DOI:10.13203/j.whugis1985.02.008

手写体汉字模式识别

王葆元　　邓铁清

摘　要

本文介绍了一种运用统计决策法和句法结构法识别手写体汉字的方法。该识别方法基于下列事实：汉字可以用汉字语法树表示。对每个汉字模式可通过字根分离和字根识别生成语法树来描述。

识别过程如下：首先，搜索汉字中的“结构缝隙”分离字根，然后，根据模糊数学的理论对分离出的字根进行模糊聚类、比较、识别，生成汉字语法树，最后，中序遍历语法树组装汉字，查找汉字字典，输出识别结果。

该识别方法在APPLEⅡ微型机上作了模拟试验，得到了初步验证。

一、引　言

模式识别是人工智能研究领域之一，它的狭义研究目标是为计算机配置各种感觉器官，以便直接接受外界的各种信息。

汉字模式识别，是把以汉字作为媒体的信息进行输入输出。在实现最自然的数据处理中，汉字模式识别是不可缺少的一门技术。在计算机控制系统中和计算机进入国家各职能部门后，汉字模式识别技术也是一个十分关键的研究课题。

近年来，中外科学家对汉字识别作了大量的研究，提出了许多识别汉字的方法，但归纳起来，不外乎两种：统计决策法（简称统计法或决策论法）和句法结构法（简称句法方法）。统计决策法是从输入的汉字图象中抽取一组称为特征的特征向量（每一汉字图象用一特征向量表示），再根据既定的规则判别该汉字的类别，达到识别的目的。句法结构法是把汉字模式分解成若干简单的元素，然后用特殊的文法规则来描述这些元素之间的结构关系进行识别。

本文提出了综合利用上述两种方法对手写体汉字进行识别的方法。通过对汉字“结构特征”的分析，发现了汉字可用汉字语法树表示这一事实，构造了一个手写体汉字的识别系统。在汉字识别过程中，采用句法结构法，搜索汉字“结构缝隙”分离字根；运用统计决策法，对变体字根进行模糊聚类、比较、识别。

二、汉字“結构特征”的分析

汉字种类众多，字体千姿百态，但汉字本身具有内在结构规律。汉字结构类型一般可分

本文1984年12月收到。

为单体字和合成字两大类。合成字的结构有十几种以上，且分法完全相同，但主要的结构类型分为三类，即上下结构，左右结构和内外结构。

以汉字的第一层结构命名汉字的结构类型有：

左右型　如：

联　街　清　刹

上下型　如：

吴　章　挚　花

内外型　如：

国　这　唐　同　凶

除第一层结构外，多数汉字还有第二层结构，第三层结构，……，汉字结构是嵌套的。在汉字识别中，把汉字看成多层次的结构关系可减小字根识别中的字根集，但同时也增加了字根分离的复杂性。

