散编码”的作用.

为此，作者建立了字根和笔画的数据库，经数据整理和繁难的统计分析，终于得到了除 5 种单笔画之外，几百个部件的频度统计表.表 4 显示出了国标一、二级 6763 个汉字中最常见的 10 个部件在各个编码位置上出现的次数.

表 4 国标 6763 个一二级汉字中，“高频部件”出现的频次统计

序号	部件	第 1 码位置	第 2 码位置	第 3 码位置	第 4 码位置	累计次数
1	口	397	227	345	224	1193
2	艹	371	115	41	26	553
3	人	61	186	164	108	519
4	日	99	162	167	86	514
5	木	261	84	82	74	501
6	土	155	122	112	92	481
7	讠	366	31	1	0	398
8	亻	242	91	54	0	387
9	月	134	59	106	47	346
10	扌	280	31	8	0	319

从表 4 可见，在汉字部件中，出现频度最高的部件是“口”，而且在各个编码位置上都经常出现，是一个构字能力最强的部件.所以，若要增加一个码元，最佳决策是优选“口”作为新增一个键位的码元.其作用有 3 种：

① 扩大了编码空间.现有技术对 6763 个汉字编码时，其编码空间及其按照上述重码率的理论计算公式的重码率为：

$$\Omega_1=\sum_{i=1}^{5}5^i=3905,\ \eta_1=\frac{6763}{3905}=173.2\%.$$

而新增一个键位码元“口”之后：

$$\Omega_2=\sum_{i=1}^{5}6^i=9330,\ \eta_2=\frac{6763}{9330}=72.5\%.$$

应该说，仅增加一个码元和一个键位，对简单、易学程度的影响很小，可是理论上的重码率却只有现有技术的 42%，这个决策是正确的.

② 降低重码率.增加“口”后，对累计出现频度为 90%的最常用的 1000 个汉字进行编码试验（如表 5 所示），不重码的字为 474 个，而现有技术则是 377 个，增加码元“口”后的不重码字，是现有技术的 1.26 倍，净增加 26%.

表 5

项 目	重码级别	不重码字数	2 个字重码	3 个字重码	4 个字重码	5 个字重码	6 个字重码	7 个字重码	8 个字重码	9 个字重码
现有技术“基础方案”	重码组数	0	99	55	15	10	4	5	1	1
	涉及字数	377	198	165	60	50	24	35	8	9
带“口”的 6 键输入法	重码组数	0	108	43	10	7	5	1	1	
	涉及字数	474	216	129	40	35	30	7	8	

③降低取码难度. 选“口”作为单占一个键位的码元之后，“口”就不再拆成 3 个单笔画. 这对于降低“取码难度”有着明显的贡献. 例如，“口”若拆成单笔画，“中”的取码难度为 $d=1+2+3+4=10$；若“口”不拆，其取码难度则为 $d=1+2=3$.

数字化汉字输入技术的设计，就是确定键位、码长及码元、制订规则，实现“多目标统一协调”的过程. 键位太少，Ω 太小，重码率就高；而键位太多，Ω 增大，必然影响效率；码元太多，必然降低易学性. 用 5 个基本笔画作为 5 个键位上的码元，Ω 显然太小，重码太多.

而且，当输入技术在计算机标准键盘右端的数字小键盘上应用时，要用食指(对应 1、4)、中指(对应 2、5)、无名指(对应 3)打键，无名指只管一个键，显然对均衡手指负荷不利. 如果能增加一个键位，把数字键 6 也用上，既有效地扩大了 Ω 值，降低重码率，又可使三个手指各负荷 2 个键，打键均衡.

由上表可见，在数百个汉字部件中，部件“口”是一个符合上述要求的高频部件.“口”其所以特殊，是因为许慎在《说文解字·序》中云：“仓颉之初作书，盖依类象形，谓之文”[7]，“画成其物，随体诘诎”[8]，而人类的两大行为“吃”、“说”都离不开“口”，在汉字造字之初，凡是与“吃、说”相关的字大都含有“口”. 在不失简单易学的情况下，最有资格被提升为“常委”的码元，当然是“口”.

