中 国 科 学
1989 年 5 月　　A 辑　　第 5 期

可回溯点跟踪包含配选法印刷汉字识别

郭宝兰　张彩录　马颖丽　李素琴
（河北大学电子系，保定）

摘　要

在汉字识别的包含配选法、双重包含配选法以及点跟踪包含配选法的基础上，提出了一种新的汉字识别的方法——可回溯点跟踪包含配选法．本文论述了该方法的基本原理及用此方法，在C1000光导摄象机与 HP-9000 小型机所组成的图象处理系统上，对4种印刷汉字的识别实验，其结果表明这种识别方法设计合理，具有较强的适应能力．

关键词：模式识别，汉字识别，印刷汉字识别，模式信息处理，中文信息处理

一、引　言

随着人工智能研究，汉字信息处理技术的进展，汉字机器识别课题提上了议事日程．这是因为：其一，汉字识别是人工视觉必不可少的组成部分；其二，汉字机器自动读入是解决汉字信息处理中，输入"瓶颈"问题的重要途径．

汉字就其存在形态而论，有印刷汉字和手写汉字．印刷汉字的特点是字型规整地分为不同的字体．每种字体具有鲜明的笔划特征．手写汉字其字型因人而异，现在多限于研究限制性手写汉字．就印刷体汉字识别而论，日本在 1977 年完成了世界上第一个实验装置，可识别 2000 个汉字，识别速度为 100 字/s．在实际阅读专利公报时得到 98.4% 的识别率[1]．1984 年，日本又研制出多种印刷体汉字识别装置，该装置可识别 2300 个汉字，识别速度大于每秒 100 个字，对印刷在 OCR 专用纸上，固定大小（4.84 mm × 4.84 mm）的文字的识别实验中，识别率为 99.88%[2]．

汉字是我国使用的主要文字，我国汉字识别的最棘手的问题是字域庞大，以信息交换用汉字编码字符集收录的汉字计算为日本机器识别字数的 3 倍，这样，如若采用与日本同样的识别方法，在技术指标上与日本相比较，标准辞书所需的存贮容量至少要大 3 倍，识别速度和识别率将大大降低．因此，为使我国汉字识别的各项技术指标赶上并超过世界先进水平，就不能照搬国外的识别方法，而应创立适合我国文字特点的新方法，可回溯点跟踪包含配选法，就是我们在这方面做的一点尝试．

人们在识别文字时，注意的是笔划排列的相对位置．在笔划结构不变的情况下，笔划的位

1988 年 9 月 20 日收到修改稿．

置、长短、粗细、走向等的变化并不影响人们对文字的识别。可回溯点跟踪包含配选法的设计，是考虑摸拟人的这种识别功能的。本方法是在汉字识别的包含配选法[3]、双重包含配选法[4]、点跟踪包含配选法的基础上发展起来的。这些方法的共性是都采用了包含配选法原理，其不同之处是辞书结构和配比方式不同。

二、基 本 原 理

可回溯点跟踪包含配选法，从原理上讲是利用了可回溯式的控制策略，和点跟踪包含配选法的原理。以下分别进行介绍。

1. 点跟踪包含配选法原理

设文字图形以 $N \times N$ 点阵取样，则有

$$S(x_i, y_i) = \begin{cases} 0, & \text{空白上,} \\ 1, & \text{笔划上.} \end{cases} \tag{1}$$