汉字结构可用汉字结构文法描述。在规定汉字结构文法之前，先定义三个关系运算符。

定义一　左右关系运算符①。

若x和y有关系①，即x①y，我们说x在y的左边，或者说y在x的右边。利用①，汉字“联”可表示为：耳①关。

定义二　上下关系运算符㊀。

若x和y有关系㊀，即x㊀y，我们说x在y的上面，或者说y在x的下面。利用㊀，汉字“青”可表示为：主㊀月。

定义三　内外关系运算符⊙。

若x和y有关系⊙，即x⊙y，我们说x在y的里面，或者说y在x的外面。利用⊙，汉字“国”可表示为：玉⊙口。

汉字结构文法定义为如下四元组

$$G=(V_T, V_N, S, P)。$$

其中　$V_T=\{$ 〈字根〉，〈单体字〉 $\}$；

$V_N=\{$ 〈部件〉，〈运算符〉 $\}$；

$S=$(汉字)；

$P=\{$ 〈汉字〉::=〈单体字〉|〈部件〉〈运算符〉〈部件〉，

〈部件〉::=〈字根〉|(〈部件〉〈运算符〉〈部件〉)，

〈运算符〉::=① |㊀ |⊙ $\}$。

根据汉字结构文法，任一汉字可以用汉字语法树表示。如图 1，图 2 所示。

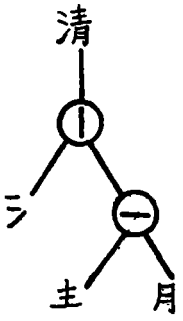

图 1 汉字“清”的语法树

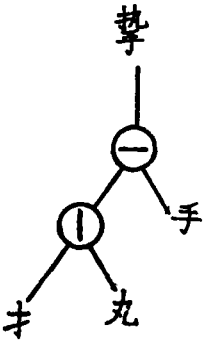

图 2 汉字“挚”的语法树

汉字语法树是一颗二叉树。这是因为汉字结构文法中三种关系运算符都是二目运算符，且都不满足交换率。在计算机中，语法树中的任一节点用三个域表示，如图 3 所示。

LCHILD	DATA	RCHILD

图 3 语法树中的节点在计算机中的存贮表示

图中 LCHILD：节点与左子节点的链接指针，无子节点时为空（Λ）；

RCHILD：节点与右子节点的链接指针，无子节点时为空（Λ）；

DATA：节点有子节点时，存放左右两个子节点间关系的运算符（①、㊀、⊙）的编码，无子节点时，存放字根（或单体字）的编码。

图 4 是汉字“挚”的语法树在计算机中的存贮表示。

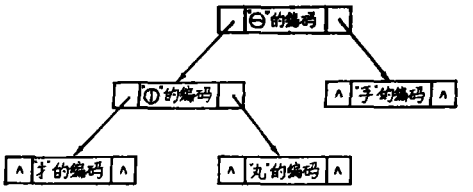

图 4 汉字“挚”的语法树在计算机中的存贮表示

三、汉字识别系統

汉字可用汉字语法树表示，基于这一事实，本文提出了如图 5 所示的手写体汉字识别系统。在汉字识别时，通过对字根分离和字根识别系统生成待识汉字的语法树，经汉字组装形成汉字编码，通过查找汉字字典得到识别结果。

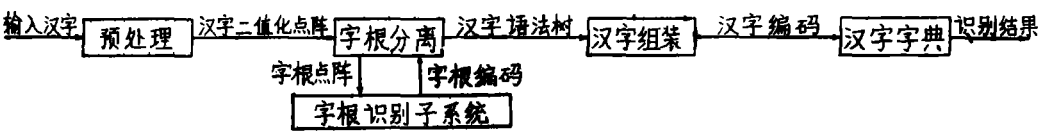

图 5 汉字识别系统

1. 预处理

用光导摄象机对输入汉字图象扫描，将扫描图片转换成64×64点阵，每一点阵取辉度值为1（黑）或0（白）。再将汉字大小作比例调整及笔划宽度细化等，进行标准化处理，最后输出汉字二值化点阵。预处理由专门课题研究，不是本文的重点，在此不作述说。

2. 汉字字根的分离

字根分离有多种方法，本文采用搜索汉字中的“结构缝隙”法，它具有边识别结构类型和边分离字根的特点。

搜索汉字中的“结构缝隙”法要求组成汉字的字根不相连。由于解决字根不相连的字根分割法不甚完善，它实际上还是一种对限定性手写体汉字的字根分离方法。

根据汉字合成字的三种基本结构，即左右结构，上下结构和内外结构，搜索汉字中的“结构缝隙”依次是上下扫描、左右扫描和内外扫描。由于汉字构形和手写笔划的畸变等问题，汉字中的“结构缝隙”除了直线型以外，还存在许多非直线型。所以，搜索汉字中的“结构缝隙”既要搜索直线型，又要搜索非直线型。

搜索直线“结构缝隙”的方法是在汉字实际书写的X_{min}～X_{max}，Y_{min}～Y_{max}范围内，当$\sum_{Y_{min}}^{Y_{max}} A_{x_i,y} = 0$存在，为左右结构，$x_i$处为缝隙(如图6)；当$\sum_{X_{min}}^{X_{max}} A_{x,y_i} = 0$存在，为上下结构，$y_i$为缝隙（如图7）。

