

第一章 汉字输入技术概述

§ 1.1 计算机与汉字输入

汉字是世界上最古老的文字之一,是中国人民宝贵的文化遗产;汉字是世界上使用人口最多的文字,也是联合国的工作语言之一。

自本世纪六十年代以来,电子计算机以其神奇的功能给人类社会带来了空前的社会文明。目前,在一些发达国家,计算机已广泛应用于政治、军事、经济、文化、科技等各个领域。无论是在科学计算等数值计算方面,还是在办公自动化等非数值信息处理方面,都发挥着越来越大的作用。在我国,电子计算机的应用从六十年代就开始了,直到进入八十年代以后,才得到广泛的应用。但是,要使电子计算机技术在我国真正的普及和推广,进入普通的办公室和家庭,就必须解决电子计算机的中文信息处理问题,并在此基础上,提高计算机对汉字中文信息的处理能力。计算机中文信息处理的关键是汉字信息的输入问题。汉字进入计算机是中文信息处理的基础,没有这个基础,其他就都谈不上。

为了使古老的汉字能够迅速、高效地适应信息技术的发展,使汉字能进入计算机并可参与信息处理,使我国的计算机应用走向普及,七十年代以来,具有聪明才智的炎黄子孙,对计算机的汉字输入技术进行了深入的研究与开发。其输入技术大致可分为如下三方面:汉字编码输入技术、汉字识别输入技术、汉字语音输入技术等。仅就汉字编码输入技术而言,到八十年代中期,据有关资料

统计,已有汉字输入编码方案和设想五百多种。这就为汉字适应电子计算机的环境,进行计算机处理创立了有效的手段。

§ 1.2 汉字编码输入技术的发展

随着计算机技术的普及和推广,经过数年的探索和实践,有关人士已逐渐认识到,要想实现真正的办公自动化,仅靠配备专业操作人员将永远无法真正提高办公效率,也无法实现中文信息处理的现代化,而必须让人们放下手中的笔和纸,将汉字直接输入计算机,使汉字输入工作基本上一次性完成。这也是汉字编码输入技术的目的和发展方向。

1.2.1 汉字编码方案的发展

汉字编码方案的发展,从总体上来说,是从无规律到有规律,从难记到易记,从低效率到高效率的。汉字编码的分类如下:

1. 无理序号编码

纯无理序号编码在我国没有出现过。这种汉字编码最短,但记忆难度极大。此种方法若经长时间的训练,熟练掌握后,其输入速度从理论上说是最快的。

2. 分类序号编码

典型的分类序号编码包括:国标码、区位码、电报码。

3. 有理编码法编码

主要分四类:拼音类、音形结合类、笔画拼形类和字元拼形类。

包括:汉语拼音、拼音联想、首尾码、五笔字型、自然码、大众码等数百种。

1.2.2 汉字输入的时代划分

第一代:以单字输入为主。

早期发明的几百种输入法大多以此为基础。

第二代：以词组输入方式为主。

以字为基础，词为主导，音形结合，字词一体，高频先见智能处理的输入方式属于第二代。象自然码基本上属于第二代产品。目前，常用的《五笔字型》、《大众码》等汉字输入方法是以单字为主，但兼有部份词的功能，是介于第一代与第二代之间的产品。

随着计算机技术的发展，汉字编码输入效率已由过去的依赖编码转向依赖系统，依赖智能化的软件设计。同样的编码在不同的软件支持下将会有明显的差异。

第三代：以语句智能化语音输入为主，必须通过第五代计算机去完成。

§ 1.3 汉字输入方法选择要点

选用哪一种方法输入汉字，使用者可根据自己的实际情况择优选用。选择时应注意如下要点：

1. 编码规则要简单明确，易学易记，便于熟练掌握。
 2. 编码容量要大，至少应覆盖 GB2312—80 基本集中规定的全部汉字(6763 个)，并易于扩充。在必要情况下，应能覆盖辅助汉字集中所包括的汉字。
 3. 重码率要低，并有简便的区分重码的方法。
 4. 编码要有唯一性。同一个汉字只能产生一个确定的编码。
 5. 尽可能使用标准字符键盘作为汉字输入的设备，键位的分布合理。
 6. 汉字输入的平均击键次数要少，尽量缩短编码长度，以提高编码效率。
 7. 输入速度要高。
- 总之，要选择易学、易记、易用、输入速度高的输入方法。