其中 $0 \leqslant x_i < N$, $0 \leqslant y_i < N$, $x_i, y_i \in$ 正整数。

定义 1. 标准辞书。用以描述一文字的部分或全部特征的一个结构，该结构作为识别时判定文字所属的依据。

定义 2. 识别字典。在一汉字识别系统中，依照一定规则将全部标准辞书组织成具有某种结构的一个集合。

任何文字识别方法中，都设有识别字典，不同的识别方法主要体现为标准辞书结构和识别算法的不同。点跟踪包含配选法的标准辞书是由特征点组成的一个单支关系树。用关系文法来描述则为

$$G = (D_0, D, S, L), \tag{2}$$

式中 D_0 为配比始点，D 为特征点集合，S 为特征点属性集合，L 为特征点连接属性集合。

定义 3. 特征点集合。

$$D = \{(x_i, y_i) \mid_{i=1,2,\cdots,M, M \in \text{正整数}}\}, \tag{3}$$

式中 M 为特征点总数。x_i, y_i 分别为第 i 个特征点的 x 方向和 y 方向的坐标。D 的元素可以是笔划的端点、折点、歧点、交点等笔划上关键部位的点，也可以是为区别其他字形而取的笔划外关键部位的点。

定义 4. 特征点属性集合。

$$S = \{0, 1 \mid 0 \text{ 为虚点(空白上点)}, 1 \text{ 为实点(笔划上点)}\}. \tag{4}$$

定义 5. 特征点连接属性集合。

$$L = \{L_0, L_1 \mid L_0 \text{ 为虚点连接属性}, L_1 \text{ 为实点连接属性}\} \tag{5}$$

点的连接属性是指 D_i 点与其前一点 D_{i-1} 点的连接关系。点跟踪包含配选法规定相邻两点间的连接必取直线段路径。

规则 1. L_1——实点连接属性，即笔划上的特征点与其前一特征点的连接关系属性。当 D_i 点的属性 $S(D_i) = 1$ 时，L_1 有 3 种连接方式。

$$L_1 = \{L_{11}, L_{12}, L_{10} \mid L_{11} \text{ 全实连接}, L_{12} \text{ 半实连接}, L_{10} \text{ 其它连接}\}. \tag{6}$$

方式 1. L_{11} 全实连接，两点间连线必处于同一笔划上。

方式 2. L_{12} 半实连接，两点间连线必不处于同一笔划上，且该连线不许穿越其它笔划。

方式 3. L_{10} 除以上两种情况外的其他连接。

规则 2. L_0——虚点连接属性，即空白上的特征点与其前一点的连接关系属性。当

$$S(D_i) = 0$$

时，L_0 有 3 种连接方式。

$$L_0 = \{L_{01}, L_{02}, L_{00} | L_{01} \text{全虚连接}, L_{02} \text{半虚连接}, L_{00} \text{其它连接}\}。 \tag{7}$$