5　数字键汉字输入技术的“初级方案”及其“增强型方案”

将 5 种基本笔画和“口”分别设置在 1、2、3、4、5、6 共 6 个键上，便构成了一个数字键输入的“初级方案”：

7	8	9
丶 4	乙 5	口 6
一 1	丨 2	丿 3

$K=6$，$L_{\max}=5$，

$\Omega=\sum_{i=1}^{5}6^i=9330$，

取码公式：$W=C_m^{4+1}$(取前 4 个和最后 1 个)，

重码率：$\eta=\dfrac{H}{\Omega}=40.24\%$，

重码字数：$H_{重}=\dfrac{H^2}{\Omega}=1511$(个).

仅仅多设 1 个键位，这个“初级方案”的各项指标就远远优于“基础方案”.

如果在 6 键上并列一个码元“日”，这个“初级方案”便成为一个增强型.

因为汉字编码的离散性，初级方案的实际重码率比理想值要高得多. 而且数字键“6”的平均负荷，也仍然比其它几个键低得多. 为此，作者将处在部件排行榜第 4 位的“日”(频度 2.14%)加以提升，并列设计在“6”键上，与“口”同一个键，不过赋予它一个双码“66”，就像人的姓氏中有“诸葛、司马”一样. 从字形上看，“日”形似两个“口”，“口”是 6，“日”自然可以是 66，形象易记. 这样(按“前四末一”)，“旦”的编码是 661，而不再是 25111 与“目”等重码了.

将“日”与“口”并列设计在“6”键上，享用“双码”66，在码元及编码设计中，是一个应用效果很好的创新案例：

① 避免了把“日”拆成单笔画(2511)，让“日”从“目、具、旦”(25111)等汉字的编码“辖区”中分离出来，明显地减少了重码和键选率；

② 以 66 为代码，就等于将“日”的 2.14%的频度，变为 $2\times2.14\%=4.28\%$，和“口”的频度迭加在同一个键“6”上，使键位负荷均衡.

这样就形成了如下所示的“增强型初级方案”，其实测的各项技术指标更接近理想值(这里不再列出).

7	8	9
丶 4	乙 5	口日 6 66
一 1	丨 2	丿 1

6　数字键汉字输入技术的“中级方案”

数字键汉字输入技术的研究是一个不断创新的过程. 作者设计的“首部-余部”编码法及其键盘(专利号：ZL00100002.0)，便是一个与“初级方案”键位相同、码元相同，却因为取码规则的创新而使得编码效果更接近理论值、实用性更强的“中级方案”.

这项设计给出了一种全新的“首部-余部”汉字编码输入技术方案，有效地克服了“初级方案”重码较多、效率不高等问题. 它揭示了合体汉字从结构上均可“一分为二”，并分为“首部”和“余部”两个部分，便于分别取码的结构规律.

定义.　对于占总数 95%以上可以“一分为二”的合体汉字，其包含第一笔的部分叫“首部”，剩余的

部分叫“余部”.

从编码学的角度来看,任何一个或一组笔画,其拓朴图形“周边”的笔画,也即外露的笔画,都比结构内部的笔画有更大的熵值,因而更便于被辨认,被识别,具有较佳的离散能力.该项发明由此首创了“首部-余部”取码法,即先取“首部”的第一笔和末笔,再取“余部”的前几笔和最末笔合在一起,作为输入编码,形成全新的编码体系.

该发明提出的“首部”和传统的“部首”,例如《新华字典》所用的 201 个部首,不是一个概念.在汉字结构中,“首部”虽然常常也是第一笔写成的“部首”,但正如“奶牛”不同于“牛奶”一样,“首部”不同于“部首”,“部首”不全是“首部”.例如“想”的包含第一笔的“首部”是“相”,“部首”却是“心”,“都”的“首部”是“者”,“部首”却是“阝”等.

使用此种全新的取码法,以“初级方案”的键位和键元为基础,便成为一个“中级方案”.

7	8	9
丶 4	乙 5	口 日 6 66
一 1	丨 2	丿 3

虽然“中级方案”的码元、键位与“初级方案”完全相同,但由于取码法的创新,加上码长从 5 增加为 6,便产生了质的进步:

$$K=6,\ L_{\max}=6,$$

$$\Omega=\sum_{i=1}^{6}6^i=55986,$$

重码率:
$$\eta=\frac{H}{\Omega}=12.08\%\ (H=6763),$$

重码字:
$$H_{重}=\frac{H^2}{\Omega}=817(字)(理想值).$$

取码公式:
$$W=C_{m_1}^{n_1+n_2}+C_{m_2}^{n_3+n_4},$$

m_1:首部笔画,$n_1=1$,$n_2=1$,

m_2:余部笔画,$n_3=3$,$n_4=1$.