§ 1.4 常见的几种汉字输入法之比较

目前,各种汉字输入法有数百种之多,但真正商品化的仅几十种;在计算机(特别是微型计算机)上应用面较广的几种汉字操作系统中,主要配备了:国标码、区位码、电报码、汉语拼音、拼音联想、首尾及快速、大众码、五笔字型、自然码等汉字输入方法。这些输入方法是本书主要介绍的内容。

为便于大家选择适合自己实际情况的一两种输入法,下面对这几种输入法进行简单的介绍和比较。

1.4.1 区位码

区位码输入法是按国标 GB1988 - 84 和国标 GB2311-84 中对各种汉字和符号的编码规定,对国标 GB2312-80 中 7445 个汉字和符号,按行、列进行划分,共划分为 94 个区(行)和 94 个位(列)。每个汉字或符号都对应于一个区号和位号。输入时,只要输入对应的区号和位号四位数字,即可得到对应的汉字或符号。

该方法的优点是:只要有一本区位码表,不需任何训练,即可输入汉字和各种符号。码长为 4 键/字。

主要缺点是:即使进行专门训练,也不可能完全记住区位码表,而查表则严重影响输入速度。

1.4.2 国标码

国标码是区位码的另一种表示方法,国标 GB2312-80 中每个汉字和符号用四个数字或字母(A、B、C、D、E、F)来表示,即用四位十六进制数来表示。输入时,对照国标码表,查出对应的编码,并依次输入,则相应的汉字或符号就输入计算机了。

例如:“摧”字的编码为 345D,依次按下数字键“345”和字母键

“D”，“摧”字就输入计算机了。

用国标码输入汉字的优缺点和区位码完全相同，只是不如区位码使用面广。

1.4.3 电报码

电报码是按电报明码编码法，用四位数字对应一个汉字。输入时，先查到汉字对应的编码，依次输入相应的数字即可。

例如：“多”字的电报码为 1122，依次按下数字键“1122”，汉字“多”就输入计算机了。

· 用电报码输入汉字的优缺点和区位码基本相同。

1.4.4 汉语拼音

· 汉语拼音方法是微型计算机上早期实现的一种汉字输入方案。现在，几乎所有的通用汉字操作系统和专用操作系统中均配有该方法。该方法是用输入汉字对应的汉语拼音字母来选择和输入汉字的。据统计，中文汉字约有 50000 个，国标码中规定的 1、2 级字库汉字有 6763 个，只有不到 500 种不同的拼音；显然，拼音相同的汉字较多，输入时只能从同音字中选字。

· 汉语拼音法的主要优点是：学习时间短，对于会汉语拼音的人来说，不需专门训练，即可向计算机中输入汉字。

· 主要缺点是：输入速度慢，不认识的汉字无法输入且汉字库中的 682 个符号无法输入。

1.4.5 拼音联想输入法

拼音联想输入法是基于拼音输入法又带有联想功能的一种汉字和符号的输入方法。当采用此法输入时，拼音联想系统自动调整汉字及联想字的位置，保证最后输入的汉字自动调换到同音字的最前面。从而大大减少选字的时间。

汉字联想输入法的主要优点是:速度比较快,学习掌握很容易,不需要专门训练,并可输入汉字库中的常用符号。对于输入汉字的同时又需要输入各种符号的操作尤为方便。

主要缺点:基本同于汉语拼音输入法。

1.4.6 首尾码及快速输入法

首尾码是用 52 个字根来概括组成汉字的基本笔画。52 个字根用键盘上的 26 个英文字母来对应。输入汉字时,先确定它的首字根和尾字根,依次按下对应的字母键,首尾字根相同的十个汉字就显示在屏幕的下方,用“>”键和“<”键向后或向前寻求。找到时,按相应的数字键,选中的汉字就输入计算机了。所谓快速输入法,实际上就是首尾码的快速输入法。

首尾码输入汉字的主要优点是:见字识码,会写就会输入。

主要缺点是:重码很多,输入速度慢。

1.4.7 五笔字型

《五笔字型》汉字输入法是目前国内影响较大、应用面很广的一种高速汉字输入法。该方法又分为三种类型:简易型、普用型、高效型。简易型名为五笔画输入法,它采用横、竖、撇、捺、折五种基本笔画来概括组成汉字的各种笔画。输入时,只要输入汉字的开始四笔和最后一笔即可。普用型名为五笔桥输入法,它采用几十个字根和单笔画来概括组成汉字的各种笔画。它的基本输入方法是输入汉字的首笔对应的字根和其他部分的四个单笔画。高效型名为五笔字型输入法,它使用 130 个字根来概括组成汉字的各种部件。基本输入方法是输入汉字的前三个字根和最后一个字根。三种方法均有词组输入功能。