式中$A_{x,y}$表示汉字二值化点阵，$A_{x,y}=\begin{cases}0\ (白点)\\1\ (黑点)\end{cases}$。

非直线“结构缝隙”，分折线“结构缝隙”和弯曲“结构缝隙”两种。折线“结构缝隙”存在于  等结构的汉字中。搜索折线“结构缝隙”和搜索直线“结构缝隙”的原理近乎相同，主要差别在于前者搜索过程中要改变扫描方向。弯曲“结构缝隙”是由汉字本身构形、笔划的伸缩、倾斜等引起的。搜索弯曲“结构缝隙”的原理如图8所示。以x_i为出发点，在Δ范围内，若存在由$A_{x_i\pm\Delta x}$，$y_{min}=0$到$A_{x_i\pm\Delta x}$，$y_{max}=0$（$\Delta x=\{0, 1, \cdots\cdots, \frac{\Delta}{2}\}$）的白点路径，则存在左右结构，该弯曲形的白点路线为分界线。

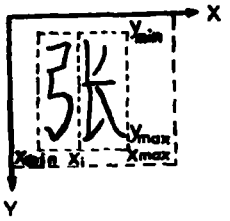

图6　x_i是汉字“张”的左右结构缝隙

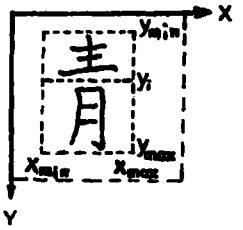

图7　y_i是汉字“清”的上下结构缝隙

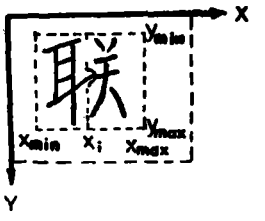

图8　汉字“联”的左右弯曲结构

汉字结构是嵌套的，因此字根分离是一个递归的过程。

3. 汉字字根识别

汉字字根有二百个左右，字根识别比英文字母和数字的识别困难得多，但和识别汉字相比，它具有下列优点：一是数量少。汉字字根仅是汉字数目的百分之一；二是图形简单。建立字根识别字典不象汉字识别字典那么庞大、惊人。本文采用统计决策法识别字根、抽取字根的笔划轮廓投影作为特征向量，以提高识别速度和减少信息存贮量。如图 9 所示。

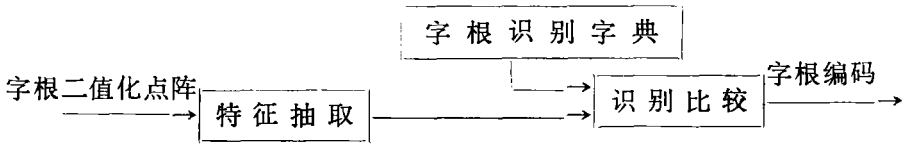

图 9 字根识别子系统

1）特征抽取

特征抽取就是抽取表征字根的信息量（或特征向量）。字根在计算机中用一个n×n的网格表达。在特征抽取前，按照比例将分离出的字根二值化点阵（$X_{max}-X_{min}+1$）×（$Y_{max}-Y_{min}+1$）变换为n×n的网格。由于特征向量总数较大，在字根识别时既耗费机时又占用存贮空间。为了压缩信息量和减少识别时间，采用轮廓投影法抽取字根的特征向量。