难度系数(设“余部”的“起始难度”为 2):

$$D=(1+2)+(2+3+4+2)=14.$$

这样,“中级方案”便成为各项指标完全符合国家标准的实用化方案了.

注 1. 对于占汉字总数 5%左右,不便于“一分为二”的独体汉字,我们采用“初级方案”的取码规则,依书写顺序取“前四末一”编码,取码公式为:$W=C_m^{4+1}$,“成”取 13554,含“日”的“重”取 31661 等.

注 2. 重码率≠键选率.“键选率”是指输入汉字的过程中,从“重码序列”中选择汉字的频率.由于编码的不均匀性往往使得方案的实测参数离理论值很远.但是仍可以通过软件设计加以改善而大大降低从提示行键选汉字的次数.当然,这就与提示行中汉字的多少和排列顺序密切相关.比如,当提示行的重码字在软件的支持下按“实用频度”排列使得“高频先见”时,其实际的“键选率”就比平均重码率要低得多.虽然上述“中级方案”的重码率(在 6763 字集内)是 12.08%,但实际的键选率要远远低于这个值.

7 数字键汉字输入技术的“高级方案”

前述之基础方案、初级方案和中级方案,都是为了便于普及,以简单易学为指导思想的,所以分别只采用了 5～6 个键位和 5～6 个码元.

国家标准《通用要求》规定,数字键汉字输入技术可以用 1～0 共 10 个键位.除 0 键因常作为“结束标志”实际上不宜安置码元外,如果采用 9 个键,在同样码长 $L=6$ 的情况下,其编码空间、重码率等技术参数将会有更大的进步,兹以国标一、二级汉字为例:

$$K=9,\ L_{\max}=6,\ H=6763,$$

$$\Omega=\sum_{i=1}^{6}9^i=597880,$$

重码率:
$$\eta=\frac{H}{\Omega}=\frac{6763}{\Omega}\approx1.13\%,$$

重码字:
$$H_{重}=\frac{H^2}{\Omega}\approx76.5(字).$$

这个理想情况下的参数指标,甚至比“25 键 4 码”的五笔字型还要好得多.这是因为:虽然键位从五笔字型的 25 个字母键减少为 9 个数字键,可“码长”却从五笔字型的 4 码变成了 6 码,编码空间 Ω 从 40 万扩大为 59 万,重码的理论值当然也就更好.

然而,仅仅有这样一个数学期望还是构不成实用技术的,要成为“高级方案”,关键还在于如何优选码元、合理地设置在 9 个键上并制订既能有效离散重码,又简明易学的取码规则.

优选码元:在数字编码设计中,码元的选取一般遵循以下原则:

① 以 5 种基本笔画为基础,这样才能做到既符合国家标准又保证体系一致,使之与前述的基础、初级、中级方案“向下兼容”,让初学者可以由基础方案入门,逻辑地递进到高级方案.

② 笔画在数字键位上的设置,“一丨丿㇏乙”对应

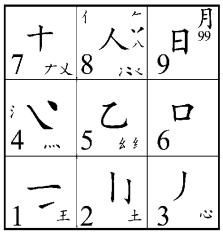
12345，是个不可变动的硬性规定，随意错位或让它们对应其它数字键，将不符合国家标准的《通用要求》.

③ 除 5 种单笔画外，还可以选取少量构字能力强、结构简单、实用频度高的部件或笔画结构，如“口日十人氵纟月…”作为码元，当然是宜少不宜多；超过 20 个将难记难用；

④ 少量被优选采用的部件，在键位上的设置，要有规律可循或者使之同类相聚，易记易用；

⑤ 取码规则要简单、易学，一听就明白，伸手便能用，操作灵巧方便，最好能够与前述之“兄弟”（中级）方案“向下兼容”.

据此，我们研究出一个“高级方案”（专利号：ZL00103505.3），这个方案的键位码元图如下图所示（图中可见，虽然 7、8 两个键位上的码元多一些，但因笔画简单且形似类聚，所以易记易用）：



这个“高级方案”，$K=9$，码元 9 组（大小共 20 个），各项理论数据如本节上文所示.