五笔字型输入法的主要优点是:见字识码,会写就会输入;输入速度很快,经过短期训练的入很容易达到 100 字/分钟以上的速

度。

主要缺点是：不会写的字和汉字库中的非汉字字符不能输入。

1.4.8 《大众码》汉字输入法

《大众码》汉字输入法也是目前国内应用十分广泛的一种快速汉字输入方法。它采用 128 个基本字元和 44 个联想字元来概括组成汉字的各种部件，共选定了 174 个字元。基本输入方法是输入汉字的前三个字元和最后一个字元。输入一个汉字的平均击键次数（包括空格键）为 3.42 键/字。

《大众码》输入法的主要优、缺点和五笔字型基本相同。

1.4.9 《自然码》汉字输入法

《自然码》汉字输入法是一种高度灵活的开放式汉字编码输入法，目前在國內的应用日趋广泛。自然码全码共五位，分为表音（声部及韵部）、表意（义部）、和表形（形部和尾部）三个独立的部分。在使用中可选用其中不同的 2 至 4 位搭配组合成字或词的输入码。由此演变出很多种输入方法。其特征为：音形结合，音里有形，形里有音，以音代形，以形表音，形音共存。

《自然码》的优点：简单易学、易掌握；输入速度快且汉字和非汉字图形符号均可输入；使用时，疲劳程度低。

没有明显的缺点。

1.4.10 综述

显然，任何一个使用计算机的人都没有必要掌握所有的汉字输入方法。一般来说，只要掌握其中一两种方法就足够了。但所有在计算机上进行中文处理的人员（使用《自然码》汉字输入法的人员除外），都应该掌握区位码、国标码、电报码这三种汉字输入方法的其中一种。特别是区位码汉字输入法。以便进行汉字输入时，万

·遇见找不到的个别汉字或用到非汉字字符,可用查表的方法找到。

在计算机(主要是微型机)上进行汉字输入,关键是微型机必须配备一个汉字操作系统(包括一个汉字库)。汉字操作系统有各种型号和版本。汉字库是把通常用的汉字聚集在一起,以备进行汉字的输入输出处理。目前,在国内使用的计算机汉字库一般都是按国标 2313—80 中规定的 7445 个字符组成。汉字库在微型机内部建立的方法有两种:软汉卡和硬汉卡。软汉卡是把汉字库放在硬盘或软盘上,启动汉字操作系统时调入微机内部。如:IBM-PC/XT, IBM-PC/AT 等微型机。硬汉卡是在生产微型机时(硬汉卡也可以单独加装),将汉字库直接装在微型机内部,如:长城 0520、长城 286 等微型机。不论哪种方法都必须和汉字操作系统联合使用。其中软汉卡是随着操作系统的启动,自动调入微型机内部的。

关于在计算机(主要是微型计算机)上选择汉字输入方法的操作(如何安装各种操作系统及汉字输入方法,这里就不赘述了,请参阅随盘带的安装说明),可大致分为两种:

1. 用“ALT+功能键”选择,如:IBM=PC/XT 等机型。
2. 用键盘上的方式键选择,如:长城系列微型机等。

第二章 五笔字型汉字输入法

§ 2.1 五笔字型概述

2.1.1

为了使用五笔字型输入法,在计算机的键面上,刻上五笔字型字根。这样,这个键盘就成了一个“字根键盘”,如图 2.1 所示。

键面上有“金木水火土,日月山石田,人口手”等字根。用户可以用这些字根象小孩子搭积木一样,拼合成许多汉字。如机器已经置在五笔字型的状态,用“日”和“月”左右拼起来就成为“明”,用“木”与“土”就可以拼合成为“杜”和“桂”。你只要在键面上轻轻地“点”一下有关字根,有关字根就会自动组合成你所要的汉字,并显示在屏幕上。

其中的奥妙也很简单。当你按了一个“木”字键时,计算机并没有如实地在机器中“写”进了一个“木”,而是记下了字根“木”所在键的英文字母 S。按了一个“土”键时,机器记下了字母 F。当你连续按了“木”、“土”时,机器记下了 SF。这时, SF 就成了“杜”字的代号(外码)。机器再按照这个代号,从存有几千个汉字图形的“字库”中,把“杜”字的图形调出来,显示在荧屏上,从而就完成了“杜”字的输入。

