中 国 科 学
1990 年 1 月　　A 辑　　第 1 期

印刷体汉字文本的微型计算机自动识别

张炘中　阎昌德　刘秀英　王　玉
（北京信息工程学院中文信息处理研究中心，北京 100011）

摘　　要

本文提出一个微机印刷体汉字文本识别系统．该系统是基于汉字识别特征点方法实现的，是一个包括版面分析、文本识别和编改处理的完整的系统．该系统在实用中，经过约 200 万字的识别得出：对中等印刷质量的书刊、文件，在纯软件条件下，达到 20 字/s 的识别速度（20MC386 微机）和 95% 以上的识别率．

关键词：印刷体汉字识别，汉字特征点，微机汉字识别系统，版面分析

一、前　　言

近年来，随着办公自动化和新一代计算机人机智能接口的需求，汉字识别的研究已经从纯方法研究开始走向实用化、商品化[1]．

特征及其抽取是模式识别的核心．抽取什么特征，用什么方法抽取，基本决定了系统可能达到的性能指标．对汉字字形分析研究，得出表达汉字结构本质的特征及抽取方法，是汉字识别中的一个十分重要的问题．

以前发表的汉字特征及抽取方法，从最早采用的轮廓投影 Fourier 变换发展到复杂指数、四边码，一直到当前常用的粗外围、粗网格、笔划密度、笔划方向、网格单元等[2-4]，这些对二、三千日本汉字识别有用的特征，对七千多中国汉字则缺乏有效性．要在微机上实现比日本汉字多一倍的中国汉字识别，而且不用专用硬件就能达到实用速度，必须研究新的特征．

新的特征怎样选择呢？分析一下以前统计法识别汉字所采用的特征的演变可以得到启发．最早采用的特征，如轮廓投影 Fourier 变换，是把汉字看成一般的二维图形来抽取特征，很少考虑汉字本身的结构特点．发展到四边码、粗外围、笔划密度、网格单元就注意到汉字结构特点．汉字特征的选择原则上有三条思路：(1)把汉字看为一般的二维图形，用通常选取图形特征的方法来选择；(2)只考虑几千或几万个汉字的特殊图形的区别来选择结构特征；(3)在汉字结构信息中再选取关键的稳定部分．我们认为，从第三条思路来考虑最有效．更多、更准确地考虑到汉字结构特点，力求从这些结构信息中找出稳定而关键、又便于提取的特征来，是不同于以前的、能解决我国七千多汉字分类、识别的特征选择的道路．据此，我们在 1987 年提出了一种汉字识别的特征点法[5]，本文先简述该方法，再详解一个基于该方法的印刷体汉字文

本文 1988 年 9 月 30 日收到，1989 年 7 月 24 日收到修改稿．

本识别系统.

二、汉字识别的特征点法

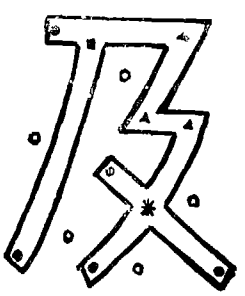
汉字基本上是一种直线型文字,一个二值化点阵汉字信息,绝大部分集中在汉字骨架上,而汉字骨架信息又大多集中在若干特征点(称为笔划特征点)上. 一旦确定笔划特征点,根据连接规则,汉字笔划以及结构形状即可确定. 一个汉字的背景部分,也包含了区别其它汉字的丰富信息. 在背景中选取若干点(称为关键背景点),可有效地区分其它汉字. 所以,汉字笔划上和背景中的若干关键点是汉字本质的字形结构特征.

定义 1. 汉字笔划特征点 t_i 为端点 D、折点 Z、歧点 Q、交点 J 的集合, $t_i=\{D, Z, Q, J\}$.

定义 2. 关键背景点 B 是在有相似和包含关系的 t_i 的汉字集中,能区别开各文字的背景点.

定义 3. 汉字特征点 $t=\{D, Z, Q, J, B\}$.