关于数字键“9”上的码元：“日”的代码从“中级方案”的 6 6，改变为单一编码“9”；“日月同辉”，“月”与“日”同键，被赋予双码“99”. 这是为了使常用部件“月”不再拆分，同时增加 9 键的负荷. 为便于记忆，“高级方案”有一首《键元歌》：

1 王 2 土 3 撇心
4 点 5 折 6 对口
7 键十叉人八 8
日出 9 键月 9 9

其实，码元选定并安置到键位上之后，最重要的是制订“取码规则”. 一套好的编码规则，往往使实际编码的结果，各项指标更加接近理想值. 本案的“取码规则”可以有以下几种供选择：

① 顺序取全码：依照书写顺序，以码元为单位，把汉字“拆解”取完，按键输入，这当然无疑于在键盘上“写字”，码长不定，无码好编. 编码空间 $\Omega=\sum_{i=1}^{n}K^i$ 随笔画数 n 而按级数陡增，冗余度可想而知，取码难度也将累计增大，此法自不可取.

② 按“前四末一”取码：因为“高级方案”的码元与五笔字型的字根相比，都是笔画数很少的结构，要“小”得多，常见的偏旁如“马、车、鱼、⻏、舟、…”等，势必还要被拆成好几个小“码元”，“浪费”了有限个“码位”而降低离散能力，重码较多，因而也不可取！

③ 按“首部-余部”取码：正如前述之“中级方案”的取码法——“首部头尾取 2 码，余部前 3 加最后”，既简单易学，又有最好的离散能力，同时，还能与上述之“中级方案”的取码法一脉相承保持一致，其取码公式为：

取码公式：$W=C_{m_1}^{1+1}+C_{m_2}^{3+1}$，

取码难度：$D=(1+2)+(2+3+4+2)=14$.

“高级方案”也被称做“王码 9 键”输入法，编码的统计分析结果，其重码率十分接近五笔字型，加上词语输入，输入日常文章时，其键选率低于 2%，其技术指标远远优于国家标准的《通用要求》（键选率≤10%），当然比各种“进口”输入法要好得多。

8　数字键上的五笔字型——“数字五笔”

25 年来，五笔字型为很多人所熟悉. 当大量的手机类数字化产品出现时，大家自然会希望在数字键上打五笔字型. 这一愿望是可以实现的.

五笔字型的 25 键-5 区 25 个键位的“字根键位图”如图 1 所示[8].

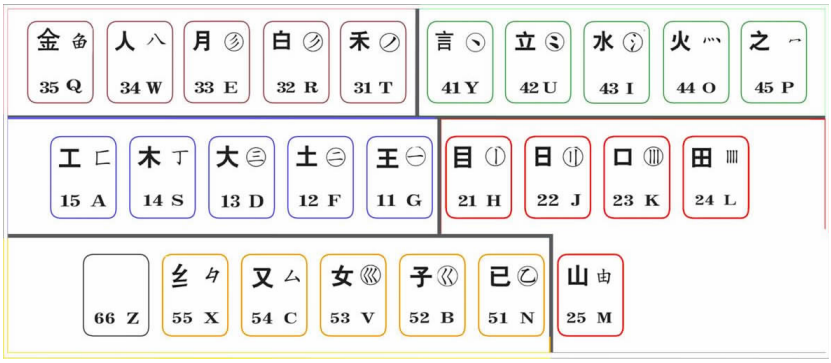


图 1　字根键位图

依照这个设计，五笔字型问世之初，已经就有 3 套并行的输入方式：即字根分解方式、区位编码方式和字母编码方式. 例如：

现 { 王 口 儿 / 11 25 35 / G M Q ；编 { 纟 、 尸 艹 / 55 41 51 15 / X Y N A

这样，在数字键上，我们可以有两种方式实现“五笔字型”输入：

① 按区位码输入，即在数字键上打字根的“区位码”：

汉→43 54 41；字→45 52 12

② 按字母编码输入，就是直接在数字键上输入五笔字型的“字母编码”，如：

信→WYG；息→THNU；公司→WCNG

这个方法叫做“数字五笔”输入法. 因为不再是一个字母一个键了，重码自然要比五笔字型多许多，但比“拼音输入”的重码要少很多.

总之，对于已经熟用五笔字型的人士，用“数字五笔”在数字键（手机）上打字，不用学习，立可上手. 但对于不会五笔字型的人士，当然还是用“数字王码”更为简便可行.