五笔字型方案采用“字词兼容”技术。在输入操作中,字与词之间没有什么换挡操作。当你连续按下“禾、石、讠、戈”四个字根时,屏幕上就会显示“科研计划”这个词。实际上,机器记下了与“禾、

金魚 金魚 35 Q	人 人 34 W	月 月 33 E	白 白 32 R	采 采 31 T	高 高 41 Y	立 立 42 U	水 水 43 I	火 火 44 O	之 之 45 P
工 工 15 A	木 木 14 S	大 大 13 D	土 土 12 F	王 王 11 G	目 目 21 H	日 日 22 J	口 口 23 K	田 田 24 L	：
乙 乙 55 X	又 又 54 C	女 女 53 V	子 子 52 B	子 子 51 N	已 已 50 N	山 山 25 M	<	>	？
乙 乙 55 X	又 又 54 C	女 女 53 V	子 子 52 B	子 子 51 N	已 已 50 N	山 山 25 M	<	>	？

图 2.1 字根键盘

石、讠、戈”这四个字根相对应的一组字母，即 TDYA，也就是这条词汇的外码。机器根据这组外码，在“词汇库”里把它对应的四个汉字的图形一次调出来。操作的过程和结果联系起来看，就好像是：取“科研计划”一词中每个汉字的第一个字根，就可以组成四个完整的字。其实质是人为地在两者之间建立了对应关系而已。

为了加强这种直观印象，下面再举几例。

例 1：S 键上有“木”，B 键上有“子”，按 SB（再加按一个空格键）就组出了“李”字。

例 2：W 键上有“人”，按 WWW（再按一下空格键）就组出了“众”字。

例 3：I 键上有“讠”，P 键上有“石”，X 键上有“乡”，D 键上有“石”，按 IPXD 就组出了“汉字编码”（即打每一字的第一部分）。

这个过程叫“拼形组字（词）”。据此，五笔字型可把成千上万的汉字和词语输入计算机，这叫“拼形输入”。

2.1.2 汉字的三个层次

汉字是一种象形文字。最古老的汉字只求图形相似而不计笔划曲直多少。秦始皇统一中国之后，车同轨，书同文，统一使用的篆书，还没有从图画文字中蜕变出来。只要画成一定的形状，多一“笔画”，少一“笔画”都没有关系。直到后来，汉字楷化之后，才对汉字图形的线条加以规范化，其笔势有了一定之规，形成了“笔划”。可见，笔划是后来确定的。

一个完整的汉字，并非笔划的线性排列，也非笔划的任意堆积。由若干笔划复合连接交叉所形成的相对不变的结构，叫做“字根”。将字根按一定的位置关系拼合起来，就构成了汉字。

可见，汉字划分为三个层次：笔划、字根、单字。

《五笔字型》方案要求：不考虑读音，不把汉字全部肢解为笔划，而是遵从人们的习惯书写顺序，以字根为基本单位来组字编

码、拼形输入汉字。

如果汉字的分解以笔划为编码单位,那么一个汉字最多的笔划可能有二十多笔,就要击键二十多次,才能完成一个汉字的输入,这样的输入速度无疑是太慢了。如果以字根为分解单位,那么平均最多只击四次键就可完成一个汉字的输入。我们以“湖”字为例来看汉字的分解,如表 2.1。

表 2.1 字根的分解

汉字	字根	笔划
湖	亻	丶、丨、㇏
	古	一、丨、乙、丶
	月	丿、乙、一、丶

由上表可见,若分解到笔划,则编码输入“湖”字要击 12 次键;若分解到字根,则只需击键三次即可。

2.1.3 汉字的五种笔划

楷化后的汉字,其笔划的分类有多种:有人将其分为二十四种,有的人将其分为三十六种。之所以混乱,是因为对笔划的概念含糊不清,没有抓住笔划的本质,只看到它的外表。

下面,给笔划下一个严格的定义:

在书写汉字时,不间断地一次连续写成的一个线条叫做汉字的笔划。有了这个定义,如“二”及“×”根本就不能称为笔划。

对成千上万的汉字来说,仅考虑笔划的运笔方向,而不考虑其轻重长短,则只有五种笔划:横、竖、撇、捺、折五种,命以代号 1、2、3、4、5。如表 2.2 所示。

表 2.2 汉字的五种笔划

代号	笔划名称	笔划建向	笔划及其变形
1	横	左→右	一、𠃍
2	竖	上→下	丨、𠄎

代号	笔划名称	笔划建向	笔划及其变形
3	撇	右上→左下	丿
4	捺	左上→右下	㇏、
5	折	带转折	乙ノㄣㄣ ㄣレレレレ

几点说明：

1. 由“现”是“王”字旁可知，提笔应属于横。
2. 由“村”是“木”字旁可知，点笔应属于捺。
3. 由旧体的“木”竖笔带钩可知，竖左钩应属于竖。
4. 一切带拐弯的笔划，都归为折类。