汉字特征点示于图 1. 这些特征点集中了主要的汉字结构信息. 端、折点决定了一个汉字的笔划位置和形状; 歧、交点决定了不同笔划间的相互连接关系. 关键背景点又弥补了区别相似笔划特征点汉字的不足. 各种印刷体汉字(宋、仿宋、楷、黑等),同一汉字的特征点很少变动,对折、交点和关键背景点则不变. 因而, 用汉字特征点识别汉字,从原理上就能很好识别多体印刷汉字,甚至可以识别手写体汉字, 把印刷体汉字和手写规整汉字识别方法统于一个. 每个汉字的特征点大约是该二值化点阵图形容量的几十分之一,用它来识别汉字,可以减少存贮量, 提高识别速度, 使系统能在微机上实现. 汉字特征点反映了浓缩的汉字结构特征,和统计特征相比,汉字中非结构信息 (如笔划粗细、字形位置变动、少量旋转等)的不稳定性,对特征点的数量和相对位置影响不大. 所以,用特征点来识别汉字,可以增加抗干扰能力,提高实用识别率.



● 端点 ▲ 折点

■ 歧点 米 交点

O 关键背景点

图 1 汉字特征点

用特征点来识别汉字,国外常规的做法或是细化后检出汉字笔划特征点, 连接成线段、子笔划、笔划,再抽取方向、长度等特征来识别;或是根据背景区提取笔划方向、段数等特征来识别. 我们则把笔划特征点和关键背景点两者结合起来, 而且直接根据特征点本身的信息 (类型、数目、相对位置等)来识别.

设 T 为汉字特征表达式, t_k 是汉字特征点, K 是特征点总数, s_k 是特征点类型, x_k, y_k 是特征点在汉字点阵中的坐标, $\{p_k\}$ 是特征点其它属性的集合. 则有

$$\begin{aligned} T &= \{t_k\}, \quad k=1, 2, \cdots, K, \\ t_k &= (s_k, x_k, y_k, \{p_k\}). \end{aligned} \tag{1}$$

汉字笔划特征点抽取的方法是: 通过四方向量化将汉字分割成方向线段, 再把方向线段划分为三个区——起点区、内点区和终点区,给每个区域以特定的象素值, 再把所有方向线段累加起来,根据累加象素值 $x(i, j)$ 确定每个黑点的分枝数 $b(i, j)$. $b(i, j)=1, 2, 3, 4$ 时分

别为 D, Z, Q, J 点.

$$b(i,j)=\begin{cases}\sum_{k=0}^{4} d_k(i,j)+2\sum_{k=4}^{7} d_k(i,j)+\sum_{k=8}^{11} d_k(i,j), \\ \quad \text{当 } \sum_{k=0}^{3} d_k(i,j)+\sum_{k=8}^{11} d_k(i,j)>0 \text{ 或 } \sum_{k=4}^{7} d_k(i,j)>1 \text{ 时}, \\ 0, \quad \text{其它情况}.\end{cases} \tag{2}$$

式中 $d_k(i,j)\in\{0,1\}$, 由(3)式确定. $k=l\sim l+3$, $l\in\{0,4,8\}$, 分别是起点区、内点区和终点区内方向线段的四个量化方向.

$$x(i,j)=\sum_{k=0}^{11} d_k(i,j)\cdot 2^k. \tag{3}$$

三、识 别 系 统

1. 系统构成

用图文扫描器扫描整页版面,对标题、正文、图象版面进行分析,用软件识别汉字, 识别结果可显示、编改、打印、存盘,也可用语音合成输出. 系统构成见图 2.

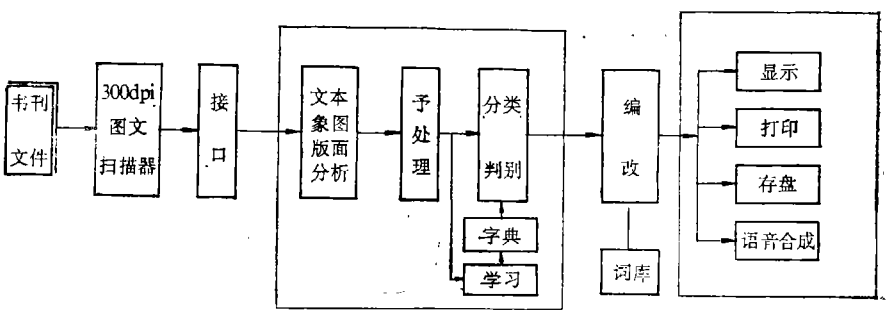


图 2　印刷体汉字文本普通微型计算机识别系统方框图

识别软件包括扫描模块、二值化模块、版面分析模块、行字切分模块、规范化模块、粗分类模块、细分判别模块、词编改模块、字典制作模块、语音合成输出模块、评价模块和学习模块.