方式 4. L_{01} 全虚连接指二点连线必处于空白上。

方式 5. L_{02} 半虚连接必须是笔划上点到笔划外点的连接，连接不允许穿越其它笔划。

方式 6. L_{00} 除以上两种连接外的其它连接。

图 1 为标准辞书关系树的一例，其中每一节点相应于一特征点，为着描述特征的需要，一个特征点在辞书中可以取用两次甚至多次。

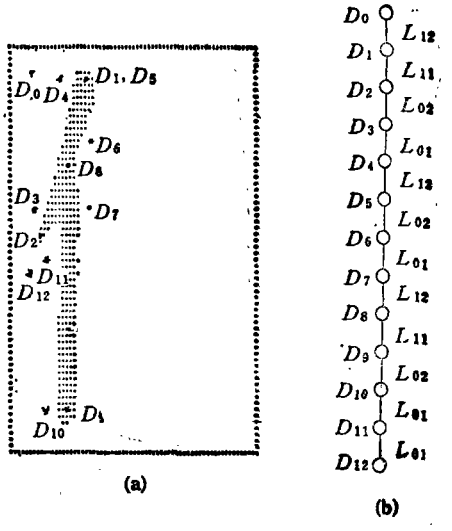


图 1　标准辞书单支关系树例

配比过程即是逐点跟踪配比，验明未知文字是否符合标准辞书的文法关系，若符合则达成匹配，判明文字所属。

2. 回溯策略

汉字特征点间的连接关系是由汉字结构所决定的，它不会因字体不同，字的风格不同而改变。也就是说，笔划位置、长短、粗细、走向等的变化，都不会影响特征点的连接属性，而只会使特征点的位置发生变动。因此，若在一定范围内找寻符合条件的特征点，将单个特征点扩展成多个节点，将单支关系树扩展成多支关系树，就可开辟多条配比路径。当循某一路径达不成匹配时，运用回溯策略，回到前面已达成匹配的某一级节点上，再循新的路径配比。一旦循某一路径达成匹配，则可判明文字的所属。若在所有路径上均不能达成匹配，即转入下一个标准辞书的配比过程。

扩展关系树要符合如下的判断：

$$\text{if } L(D_i) = L_{12},$$
$$\text{then } R\text{: } D_i \rightarrow (D_{i0}, D_{i1}, D_{i2} \cdots\cdots D_{in}). \tag{8}$$

这里 R 表示扩展规则，扩展规则可以有多种，如差二四邻接点、差二八邻接点、差二差四四邻接点等。扩展的点数 n 越大，则允许笔划绝对位置移动的范围也越大，与一辞书配比所需的时间越长。因此，具体采用何种扩展规则需通过实验确定。

举例而言，在连线属性为 L_{12} 的实点上，以差二四邻接点规则进行扩展有

$$R\text{: } \begin{cases} D_{i0} = D(x_i, y_i), \\ D_{i1} = D(x_{i-2}, y_i), \\ D_{i2} = D(x_i, y_{i-2}), \\ D_{i3} = D(x_{i+2}, y_i), \\ D_{i4} = D(x_i, y_{i+2}). \end{cases} \tag{9}$$

这时辞书将扩展为具有 5 个枝的关系树。图 1 给出的单支关系树，此时将扩展为多支关系树

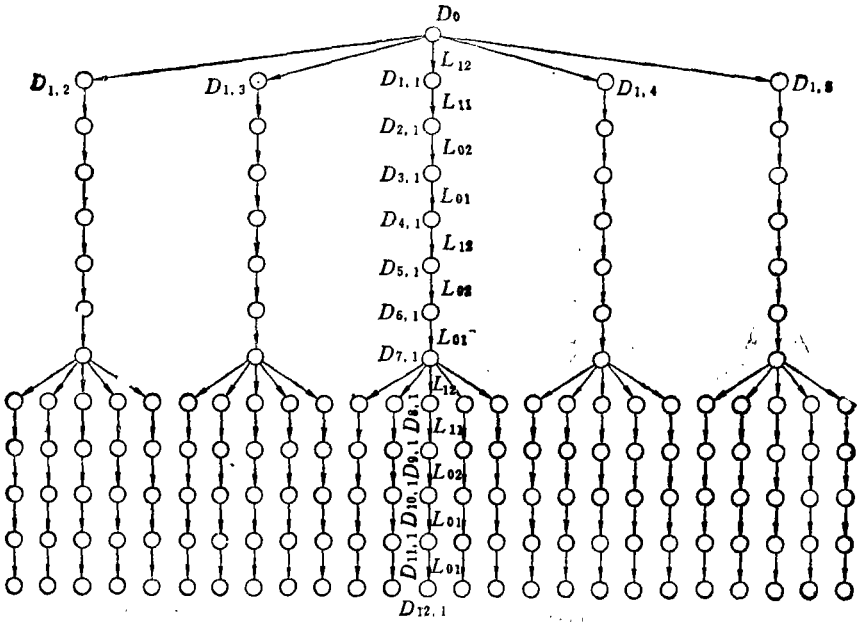


图 2 经扩展的多支关系树

如图 2 所示.