五种笔划组成字根时，其间的关系可分为四种情况：

1. 单：即五种笔划自身。
2. 散：组成字根的笔划之间有一定的间距，如三、八、彳、心等。
3. 连：组成字根的笔划之间是相连接的，可以是单笔与单笔相连，也可以是笔笔相连，如厂、人、尸、弓等。
4. 交：组成字根的笔划是互相交叉的，如十、力、又、车等。

还有一种混合的情况，即一个字根的各笔划间，既有连又有交或散，例如：“彡”是有连有散，“禾”是有连又有交等。

2.1.4 汉字的三种字型

根据汉字的整体轮廓和组成汉字的各个部分之间的关系，我们可以把成千上万的汉字分为三种组成类型：左右型、上下型、杂合型。并视各型汉字的使用频率的多少，分别给它们编写为 1、2、3。如表 2.3 所示。

表 2.3 汉字组成类型

字型	字型代号	字例
左右	1	江 洲 给 到
上下	2	字 空 花 型

字型	字型代号	字例
杂合	3	因内道乘 太重天日

由表可见：所谓左右型，是指汉字基本上由左右部分组成；上下型则是汉字由上下部分组成；而杂合型是指没有明显的上下或左右结构的那种汉字。表中的杂合型又叫独体字。左右、上下型为合体字。两部分合在一起的汉字叫双合字、三部分合在一起的汉字叫三合字。一般合体字只分到三合字。由于二、三合体字的字根较少，不足四码，所以有必要补加一个字型信息。对于可以拆分为四个（至少四个）部分的汉字，其信息量已足够，就不必考虑字型信息了。

一型：左右型汉字。左右型汉字有以下两种情况：

1. 双合汉字的左右结构。左右字根间有一定的距离。如：明、现、肚、理、咽。而咽的右边是杂合型，但整个汉字是左右型结构的。
2. 三合汉字的左中右结构。这种汉字由左中右三部分构成。如：侧、哪等。另一种三合汉字，虽然是三个部分构成，但并不是左中右结构。如别、谈等。这两种三合汉字的字型结构都归为左右型结构。

二型：上下型汉字。上下型汉字也分为双合字和三合字：

1. 双合字的上下结构。字根的上下间有一定的距离，如：军、声、字、节、旦等。
2. 三合字的上、下结构。这种汉字通常由上中下三部分或者三部分按上下排列而构成，如：竞、想、花等。

三型：杂合（单体、内外、包围）型。三型汉字是指汉字没有明显的左右或上下关系，如：团、同、这、斗、周、建、飞等。

汉字的字型特征，可以作为识别汉字的一个重要依据。如：吧、邑、它们的字根一样，但结构不一样。在输入不足四码的汉字时，除字根输入外，还必须输入字根的排列方式，这样才能准确地识别汉

字。

由前述可知：笔划可以组成字根，字根可以组成汉字。笔划和字根之间存在单、散、连、交的关系。同样，字根与汉字之间也存在单、散、连、交的情况。

1. 单：如“口、木、山、田、马、寸”等，即基本字根本身就单独成为一个汉字。

2. 散：指构成汉字的基本字根之间可以保持一定的距离，如：吕、足、困、识、汉、照等。注意：既然字根间可以保持一定距离，那么它们就有一个相互位置关系问题。要么左右、要么上下、要么杂合，总归要属于一种，从而形成三种不同的字型。

3. 连：指一个基本字根连一单笔划，如“丿”下连“目”成为“自”，“丿”下连“十”成为“干”，“月”下连“一”成为“旦”，“勹”下连“丶”成为“久”等。其中，单笔可连前也可连后。注意，这种情况下的字根与单笔之间，不能当作散的关系。

连的另一种情况是所谓“带点结构”。例如：勾、术、太、主等字中的点，近也可，稍远也可，连也可，不连也可。为了使问题简化，我们规定：一个基本字根之前或之后的孤立点，一律视作是与基本字根相连的。