2. 主要技术指标

(1) 识别字数: 3755—6763.

(2) 识别字体: 宋、仿宋、报宋、黑、楷、打印、老繁宋体等.

(3) 识别字号与页面: 3,4,5,6 号, A_4—B_6.

(4) 识别速度: 20 字/s (20MHz COMPAQ 386 型微型计算机), 11 字/s (16 MHz Intel 286 为 CPU 的微型计算机、样张),6—8 字/s (6MHz IBMPC/AT).

(5) 识别率:99.2—99.65%(样张),95.2%(三本书,50 万字),97.9%(20 份文件, 6 万字). 用户实际使用时,输入 100 多万字,平均识别率>95%.

(6) 有汉字文本、标题、图象版面分析切分功能,可把分栏的文本连接成一个文件.

(7) 有语音合成输出功能,可用语音校对文章.

(8) 识别结果可在 CCDOS (Chinese Character Disk Operating System) 下用词上、下

文修改,结果作为文本文件存盘.

(9) 有学习功能,对其它字体(如隶书)经过学习后,系统也可识别.

四、文本图象版面分析[6]

1. 文本图象版面构成

文本图象版面由标题域、文本域和附属域组成,如图 3 所示. 这些域用域分离器区分开.

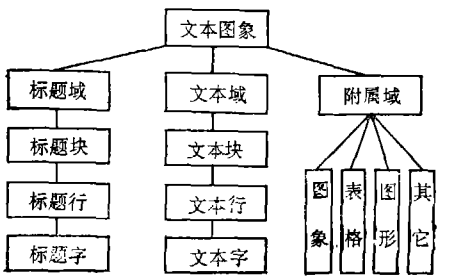


图 3 文本图象的构成

文本图象 F 可表示为

$$F(i,j)=\begin{cases}0, & \text{白象素},\\ 1, & \text{黑象素},\end{cases} 1\leqslant i\leqslant N, 1\leqslant j\leqslant M,$$

式中 M, N 是 F 的行数和每行象素的数目. 文本图象中任意一块的位置可以用一四元组 $A(x_s, x_e, y_s, y_e)$ 表示,其中 x_s, x_e, y_s, y_e 分别为文本块的 x 和 y 的起始和终结坐标.

2. 附属域与其它域的分离

采用附属域优先提取的算法. 附属域投影映象临界值

$$\rho = H_{\max}(\text{mm})\frac{R(\text{dpi})}{25.4\ (\text{mm}\cdot\text{dpi})}, \tag{4}$$

式中 R 为扫描器的分辨率, $H_{\max}$ 为文字最大高度.

附属域判定规则 1. 对 $A(x_s, x_e, y_s, y_e)$, 若块中文本图象在 x 方向的投影映象满足

$$\sum_{i=x_s}^{x_e} E(i, y_s+j)\geqslant \text{MINLINESUM},\ (0\leqslant j\leqslant y_s-y_e), \tag{5}$$

则该块中可能存在附属域. 其中 MINLINESUM 为常数, $E(i,j)=\sum_{k=0}^{7} e(k)$, $e(k)$ 为第 j 行第 i 个字节的第 k 位的值, $e(k)=0$ 或 1.

附属域判定规则 2. 若在可能存在附属域的块 $A[x_s, x_e, y_s, y_e]$ 中存在一子块 $A_1[x'_s, x'_e, y_s, y_e]$, 该子块中的文本图象在 y 方向的投影映象满足

$$\sum_{j=y_e}^{y_s} E(x'_s+i, j)\geqslant \text{MINCOLSUM}\ (0\leqslant i\leqslant x'_e-x'_s), \tag{6}$$

其中 $x_s\leqslant x'_s$, $x'_e\leqslant x_e$, MINCOLSUM 为常数.

又, 在该块中存在另一子块 $A_2[x'_s, x'_e, y'_s, y'_e]$. 其中文本图象在 x 方向的投影映象满足

$$\sum_{i=x'_s}^{x'_e} E(i, y'_s+j)\geqslant \text{MINLINSUM}\quad (0\leqslant j\leqslant y'_e-y'_s), \tag{7}$$

且 $y'_e-y'_s>\rho$, 其中 $y'_s\geqslant y_s$, $y'_e\leqslant y_e$, 则块 $A_2[x'_s, x'_e, y'_s, y'_e]$ 中存在附属域,该附属域的位置为 $[x'_s, x'_e, y'_s, y'_e]$.