9 数字键汉字编码技术体系结构

下面以“数字王码”为例，给出数字键汉字编码输入技术“递进式”体系框图.

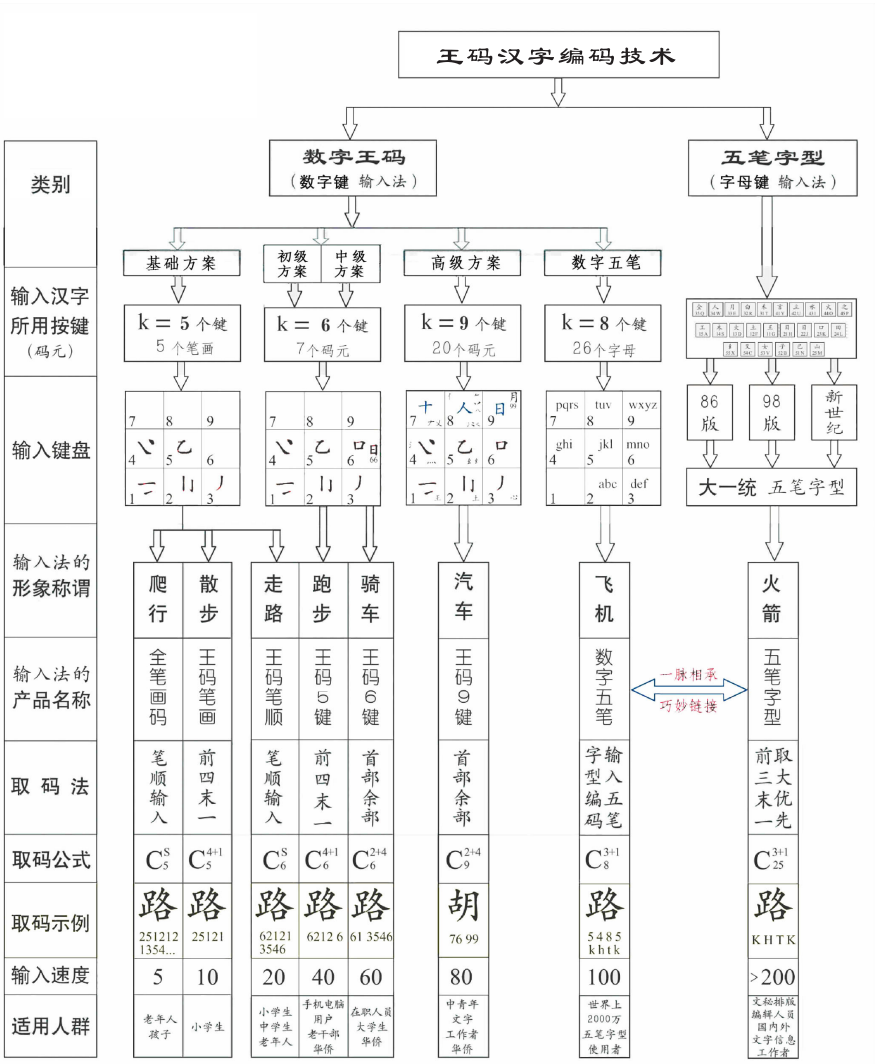


图 2 数字王码输入法体系结构示意图

10 结 语

近年来我国手机类数字产品大量涌现，在海外的数字式汉字输入法“乘虚而入”的情况下，实现我国汉字输入技术的规范化和国产化，是一个亟待解决的问题. 本文按照国家标准（GB/T18031—2000）《数字键盘汉字输入通用要求》，对数字键汉字输入的键位设计和编码设计进行了理论探讨，并以“数字王码”为例，提出了键位、码元和取码规则的匹配策略，详细

介绍了“数字王码”的 4 套方案实例。该“数字王码”方案已于 2006 年 11 月 14 日通过了国家信息产业部的电子产品国家标准技术检测和认证，完全符合国家标准 GB/T18031—2000 和 GB18030—2000 的要求，其认证证书号为 No. GES101106P10122ROS. 作为本项理论研究在实践中的应用，本文在最后概括性地给出了数字键汉字编码技术的体系结构，以期作为我国数字键汉字输入技术研究应用及实现标准化的参考.