轮廓投影法是将字根的笔划投影在垂直和水平坐标轴上，以获得字根轮廓投影信息，将二维图形变换为一维轮廓投影向量值的投影法。

水平轮廓投影为：$X(i)=\sum_{j=1}^{n} f(i,j)$

$i=1,2,\cdots\cdots,n$

垂直轮廓投影为：$Y(j)=\sum_{i=1}^{n} f(i,j)$

$j=1,2,\cdots\cdots,n$

其中f（i，j）是字根在机内的网格表示。图 10 为轮廓投影法之一例。

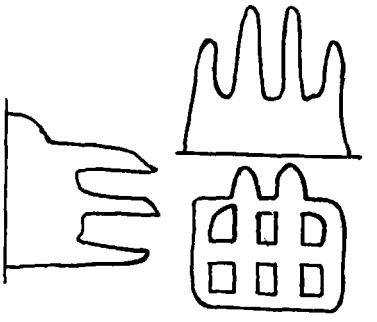

图10 单体字“曲”二值化网格图形及其在水平和垂直方向轮廓投影

利用数学归纳法可以证明，采用轮廓投影法后，信息量由原始的n^2比特压缩到现今的$2n\log_2(n+1)$比特。

虽然用轮廓投影法能压缩信息和提高识别速度，但其抵抗字根位移的能力很低，可采用付立叶变换技术，只取其频率域的振幅为特征向量。

其振幅的频谱用离散形式写为：

$$F'(u)=\left|\frac{1}{n}\sum_{i=1}^{n}F(i)*EXP(-j2\pi ui/n)\right| \qquad (u=1,2,\cdots\cdots,n-1)$$

式中$j=\sqrt{-1}$，u为频率变量，F(i)为变换前的一维特征向量，F′(u)为变换后的幅值谱。在实际计算中，可用n点的快速付立叶变换（FFT）进行。

分别用X(i)和Y(i)替换F(i)，得到

$$X'(u)=\left|\frac{1}{n}\sum_{i=1}^{n}X(i)*EXP\ (-j2\pi\ ui/n)\right|$$

$$Y'(u)=\left|\frac{1}{n}\sum_{i=1}^{n}Y(i)*EXP\ (-j2\pi\ ui/n)\right|$$

字根的特征向量P(k)定义为

$$\begin{cases}P(k)=X'(k)\ ; & k=1,2,\cdots\cdots,n\\ P(k)=Y'(k-n)\ 。 & k=n+1,n+2,\cdots\cdots,2n\end{cases}$$

2）字根集的模糊聚类分析

聚类分析是近十几年来发展很快的一种数学方法，其基本任务是对所考察的对象进行合理地分类。因手写字体笔划的不规则性带来字体本身的模糊结构。本文采用基于模糊等价关系的系统聚类法对字根集进行分类，以提高识别比较的速度和降低字根识别的误识率。分类的步骤分述如下：

① 求模糊相似关系矩阵$\underset{\sim}{R}$。

设字根集中不同字根的个数为M，各字根的特征向量分别为$\overline{P}_1$，$\overline{P}_2$，……，$\overline{P}_M$，每个特征向量又有2n个指标，则M个字根构造的相似关系矩阵$\underset{\sim}{R}$定义为

$$\underset{\sim}{R}=\begin{pmatrix}S_{11} & S_{12}\cdots\cdots S_{1M}\\ S_{21} & S_{22}\cdots\cdots S_{2M}\\ \vdots & \vdots\ \ \vdots\\ S_{M1} & S_{M2}\cdots\cdots S_{MM}\end{pmatrix}$$

其中$S_{ij}=1-\frac{1}{2n}\sum_{k=1}^{2n}\left|\frac{P_{ik}-P_{jk}}{P_{ik}+P_{jk}}\right|$表示分离出的字根i和字根j的类似度。$\overline{P}_i=(P_{i1},P_{i2},\cdots\cdots P_{i2n})$；$\overline{P}_j=(P_{j1},P_{j2},\cdots\cdots P_{j2n})$。

用上述方法建立起来的关系$\underset{\sim}{R}$，一般说来只满足自反性和对称性，不满足传递性，不是模糊等价关系，需要将$\underset{\sim}{R}$改造成模糊等价关系矩阵Q，然后得到聚类图，在适当的阀值上进行截取，便可得到所需要的分类。

② 求模糊等价关系矩阵Q。

将$\underset{\sim}{R}$改造为模糊等价关系矩阵Q的方法，即对相似矩阵求传递闭包的方法，是采用平方法，也就是将$\underset{\sim}{R}$自乘得$\underset{\sim}{R}\circ\underset{\sim}{R}=\underset{\sim}{R}^2$，再自乘$\underset{\sim}{R}^2\circ\underset{\sim}{R}^2=\underset{\sim}{R}^4$，……这种过程继续进行到相邻两次所得到的合成矩阵完全重合为止，最后得到的合成矩阵Q就是模糊等价关系矩阵。

