嵌入式中文输入法设计

闵华松[1,2]　童学才[2]　刘光临[1]

(1.武汉大学动机学院，湖北 武汉 430072；2.武汉科技大学计算机学院，湖北 武汉 430081)

摘要：本文比较了嵌入式系统中拼音输入法和笔画输入法的优缺点，介绍了嵌入式系统中的中文输入法的基本设计思想。本文重点介绍拼音输入法的设计思想，并针对传统拼音输入法设计的弊端，提出了一种折中的思想，不仅可以很大程度地提高查找效率，而且占用的存储器空间大小也比较合理，在输入法设计的过程中考虑到了输入法的可扩充性。

关键词：嵌入式系统；中文输入法；拼音输入法；笔画输入法

中图分类号：TP311.11　　**文献标识码**：A

Design Chinese Input Method in Embedded System

MIN Hua-song[1,2]　TONG Xue-cai[2]　Liu Guang-lin[1]

1.Wuhan University, Wuhan Hubei, 430072

2.Wuhan University of Science & Technology, Wuhan Hubei, 430081

Abstract:This paper compared the advantage and disadvantage between Chinese PinYin input method and BiHua input method in embedded system. This paper have put forward a new kind of method, which can not only improve the efficiency of the input method, but also can limit the memorizer in a reason size. Furthermore，this method have fully considered the extension of the new input method.

Key words:embedded system;Chinese input method;PinYin input method;BiHua input method

引言：目前嵌入式领域所使用的中文输入法一般都是诺基亚的 T9、摩托罗拉的 iTAP、爱立信的字能，虽然现在很多公司都在进行中文输入法的开发工作，但是这种“三分天下”的局面还没有改变。这些嵌入式中文输入法除了支持中文的拼音和笔画外还支持众多国家语言，比如日文、韩文，而且使用费用高。虽然现在很多流行的 GUI 都融合了输入法，但是出于成本的考虑，自己开发简单实用的 GUI 和输入法目前也为不少公司采用。本文主要介绍嵌入式领域中一种兼顾时间和空间的拼音输入法设计思想。

1 嵌入式拼音输入法和笔画输入法的比较

嵌入式产品一般具备 0－9 几个数字键和一些辅助功能键，输入法的设计是通过逐步查找匹配建立一种汉字内码到汉字点阵的映射关系。

（1）易用性比较。拼音输入法简单易用，不需要用户耗费额外学习与适应的时间，而用户在使用笔画输入法时，要根据具体的输入法学习和适应。

（2）重码率比较。在嵌入式中文输入法中，按键的多少一般与重码率成反比，拼音输入法重码率较高，笔画输入法根据其具体的实现方式重码率也不一样。

（3）效率比较。一般情况下，拼音输入法比笔画输入法的效率低。

（4）程序设计难易比较。拼音输入法设计思想简单，但是具体程序实现比较复杂，因为拼音输入法中每个汉字需要按键次数不一样。

（5）存储空间比较。它们所需要的存储空间大致相当，这个空间主要用来存储汉字的内

码和点阵。

总之，笔画输入法和拼音输入法有各自的特点。

2 嵌入式拼音输入法设计思想

2.1 传统拼音输入法

拼音输入法的设计思想是通过用户的按键操作得到一组数字组合，由这个数字组合得到可能出现的拼音组合，而每组拼音组合对应了一组同音汉字组，这种结构当中实现的是一个二级对应表，如图 1 所示。

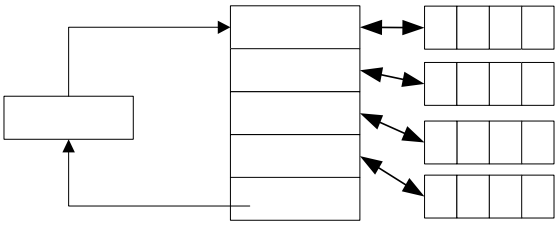

图 1 二级映射表

这个二级对应表一般按照树型结构进行组织，在程序的设计中一般要使用到二叉树的遍历和指针的操作，这个过程需要大量的 CPU 时间，这对嵌入式设备而言是不利的。

2.2 改进的拼音输入法

本文采用了一种比较简单的数据结构来实现嵌入式中文输入法：在存储器中存放两张表格——“检索表”和同拼音汉字组的“内码表”。检索表和内码表在存储器中的具体存放形式如图 2 所示。左半部分是检索表，右半部分是内码表。