3. 上、下文有关文本块的连接

在进行连接前,先要根据 ρ 进行文本行的合并和切分.

文本块连接规则．设有二列属于同一标题的文本块 A_i 和 G_j，

$$A_i = [x_{s_i}, x_{e_i}, y_{s_i}, y_{e_i}], \quad i = 1, 2, \cdots m,$$
$$G_j = [x_{s'_j}, x_{e'_j}, y_{s'_j}, y_{e'_j}], \quad j = 1, 2, \cdots n,$$

设

$$x_{s_\cdot} = \min_{1 \leqslant i \leqslant m} \{x_{s_i}\}, \quad x_{s_\cdot} = \max_{1 \leqslant i \leqslant m} \{x_{e_i}\},$$
$$x_{s_\cdot} = \min_{1 \leqslant j \leqslant n} \{x^{s'}_j\}, \quad x_{s_\cdot} = \max_{1 \leqslant j \leqslant n} \{x^{e'}_j\}.$$

(1) 如果 $x_{s_\cdot} < x_{s_\cdot}$，则 A_i 列块都位于 G_j 列块的左侧，连接时，则位于 G_j 列块的上方．

(2) 如果 $x_{s_\cdot} < x_{s_\cdot}$，则 A_i 列块都位于 G_j 列块的右侧，连接时，位于 G_j 列块的下方．

五、识别预处理

预处理包括二值化、行字切分和规范化．

1. 二值化

设汉字文本图象象素 (i, j) 的灰度值为 $g(i, j)$，则象素 (i, j) 二值化为

$$c(i, j) = \begin{cases} 0, & \text{白点}, \quad g(i, j) < TH, \\ 1, & \text{黑点}, \quad g(i, j) \geqslant TH, \end{cases} \tag{8}$$

式中 $TH = K_1(K_2 A_v + K_3)$，A_v 是 (i, j) 周围某一区域内的灰度平均值，K_1，K_2，K_3 是常数．

2. 行切分

若象素行 j 满足(9)式，则 j 为行上界 j_A，满足(10)式为行下界 j_B．

$$\left(\sum_{i=1}^{N} F(i, j) \geqslant p_1\right) \wedge \left(\sum_{i=1}^{N} F(i, j+1) \geqslant p_2\right) \wedge \cdots \wedge \left(\sum_{i=1}^{N} F(i, j+k) \geqslant p_{k+1}\right), \tag{9}$$

$$\left(\sum_{i=1}^{N} F(i, j) \leqslant q_1\right) \wedge \left(\sum_{i=1}^{N} F(i, j+1) \leqslant q_2\right) \wedge \cdots \wedge \left(\sum_{i=1}^{N} F(i, j+k) \leqslant q_{l+1}\right), \tag{10}$$

式中 $p_i (i = 1 \sim k+1)$，k，l 是大于零的常数，$q_j (j = 1 \sim l+1)$ 是大于等于零的常数．

3. 字切分

在实际汉字文本中，字切分要考虑以下情况：

- 汉字
  - 左右分离字：川、非、何，
  - 整体字
    - 正常字：国、文，
    - 窄字：目、卜，
- 标点
  - 长标点：!?()
  - 短标点：，。、
  - 宽标点：——……
- 数字：0，1，…，9，
- 字母
  - 大写：A，B，…，Z。
  - 小写：a，b，…，z，
- 干扰
  - 字边界上存在污点，
  - 字与字连接，
  - 字断笔．

近年来，在各种各样的货物运输中，已广泛地采用统一规格的集装箱来装运零散的或外形、尺寸不规则的货物，这种办法既方便装卸，又提高了效率。在军事空运中，也采用集装箱装运小件、散装军用物资和零散装备，这样，就可不必受运输飞机型号、货舱规格等条件的限制，因而提高了空运的机动能力。若空运一些体大而外形复杂的器物，例如大型火箭、飞船和飞机的大部件，各种大型的建设器材，设备，以及巨型整体预制构件等，就必须研制专用的特殊运输机。

图10-16示出的就是一种特殊运输机。这是美国的一种机身很大的巨型运输机。它由现有的运输机改装而成，把机舱的大部改装包围容器，以便于装运大型火箭、宇宙飞船等。相对于如此胖大的机身来说，装置四台涡轮螺桨发动机能机翼就显得短小了。它的机身头部很钝，不象一般飞机那样尖