这里$\underset{\sim}{R}\circ\underset{\sim}{R}$是一个新矩阵，它的第i行、第j列元素$q_{ij}$定义为

$$q_{ij}=\max_{1\leqslant K\leqslant M}\left[\min\ (S_{iK},S_{Kj})\right]$$

同样，$\underset{\sim}{R}^2\circ\underset{\sim}{R}^2$、$\underset{\sim}{R}^4\circ\underset{\sim}{R}^4$、……也都采用上述取大、取小运算法则求$q_{ij}$。

③ 截集λ值的选定

正确确定字根集的分类，关键在于选定合适的λ值。但这是一件困难的事情，需要对分类结果有个大致的估计，对统计资料作全面考查，反复修改λ值，在不断比较分类结果的基

础上，才能选取最佳的λ值。为便于分类，可选取一个λ递增序列，将分类结果绘成聚类图（如图11）。从图中可以看出各类归并情况。

由于字根集比较大，选定一个λ值得到一个分类往往不能得到满意的结果，一般需要多次选定λ值进行多重分类。

图12是按上述步骤画出的字根集的模糊聚类分析粗框图。

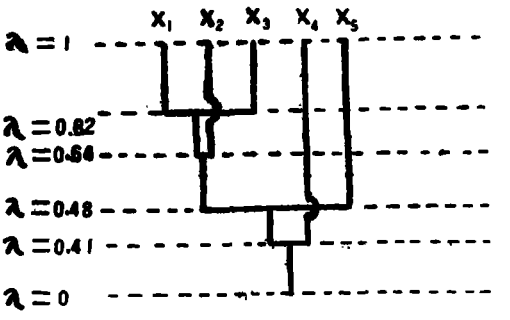

图11 字根集的聚类图

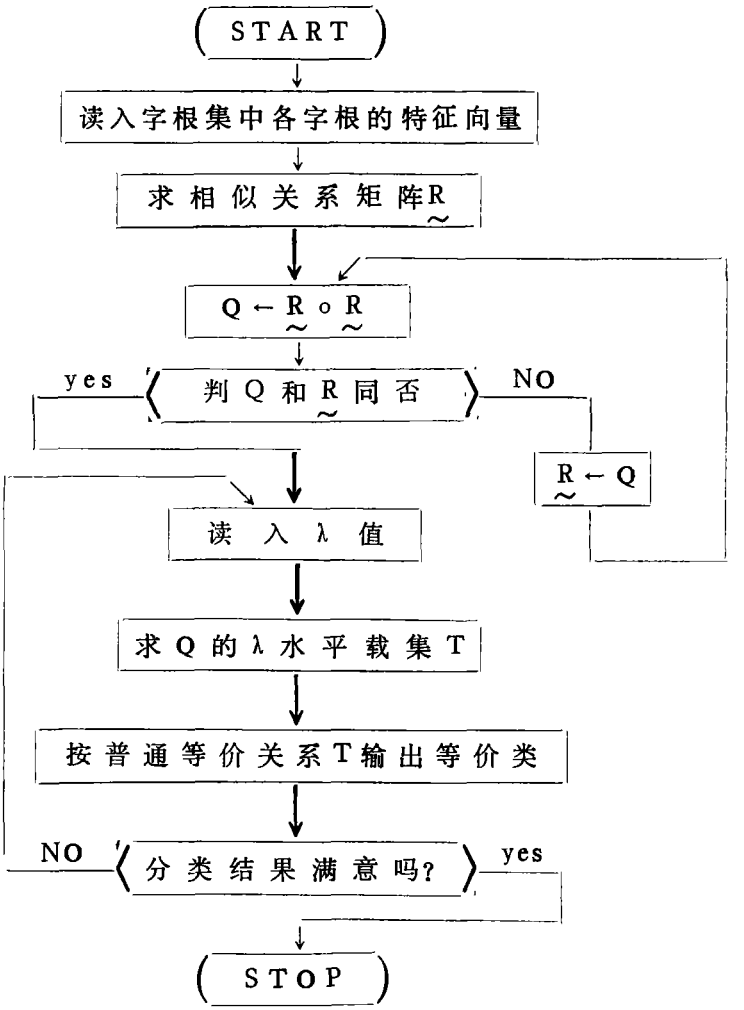

图12 字根集的模糊聚类分析粗框

3）字根识别字典的自动生成

手写体汉字因存在不同的书写者书写同一汉字造成的各种变体的问题，因而字根识别字典不能仅仅存贮一种书写体的字根的特征向量，需要计算机对字根进行学习和训练，以便取

得各种书写体的特征向量的平均值而组成字典，如图13。

字根 1	平均特征向量$\overline{P_1}$	类似度阈值$\overline{S_1}$	编　码	链 接 指 针
	⋮	⋮	⋮	⋮
字根 i	平均特征向量$\overline{P_i}$	类似度阈值$\overline{S_i}$	编　码	链 接 指 针
	⋮	⋮	⋮	⋮
字根 m	平均特征向量$\overline{P_m}$	类似度阈值$\overline{S_m}$	编　码	链 接 指 针

图13　字根识别字典

①　平均特征向量。设字根 i 的训练集 $C_i=\{P_{i1}, P_{i2}, \cdots\cdots, P_{im}\}$，其中 P_{ij} 是字根 i 的第 j 种书写体的特征向量，m是字根 i 的训练集大小。字根 i 的平均特征向量

$$\overline{P}_i=\frac{1}{m}\sum_{j=1}^{m}P_{ij}。$$