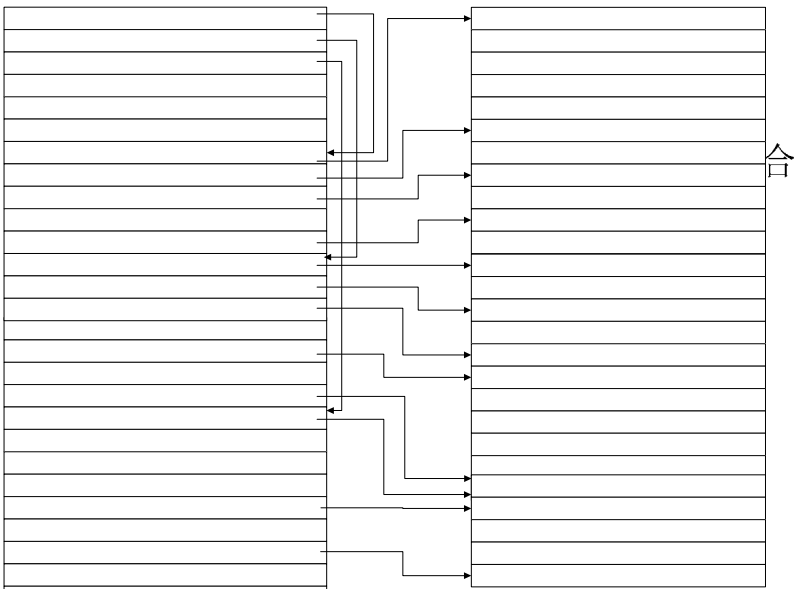

图 2 表格存储结构示意图

2.2.1 检索表

检索表存储的是同拼音汉字组的首地址，它按字母的个数把拼音分为六类（汉字拼音由 1—6 个字母组成），每类对应检索表中的一个子区域，在每个子区域中，首先按照首字母的顺序顺次排列，再按照第二、第三个字母的顺序依次往后排列。

在拼音输入过程中存在很多无效输入，比如字母数量为 1 的拼音中只存在“a”、“e”、“m”、“n”、“o”这五个，又如字母数量为 2 的拼音当中，不存在“aa”、“ab”、“ac”等这

样拼音的汉字，但是在上面表格的设计过程中采用了冗余的做法，仍然在表格当中为它们分配了指到汉字组的地址单元，这样做虽然浪费了存储空间，但是可以很大程度提高检索效率。检索表的这种结构只需使用相对偏移量进行简单的相加操作就可以实现快速查找，而这个偏移量又可以使用它们各个字母的 ASCII 码进行简单的转换来实现。具体地址计算公式如下：

假设输入拼音中字母个数为 n，输入顺序为 $a_1a_2…a_n$（$1 \leqslant n \leqslant 6$），检索表首地址为 addr，函数 f（x）将字母 x 的 ASCII 转换成对应的数字并且减去常数 97，则这个拼音对应检索表中的地址：

$$addr + （6 + 26^1 + 26^2 + … + 26^{n-1} + f（a_1）\times 26^{n-1} + f（a_2）\times 26^{n-2} + … + f(a_n) \times 26^0）\times 4$$

在具体程序中编写这个函数的时候可以使用内联函数来实现，并将计算过程优化，此函数也可以用汇编代码实现并优化，由于这个函数使用频率非常高，所以它的性能对整个输入法的整体性能有直接的影响。

比如在 32 位 MCU 中，检索表的起始地址为 200000H，每个地址都用一个字的空间来存放，当输入拼音“hao”，程序首先找到字母个数为 3 的拼音组存储区域的首地址——距离检索表首地址 6＋26×26 字开始的存储区域，即 200000H＋（6＋26＋26×26）×4H=200B10H；接着计算首字母为‘h’的拼音在字母个数为 3 的拼音区域中的偏移量，为（7×26×26）×4＝127CH（‘h’是 26 个字母当中的第 8 个，排在它前面的还有 7 个，32 位 MCU 中一个字长为 4 个字节）；再计算第二个字母为‘a’的拼音在字母个数为 3 且首字母为‘h’的存储区域中的偏移量，为（0×26）×4＝0000H（‘a’为 26 个字母中第 1 个，没有字母排在它前面）；最后计算‘o’字母在字母个数为 3 且一二字母为‘h’和‘a’的区域中偏移地址，为 14×4＝0038H（排在‘o’前面的有 14 个字母）。所以拼音“hao”对应的汉字组地址存放在以 200B10H＋127CH＋0000H＋0038H＝201DC4H 开始的 4 个连续字节单元中。在这个转换过程中，运用到了每个字母在 26 个字母中的顺序，而这个顺序可以通过每个字母的 ASCII 码字节转换得到：将每个字母（小写）对应的 ASCII 码转换为对应的数字，然后减去 97 即可得到（单个字母在 26 个字母中的顺序编号为 0—25，‘a’ASCII 码对应的十进制数为 97）。