回溯策略，就回归点而言可分为单级回溯和多级回溯. 配比方式可分为宽度优先方式和深度优先方式. 在汉字识别中，对不属于该辞书范围的字形来说，要经过关系树中的各条路径，且每条路径均不能达成匹配；对属于该辞书范围的字形，只要关系树各条路径中有一条能达成匹配，则可判定文字所属. 考虑到上述特点，采用深度优先方式，既符合点跟踪方式，又能提高属于该范畴字形的识别速度. 因此，识别算法中采用总体上深度优先方式. 再考虑到为避免因配比途中反复取用某一特征点而占用较多的时间，在技术上采用单节点宽度优先的办法. 即凡配比到有分支的节点时，按宽度优先处理与该节点相连接的各下一级节点的配比，从中选出并记录符合语法关系的下一级节点，而后循第一个符合条件的节点，再依深度优先原则往下配比，这样做有利于提高识别速度.

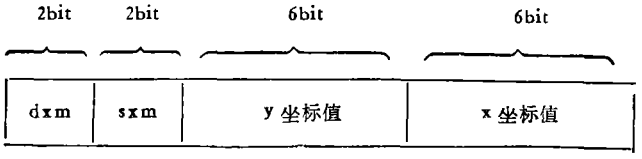
由上述可见，通过扩展关系树、采用回溯策略，使之可以在一定范围内寻找达成匹配的特征点，这相当于在配比过程中标准辞书局部的扭动和整体的浮动，正是由于这种效应，使得本方法对不同的字体及不同风格的文字，有较强的适应能力.

三、标准辞书和识别字典

1. 标准辞书

在标准辞书中，一个特征点的描述符占一个字长 (16bit)，点描述符包括点的属性、点的连接属性、以及点的 x, y 坐标值，其比特分配如图 3.

若一文字图形的辞书由 M 个特征点组成，则该辞书占用 $M+2$ 个字. 辞书的第一个字为该辞书的长度参数，第二个字为该辞书的代码，由第三个字开始依次排列各个特征点的描述符，一个标准辞书结束后，继之为下一个标准辞书，整个辞书文件以 0 作为结束标记.



dxm：点的属性，　sxm：连线的属性.

图 3　点的描述符

可回溯点跟踪包含配选法的辞书，分为分类用标准辞书和识别用标准辞书，它们的不同仅在于其代码值，用 L_c 代表类别码，Z_c 代表国标区位码，则有

$$\text{Code} = \{L_c,\ Z_c \mid 0<L_c<1600,\ Z_c>1600\}. \tag{10}$$

本识别方法，使用上述的标准辞书时，需对辞书做适当的处理．处理的第一步是找出辞书中多次使用的特征点，对这类点只是在第一次出现时，允许在一定范围内扩展，其他次出现只取原来的点，不再作扩展，以 $\text{nod}(i)$ 表示第 i 点的出现状态．处理的第二步是将特征点的绝对坐标值改变为相对坐标．辞书中给出的是

$$D = \{(x_i,\ y_i) \mid i = 1,\ 2\cdots M,\ (0\leqslant x_i,\ y_i)<N;\ (x_i,\ y_i)\in \text{正整数}\}. \tag{11}$$

为扩展特征点的需要，取

$$D' = \{(\Delta x_i,\ \Delta y_i) \mid i = 1,\ 2\cdots M;\ -N<(\Delta x,\ \Delta y)<N;\ (\Delta x,\ \Delta y)\in \text{整数}\}, \tag{12}$$

式中 $\Delta x = x_i - x_{i-1}$，$\Delta y = y_i - y_{i-1}$。

2. 识别字典

可回溯点跟踪包含配选法的识别字典，是依照一定规则，将全部标准辞书组织成的一树状结构．相应于根、节点和叶的位置，为识别标准辞书或分类标准辞书所占据．在识别字典中，凡出现概率大的字，一般排在靠近起始根的部位．因为，这样的字对识别率与识别速度有较大的影响[5]．