②　类似度阈值。字根 i 的类似度阈值 $\overline{S}_i$ 定义为：当且仅当待识字根与字根 i 的类似度大于 $\overline{S}_i$ 时，字根 i 才可能是该待识字根的识别结果。

字根 i 的类似度阈值 $\overline{S}_i$ 仍由训练集 C_i 确定。设 S_j 表示 P_{ij} 和 $\overline{P_i}$ 的类似度，则 $\overline{S}_i=\frac{1}{m}\sum_{j=1}^{m}S_j$。根据需要，$\overline{S_i}$ 可向上或向下浮动一个Δ值。

③　编码。字根 i 的编码是人为规定的代表字根 i 的数字串。

④　链接指针。它是用来链接同一聚类（由字根集的分类确定）中的下一字根。

4）字根识别算法

字根识别就是抽取待识字根的特征向量与识别字典比较、输出的过程。字根识别算法如下：

①　将输入字根的二值化点阵变换为n×n网格。

②　特征抽取。求字根笔划的轮廓投影，作FFT计算幅值谱，得到输入字根的特征向量 $X=(x_1, x_2, \cdots\cdots, x_{2n})$。

③　将输入字根与识别字典中的字根顺序比较，直至输入字根和字典中某个字根（设为字根 i）的类似度（S_i）大于其类似度阈值（$\overline{S_i}$）为止，此刻进行第④步。若顺序比较整个识别字典还找不到这样的字根，则打印拒识标志，停机。

④　取出字根 i 的链接指针送 j。

⑤　判 j 空否？若空，取出字根 i 的编码，并返回；若不空，转下一步。

⑥　计算输入字根和字根 j 的类似度 S_j。若 $S_j-\overline{S_j}>S_i-\overline{S_i}$（$\overline{S_j}$ 表示字根 j 的类似度

阀值），则$S_i \leftarrow S_j$，$i \leftarrow j$，返回④步；否则，取出字根 j 的链接指针送 j，返回⑤步。

4. 汉字语法树的生成

汉字语法树是在字根分离和字根识别过程中建立的。字根分离搜索出结构类型或者字根识别得到字根编码信息，调用建树子程序建立树节点。在建树过程中，为了找到节点在树上的正确位置，用一个路径指示栈（STACK）指示从根节点到挂接位置的搜索路线。若栈指针指示栈的内容为真，沿着当前节点的左子树向下搜索；若栈指针指示栈的内容为假，则沿当前节点的右子树向下搜索。栈指针的初始状态如图14。

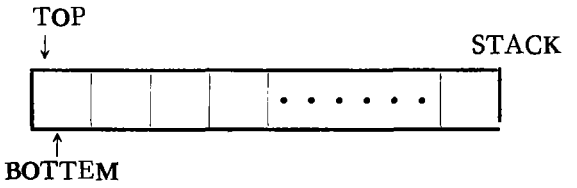

图14 路径指示栈（STACK）的初态

5. 汉字组装

汉字组装旨在形成汉字编码。汉字语法树生成后，汉字组装的工作就是遍历汉字语法树。树的遍历一般分为三种，即先序遍历、中序遍历和后序遍历。在此，为了和汉字结构文法给出的汉字句子表示相对应，采用了中序遍历，其算法描述为：