检索表冗余的做法虽然提高了检索效率，但同时也带来一个严重的问题——检索表格需要大量的存储空间来存放：（$6+26^1+26^2+26^3+26^4+26^5+26^6$）×4 字节的存储空间，这个很不现实。实际上字母个数为 5、6 的拼音很少（字母为 6 个的拼音只有“chuang”、“shuang”、“zhuang”三个，字母为 5 个的拼音也只有 24 个），使用频率低，但是它们耗费了大量的存储空间，所以对字母个数为 5、6 的拼音并没有采用冗余的做法，而采用顺序查找的办法，这个时候需要用到内码表中存储的字母 ASCII 码来进行比较。当然，这些字母的 ASCII 码也可以存放到检索表中，这样可以避免地址跳转操作，节省 CPU 时间。

2.2.2 内码表

内码表中，同音汉字组的开头部分存放拼音字母的 ASCII 码，用于查找比较，结尾部分存放“FFFFH”，表示该同音汉字组的结束。

至于汉字的编码，现在广泛使用的有 GB 码、GBK 码、BIG5 码、HZ 码等，在没有操作系统支持的情况下，这些编码可以根据实际需要编码，但是现在的嵌入式设备一般都少不了操作系统的支持，不同的操作系统支持的汉字编码不一样，开发人员需要对所使用的操作系统有一定的了解。

在设计程序的时候，由于这两个表格需要的空间较大，需要注意寻址空间的问题。

3．内码检索的程序流程图：

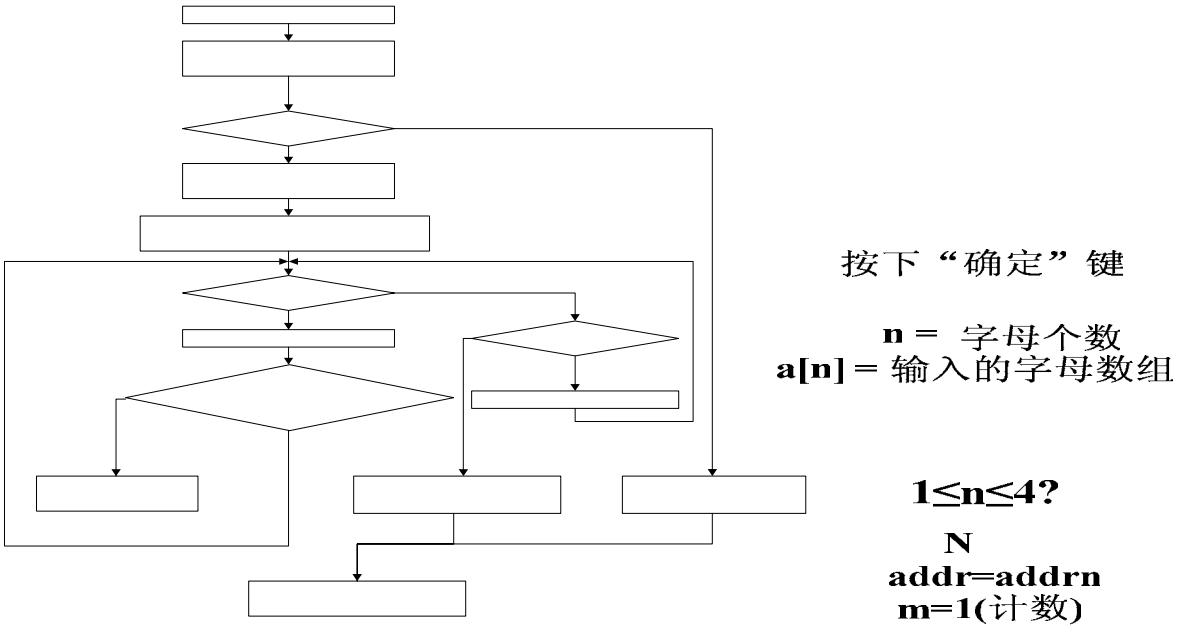

图 3 内码检索流程图

在用户按下“确定”键后，系统首先取得拼音中字母的个数 n，并将拼音字母存到数组 a[n]中，然后判断字母个数是否在 1－4 之间，如果在，则可以直接运用公式计算出检索表地址，通过此地址找到汉字内码并显示，如果拼音中字母个数不在 1－4 之间，这个时候需要使用逐个查找的办法得到此拼音对应汉字组的地址。