由此，我们可以看到，一切基本字根与单笔连了之后形成的汉字，都是不能分为几个保持一定距离的部分的。因此，在判断这一类汉字的字型时，它们不可能是左右、上下型，而只能是第三类字型了。

4. 交：是指几个基本字根交叉套迭之后构成的汉字，如：“农”是由“丷·衣”，“申”是由“日丨”，“里”是由“日土”，“夷”是由“一弓人”交叉构成的，等等。

由此，我们可以看到，一切由基本字根相交叉构成的汉字，基本字根之间是没有距离的。因此，在判断这一类汉字的字型时，毫无疑问，它们都是属于第三型了。

字根组字中,还有一种情况是混合型,即几个字根之间既有连的关系,又有交的关系。如“丙”,是“一”下边连了一个“内”,而“内”又是由“冂”与“人”相交形成,自然这类字也属于第三型。

§ 2.2 五笔字型字根

2.2.1 字根键盘

汉字可用几个基本的部分拼合而成,这些用来拼字的基本部分叫做“字根”。能组成尽可能多的常用的字根,叫做“基本字根”。基本字根共有一百三十种(一百二十五种字根,五种基本笔画)。把这些基本字根安排在英文键盘上,就形成了“字根键盘”。五笔字型在键盘上安排字根的方式是:

1. 将 A—Y 共 25 个英文字母键分成五个区,区号为 1—5。每个键称为一个位,位号为 1—5。如果将每个键的区号作为第一个数字,位号作为第二个数字,那么用两位数字(称为区位号)就可以表示一个键。

2. 将一百三十种基本字根按首笔笔画,并兼顾键位设计需要,划分作五大类。每大类各对应键盘上的一个区;每大类又分作五小类,每小类各对应相应区上的一个位。这样,键盘上的一个键就与一小类基本字根相对应。因此,用一个键的区位号或字母就可以表示该键对应的一小类基本字根。

3. 从每个键所表示的基本字根中选出一个组字频度较高而形体上又有一定代表性的字根,作为“字根键盘”的键名,每一个键名也表示一个键。左上角的大写字根为该键的键名。键盘分区及键位安排情况如图 2.2 所示。

五笔字型编码方案字根区位表如表 2.4 所示。

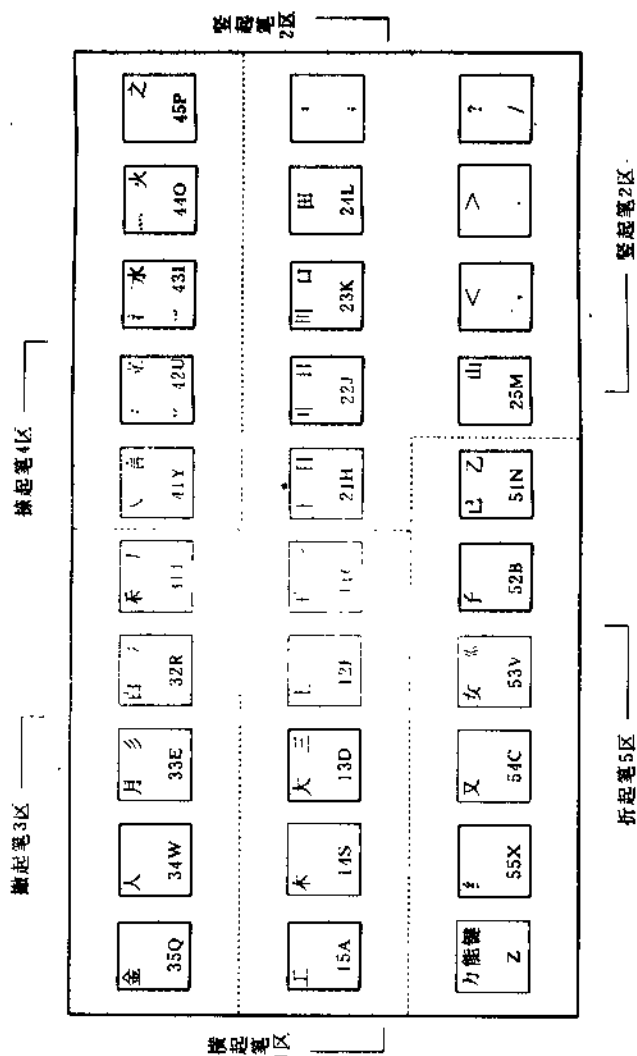


图 2.2 键盘分区及键位安排