四、识 别 算 法

可回溯点跟踪包含配选法，在配比过程中，全实连接和全虚连接采用同域包含算法．半实连接与半虚连接采用异域跳变算法．以下分别予以说明．

1. 同域包含算法

在原理中指出，标准辞书中相邻两点间必取直线段连接．以 T_i 表示 D_{i-1} 到 D_i 点的配比路径，该路径上总的配比点数为 ZDS_i，总实点数为 BS_i，总虚点数为 BK_i，则有

$$\begin{cases} ZDS_i = \sum\limits_{i\in T_i} S(x_i,\ y_i) + \sum\limits_{i\in T_i} \overline{S(x_i,\ y_i)}, \\ BS_i = \sum\limits_{i\in T_i} S(x_i,\ y_i), \\ BK_i = \sum\limits_{i\in T_i} \overline{S(x_i,\ y_i)}. \end{cases} \tag{13}$$

其中 $\overline{S(x_i,\ y_i)} = 1 - S(x_i,\ y_i)$，函数 $S(x_i,\ y_i)$ 的定义参见公式 (1)。

定义 6.

(1) 笔划上的包含度

$$\mathrm{BHD}S_i = \frac{BS_i}{ZDS_i}, \tag{14}$$

(2) 空白上的包含度

$$\mathrm{BHD}K_i = \frac{BK_i}{ZDS_i}, \tag{15}$$

显然有

$$0 \leqslant \mathrm{BHD}\,S_i \leqslant 1, \quad 0 \leqslant \mathrm{BHD}\,K_i \leqslant 1, \tag{16}$$

当且仅当 (17) 式

$$((\mathrm{BHD}S_i = 1) \wedge (L(D_i) = L_{11})) \vee ((\mathrm{BHD}K_i = 1) \wedge (L(D_i) = L_{01})) \tag{17}$$

为真时,即达成在同域的匹配.

2. 异域跳变算法

在配比路径 T_i 上,若 (18) 式为真,

$$(S(x_i, y_i) = 1) \wedge (s(x_{i-1}, y_{i-1}) = 0), \tag{18}$$

说明在异域的配比过程中,发生了一次笔划外到笔划上的跳变.