对字母个数为 5、6 的拼音进行查找时，首先将字母个数为 5 或 6 的子区域首地址赋给 addr，然后逐步查找，将这个地址逐步精确化，精确化的过程就是通过每个字母逐步比较。其中“（addr－4）中的 Am 的 ASCII 码”和“addr 中的 Am 的 ASCII 码”的比较用于判断用户输入拼音的第 m 个字母是否有效——因为‘a’－‘z’的 ASCII 码是从小到大排列的，如果查找完一次之后，检索表中对应的这一位没有和它匹配的，则说明这个时候的输入无效，程序返回。“m＜n？”用于判断已经比较了几次，如果每个字母都比较了一次，这个时候说明匹配成功，此地址就是需要查找的地址。

4．需要注意的几个问题：

（1）输入过程中无效输入的处理。比如当拼音输入到“duan”时，“k”是一个无效的输入，但是当输入到“duan”、“kai”的时候又成为了一种有效输入，对于这种无效的输入，在嵌入式手持设备的设计中，一般是不允许将无效输入显示在液晶屏上的，这样的无效输入可以在检索表中查找判断。

（2）输入的联想功能。实现联想功能需要添加词库，这样会增加大量的存储空间来存放这样的词库，词库的存放顺序直接影响到用户，PC 机上很多输入法都实现了词库的动态更新，随着嵌入式 MCU 性能的不断提高和存储容量的不断增大，这个技术在嵌入式输入法当中必将得到应用。

（3）关于特殊符号、英文、数字的处理。关于这些特殊符号的输入，一般是用 2 个特殊功能键来实现，一般情况下一个按键用来进行输入法切换，另外一个用来翻页此输入状态下允许的特殊符号，用户再通过按键来选择这些特殊符号。

（4）关于汉字的点阵。点阵文件可以自己制作，也可以选择现成的文件，比如 UCDOS 中的汉字点阵文件，一般情况下由内码可以通过公式直接算出点阵数据在文件中的位置。一个点阵文件可以实现一种字体的显示，多个点阵文件可以实现多种字体的显示。

（5）关于笔画输入法。笔画输入法和拼音输入法的设计思想相同，但是笔画输入法的设计比拼音输入法的设计要灵活，因为它的笔画划分方式以及输入方式各异，现在出现的笔画输入法种类繁多，设计时要注意它的易用性，减少用户按键的疑虑。
（6）关于输入法的可扩充性。按照上面的思想设计出的嵌入式中文输入法，具有良好的可扩充性，任意一种输入法所需要进行修改的地方主要是它的内码表，这个也直接关系着输入法的易用性。

5．结束语：

本文作者创新点：针对传统输入法设计的弊端，不仅很大程度地提高了程序查找效率，而且软件设计上占用的存储器空间也比较合理，在输入法设计的过程中考虑到了软件的可扩充性。

参考文献:

[1] 电子产品世界.常江.嵌入式系统中文输入法的设计[J]. 2004.9:70-74.
[2] 现代计算机.蔡如海.LED 显示屏拼音输入汉字的实现[J]. 2001.5:61-63.
[3] 信息技术.陈天鹅,赵曾贻.朱兰.数字键中文输入的研究[J]. 2002.1:49-51
[4] 上海计量测试.裘东.一种快速匹配算法在拼音输入整句翻译中的应用[J]. 2003年30卷第一期:17-19
[5] 微计算机信息. 刘久富等. 嵌入式软件的动态测试[J]. 2006年第1-2期:82-84

基金项目：

湖北省机械传动与制造工程省重点实验室基金项目（2003A02）

作者联系方式:

电话: 13797000784 13971365898
邮箱: zogtong2003@163.com
地址: 武汉市青山区和平大道947号#47信箱 邮编:430081

作者简介:

闵华松(1969—)，男(汉族)，湖北武汉人，副教授，博士，研究方向：嵌入式系统；童学才(1981—)，男(汉族)，湖北武汉人，硕士研究生，研究方向：嵌入式系统；刘光临，男(汉族)，教授。

MIN Hua-song(1969—),male(the Han nationality),Wuhan Hubei, associate professor, research direction:embedded system,;TONG Xue-cai(1981—), male(the Han nationality),Wuhan Hubei, master, research direction:embedded system;Liu Guang-lin,male(the Han nationality), professor.

1.文章题目：嵌入式中文输入法设计
（Design Chinese Input Method in Embedded System）
2.作者：闵华松 童学才 刘光临
3.中图分类号: TP311.11
文献标识码: A
4.基金项目：湖北省机械传动与制造工程省重点实验室基金项目
（2003A02）

5.作者联系方式:
邮箱: zogtong2003@163.com
地址: 武汉市青山区和平大道947号#47信箱
邮编:430081