1）访向左子树

2）访向根节点

3）访向右子树

6. 汉字字典的查找

遍历汉字语法树得到汉字编码后，查找汉字字典即可输出识别结果。由于汉字编码不仅冗长，而且编码的长度又因字而异。若采用顺序查找法，不仅字典的构造不经济，而且查找速度太慢，因此采用了HASH查找法，HASH冲突用多重HASH表解决。如图15所示。

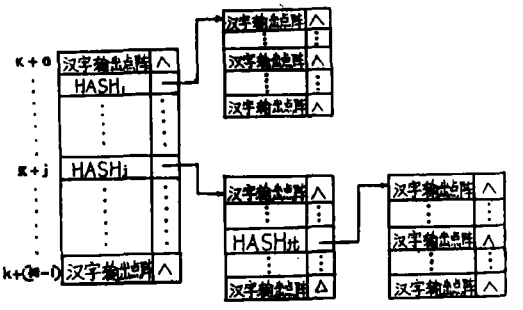

图15 汉字字典

若已知汉字编码 α，HASH 查找汉字字典的过程是：

1）计算A = K + HASH(α)，其中HASH(α)：0～N－1。

2）按地址A访向字典，取出链接指针送LINK判 LINK空否？若空，取出汉字输出点阵输出；若不空，调用相应的HASH函数过程，设为 $HASH_j$，计算 $A = LINK + HASH_j(\alpha)$，返回 2）重复下去。