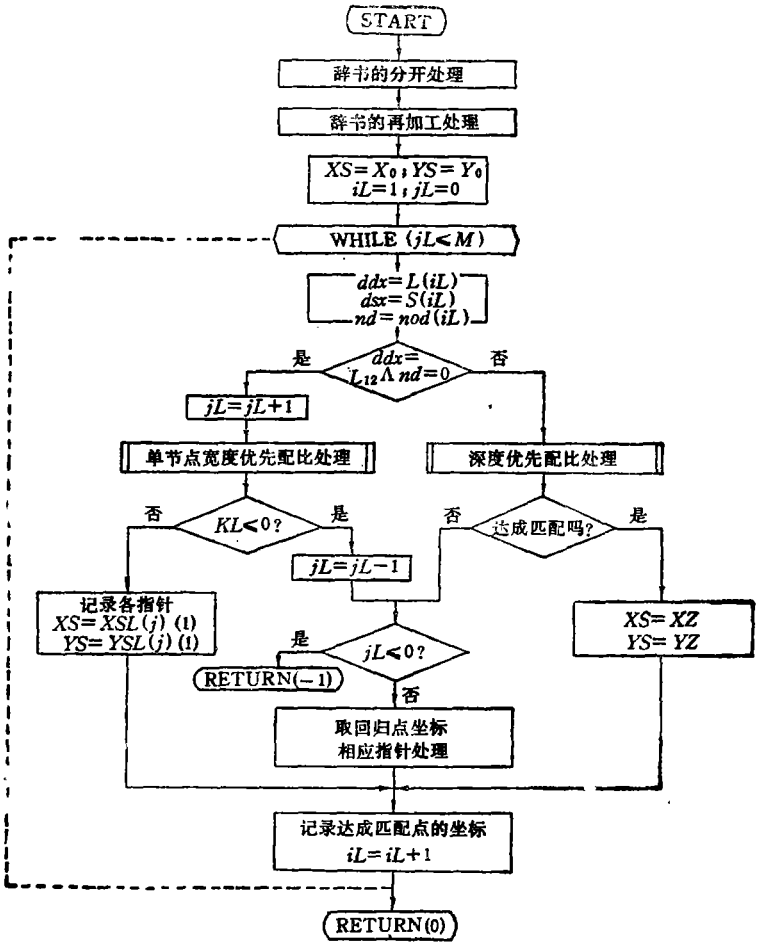


图 4 识别算法流程图

定义 7. 跳变次数

令

$$JP_j = \begin{cases} 0, & (18)\text{ 式不为真时,} \\ 1, & (18)\text{ 式为真时,} \end{cases} \tag{19}$$

我们将下式

$$\mathrm{TB}D_i = \sum_{j \in T_i} JP_j \tag{20}$$

称为跳变次数．当且仅当下式

$$((\mathrm{TB}D_i = 1) \wedge (L(D_i) = L_{12}) \vee (\mathrm{TB}D_i = 0) \wedge (L(D_i) = L_{02})) \wedge (BS_i \neq 0 \wedge BK_i \neq 0) \tag{21}$$

为真时，即达成在异域的匹配．

3. 识别算法流程图

识别算法流程图如图 4 所示．图中各变量的意义如下：

XS——配比始点 x 坐标，YS——配比始点 Y 坐标，XZ——配比终点 X 坐标，YZ——配比终点 Y 坐标，iL——配比指针，jL——回溯点指针，KL——单节点宽度优先处理达成匹配的点数；ddx——连线属性，dsx——点的属性，nd——点的出现状态．

此流程图用来做配比处理，判明未知文字是否满足标准辞书中给出的文法关系．若满足即达成匹配，则 RETURN(0)；若不能达成匹配则 RETURN(−1)．

五、印刷体汉字识别实验

1. 实验系统

实验系统如图 5 所示，文字图形输入部分是 C1000 光导摄象机系统，该系统可将图形转化为二值数字图象．

识别软件是在 HP-9000 小型计算机系统上开发的，该系统采用 unix 操作系统，多终端、多用户．

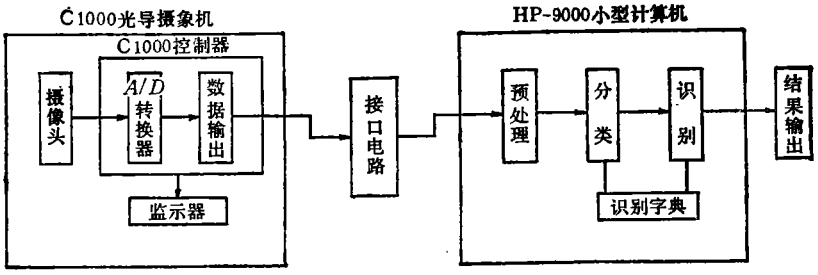


图 5　实验系统框图

由于识别实验中采用的字样来自市场上购入的字典、辞海等实际印刷物品，文字的大小排版的位置迥然各异．本实验的目的在于研究识别方法的有效性．因此，实验中采用单字输入方式．

2. 识别实验

实验开始，首先将我们所能收集到的 4 种不同字体，10 种不同样本的国标一、二级印刷汉

字 57798 个输入计算机，建立了识别用印刷体汉字库. 字库组成如表 1；10 种不同样本的字库字样如图 6 所示.