参 考 文 献

[1] Wang Yong-Min. WuBiZiXing Chinese Input Computer Technology. Zhengzhou: Henan Science and Technology Publishing House, 1985: 34-35(in Chinese)
(王永民. 五笔字型计算机汉字输入技术. 郑州：河南科学技术出版社，1985：34-35

[2] Wang Yong-Min. Three principals in computer keyboard design for Chinese input. Chinese Journal of Computers, 2005, 28(5): 879-880(in Chinese)
(王永民. 计算机汉字键盘设计“三原理”. 计算机学报，2005，28(5)：879-880)

[3] Wang Yong-Min. Theory and practice in coding Chinese. Chinese Information International Conference Papers(2), 1983: 29-31(in Chinese)
(王永民. 汉字编码的理论与实践. 中文信息国际研讨会论文集(2)，1983：29-31)

[4] He Jiu-Ying et al. Chinese Characters and Culture. Beijing: Peking University Press, 1995: 79(in Chinese)
(何九盈等. 中国汉字文化大观. 北京：北京大学出版社，1995：79)

[5] Wang Yong-Min. China Patent No: ZL85100837.2, Description(in Chinese)
(王永民. 中国专利号：ZL85100837.2，说明书)

[6] Wang Yong-Min. China Patent No: ZL03150281.4, Description(in Chinese)
(王永民. 中国专利号：ZL03150281.4，说明书)

[7] Rao Zong-Yi. Signal, Character and Letter—Tree of Chinese Characters. Shanghai: Shanghai Shudian Press, 2000: 23(in Chinese)
(饶宗颐. 符号·初文和字母——汉字树. 上海：上海书店出版社，2000：23)

[8] He Jiu-Ying. Culture of Chinese Characters. Shenyang: Liaoning People Publishing House, 2000: 185(in Chinese)
(何九盈. 汉字文化学. 沈阳：辽宁人民出版社，2000：185)

[9] Wang Yong-Min. User Manual for WuBiZiXing. Beijing: China Science and Technology Publishing House, 1993: 4(in Chinese)
(王永民. 五笔字型用户手册. 北京：中国科学技术出版社，1993：4)

[10] Lucas W F. Discrete and Systematic Model. Changsha: National University of Defense Technology Publishing House, 1996(in Chinese)
(Lucas W F. 离散与系统模型. 长沙：国防科技大学出版社，1996)

[11] Kapur J N. Mathematical Modeling. John Wiley & Sons, 1988

[12] Zhong Chu-Qian et al. Foundations of Applied Statistics. Guanzhou: South China University of Technology Publishing, 1992(in Chinese)
(庄楚强等. 应用数理统计基础. 广州：华南理工大学出版社，1992)

[13] Yuan Yin-Shang. Probability and Statistics. Beijing: China Renmin University Press, 1985(in Chinese)
(袁荫棠. 概率论与数理统计. 北京：中国人民大学出版社，1985)

[14] Sakai, Nagao, Terai. A description of Chinese characters using sub-patterns. Information Processing, 1969, 10(5): 285-293(in Japanese)

[15] Myers W. Key developments in computer technology: A survey. Computer, 1976: 48-78

[16] Berry J S. Teaching and Applying Mathematical Modeling. John Wiley & Sons, 1984

WANG Yong-Min, born in 1943, professor. Based on the research from 1978 to 1983 he invented WuBiZiXing—A patented Chinese input technology using keyboard. From 1998 to 2003, he was engage in the theoretical research on inputting Chinese by numerical keys, invented the Numerical WangMa which includes five progressive technologies of Chinese input by numerical keys, and developed series of software applied in Mobile phones and computer.

Background

As there are big differences between Chinese characters and culture and English ones, inputting Chinese into computer has been a big challenge for the last 30 years. In 1983 the author invented WuBiZiXing, and achieved the technological breakthrough of inputting Chinese characters efficiently by using a standard keyboard. As digital products like mobile phone getting rapidly popularized, inputting Chinese characters by the numerical keys in an easy and efficiently way has become a new challenge. Some input methods from abroad, not in line with requirements of national standards came into the market. Thence the author spent six years on fundamental and theoretical research, setting up a mathematical model, and completing the development of the progressive input system and Numerical WangMa software protected by six patents. This technology passed the assessment of national standard, and contributes to the standardization and improvement of domestic technology for inputting Chinese character by numerical keys.