7. 汉字识别系统框图

汉字识别系统主要由汉字识别主程序，字根分离子程序和字根识别子程序等组成，各部

分的框图如图16所示。

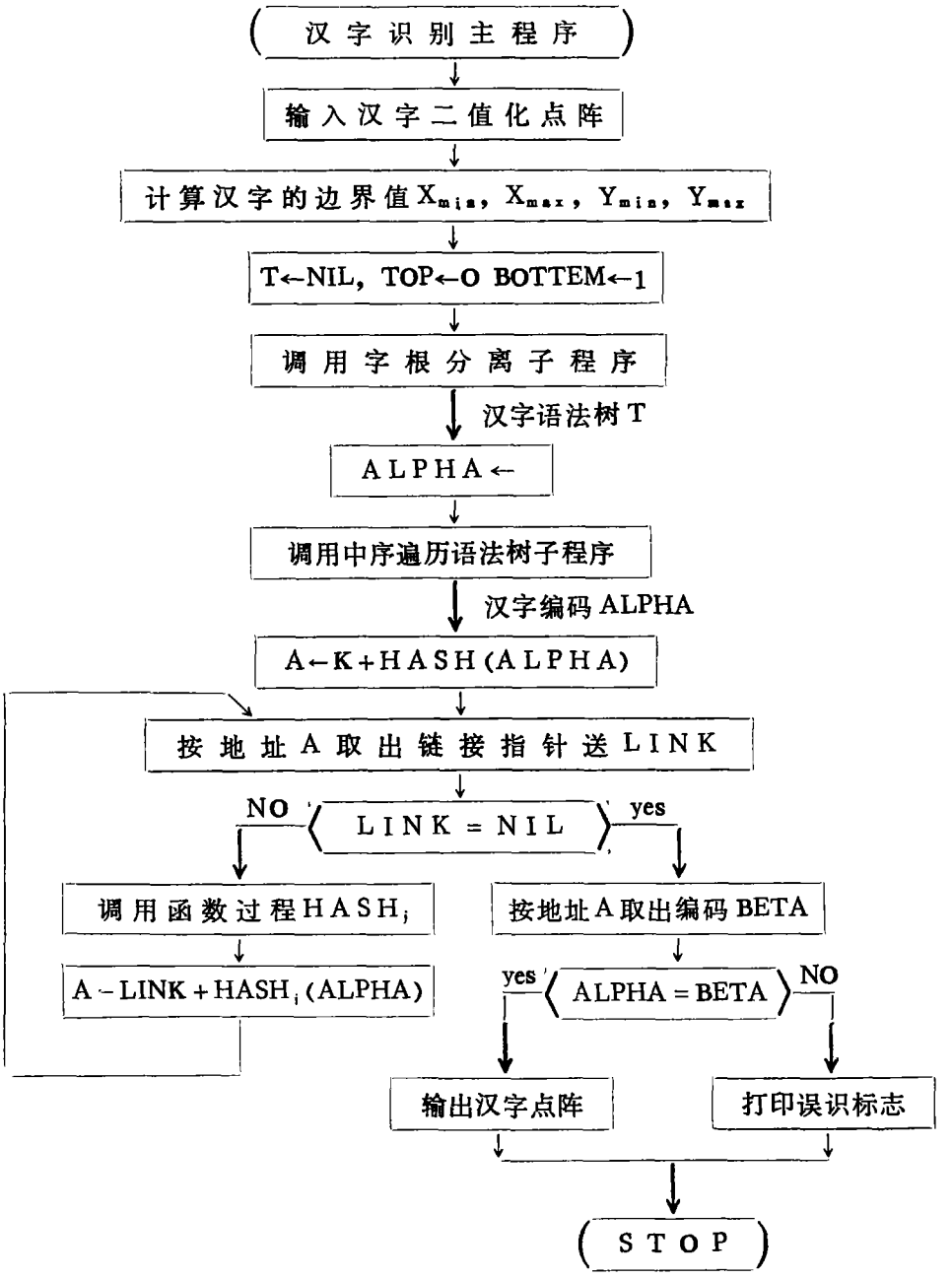

图16　汉字识别主程序

四、识别试验

应用 BASIC 和 PASCAL 语言在APPLE Ⅱ 微型计算机上对本文提出的识别方法作了模拟

试验。试验共分三步，第一步：字根集的聚类分析；字根识别字典的自动生成；字根识别。第二步：搜索汉字中的“结构缝隙”，分离汉字字根。第三步：汉字识别。图17、18分别是两个手写体汉字“汩”和“晶”的识别结果。

a b

a。输入字体“汩”

b。输出字体“汩”

图17

a b

a。输入汉字“晶”

b。输出汉字“晶”

图18

五、結 束 語

本文提出的手写体汉字的识别方法是统计法和句法方法的综合运用。它具有下列优点：（1）仅建立字根识别字典，比传统的建立汉字识别字典大大节省了存贮空间；（2）字根识别字典是通过字根训练集自动生成的，（3）汉字识别算法简明；（4）和拼字法键输入汉字兼容。

参 考 文 献

〔1〕J•E•Hopcroft, J•O•Vllman，形式语言及其与自动机的关系，科学出版社，1979.

〔2〕冯德益、博世楼等，模糊数学方法与应用，地震出版社，1983.

〔3〕中野康明等，作为电子计算机输入的汉字识别，信息与控制，3，1979.

〔4〕蔡国廉，用模糊子集理论识别手写印刷体汉字，高教系统人工智能会议论文，1983.

〔5〕夏莹、张炘中，自动识别手写印刷体汉字系统中的部件分离问题，高教系统人工智能会议论文，1983.

Pattern Recognition of Chinese Characters in Handwriting

Wang Baoyuan *Deng Tieqing*

Abstract

This paper presents a way which recognizes the Chinese characters in handwriting on the basis of statistical decision making and sentence structure. The method of recognition can work because a Chinese character can be represented by a grammar tree. Each pattern of Chinese character can be described by generating a grammar tree through separating the roots of the character and the recognition of them.

The procedure of the recognition is as follows: firstly, "gaps of structure" of a Chinese character are searched to separate the roots of the character, and then, the roots separated are clustered fuzzily, compared and recognized according to the theory of fuzzy mathematics, finally, the grammar tree is traveled symmetrically, the Chinese character is assembled, the Chinese dictionary in the computer is looked upon and the result of recognition is output.

This method of recognition has been proved through the analogous experimentation on micro-computer Apple Ⅱ.