表 1 字库的组成

样本名称	字 体	字 号	排列顺序	字数
新华字典	老宋体	3 号	拼音	6895
辞海	老宋体	2 号	偏旁部首	6244
印刷字	宋体	头号	偏旁部首	5397
中学生字典	黑体	4 号	拼音	3365
国标集、部首索引	宋体	小 4 号	偏旁部首	7253
国标集、拼音索引	宋体	小 4 号	拼音	7705
国标集、国标索引	宋体	3 号	国标码	6763
现代汉语辞典	老宋体	3 号	拼音	7127
汉英词典	扁宋体	2 号	拼音	6079
杭州 52 所印刷字	黑体	3 号	偏旁部首	4274

图 6 不同样本的字库字样

建立上述的字库，是作为实验系统学习的字库. 标准辞书的制做，是利用该字库采用人机结合的方式完成的.

在完成标准辞书，识别字典的制做之后，对印刷汉字进行了识别实验，其结果如下：

(1) 对经过学习的字库(不同的 4 种体，57798 个汉字)识别率大于 99.87%.

(2) 对未经学习的样张(宋体，3号，国标一、二级字)平均识别率大于 98%；对未经学习的样张(黑体 3 号，国标一级字)平均识别率大于 95%.

(3) 对实际印刷物品的实时抽测，平均识别率为 96.7%.

(4) 对本识别方法适应能力的测试 被识别汉字在识别窗口范围内，放大、缩小、移位、笔划粗细变化、左右倾斜等，都体现较强的适应能力. 图 7 给出了经各种变化后仍能正确识别的例子.

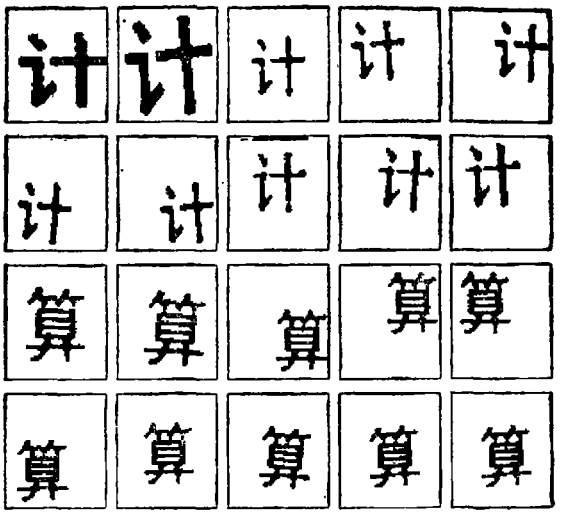


图 7　识别适应性测试例

(5) 识别速度　在多终端多用户的条件下平均为 3.2s/字.

(6) 软件容量　用C语言开发识别软件，编译成可执行程序小于 35 KB；识别字典小于 900 KB.

六、结束语

汉字识别的可回溯点跟踪包含配选法，是为适应我国字域庞大、字型复杂、相似字多的特点而设计的. 经过在 C1000 光导摄象机与 HP-9000 小型机构成的图象输入、处理系统上，完成识别实验及测得各种实验数据，说明这种方法有较强的生命力. (1)它适应多字体的汉字识别，实验中的黑体、宋体、扁宋体，字型差别较大，这种方法能较好的适应；(2) 这种识别方法具有较强的抗位移、抗扭转、抗干扰能力，能识别实际印刷物品上的汉字；(3) 这种识别算法简单，没用超越函数，便于实现并行处理；(4) 识别软件内存开销少，便于在微机上移植.

正如上面已经指出的，该项研究是为了验证本识别算法的有效性. 为将实验成果推向实用尚需作大量的具体工作.

在研究中曾得到张新中副教授、刘源和张淞芝研究员的有益帮助，在此深表感谢.

参考文献

[1] 森、坂井，日经エレクトロクス，日本，1977，10:101.
[2] 目黑、梅田，信学論 (D) J67-D0，日本，1984，8:908—915.
[3] 郭宝兰、松本欣二，通信学报，4(1983)，4:52—58.
[4] モシエ、郭宝兰、松本欣二，信学論 (D) J65-D，日本，1982，8:1011—1017.
[5] 郭宝兰、张彩录，通信学报，7(1986)，5:57—62.