图4 行、字切分实例

字切分大致分为两个步骤，首先是求出一行中各个分离部件的左、右边界，这可以用与行切割求行上、下界相类似的算法实现．其次是根据汉字的平均宽度将几个部件合并成一个字或将一个部件切分成几个字．

(1) 根据汉字文本每行的上下界 i_A, i_B，估算每行的字宽 W，$W = i_B - i_A + 1$．

(2) 左右分离字的合并和交连字的分割．

在分析了汉字本身的特点和排版先验知识的基础上，设计了一个自动机完成文字部件的合并和分割．

$$M = (s, \Sigma, f, s_0, z), \tag{11}$$

式中，s 是有限状态集，$s = \{s_0, s_1, \cdots s_n\}$；$\Sigma$ 是字母表，表示几种部件，如空部件 ε，整字宽部件 I_1，半字宽整字高部件 I_2 等；f 是产生式，如 $f(s_0, I_1) = z$，$f(s_0, I_2) = s_1$，$f(s_1, I_2) = z$ 等；z 是终态，在该状态，经过合并或分割，切分出单字．

4. 规范化

按重心分块线性变换将字形规范化，把不同大小的输入文字变换成同一大小尺寸．一个行字切分的实例见图 4．

六、匹 配 判 别

采用自上而下浮动匹配判别．

1. 判别被识候补文字

(1) 计算输入文字点阵 $c(i, j)$ 和字典特征点 t_k 的距离 D_k．

$$D_k = \frac{1}{L}\left\{\min_{\substack{(i,j=1,2,\cdots,N)\\(c(i,j)=p)}} [(i - x_k)^2 + (j - y_k)^2]^{\frac{1}{2}}\right\}, \tag{12}$$

式中 $p = \begin{cases} 1: & \text{笔划特征点} \\ 0: & \text{关键背景点} \end{cases}$，$L$ 是常数．

(2) 计算输入文字 $c(i, j)$ 和字典汉字特征点的一致度 C．

$$C=\sum_{k=1}^{K} W_k(1-D_k)\cdot\alpha(D_k)/K, \tag{13}$$

式中 $\alpha(D_k)=\begin{cases}0, & \text{当 } D_k>1\\ 1, & \text{当 } D_k\leqslant 1,\end{cases}$ W_k 为对不同特征点的加权系数. 当 $C>\varepsilon(0\leqslant\varepsilon\leqslant 1)$，则判定输入文字为被识候补文字.

2. 判别被识文字

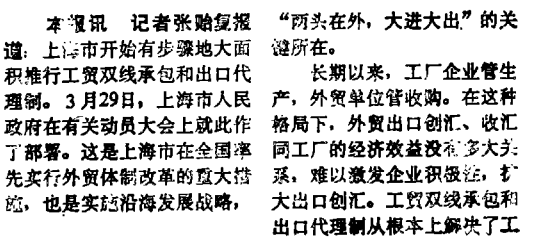
产物，而精神却只是物质的最高产物。这自然是纯粹的唯物主义。但是费尔巴哈到这里就突然停止不前了。他不能克服通常的哲学偏见，即不反对事情本质而反对唯物主义这个名词的偏见。他说：

"在我看来，唯物主义是人类本质和人类知识的大厦的基础；但是，我认为它不是象生理学家、狭义的自然科学家如摩莱肖特所认为的那样，不是象他们从他们的观点和专业出发所必然主张的那样，即不是大厦本身。向后退时，我同唯物主义者完全一致；但是往前进时就不一致了。"14

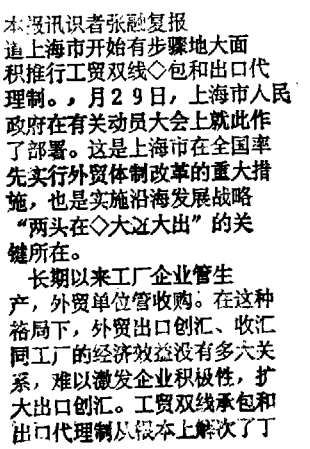
产物，而精神却只是物质的最高产物。这自然是纯粹的唯物主义。但是费尔巴哈到这里就突然停止不前了。他不能克服通常的哲学偏见，即不反对事情本质而反对唯物主义这个名词的偏见。他说：

"在我看来，唯物主义是人类本质和人类知识的大厦的基础；但是，我认为它不是象生理学家、狭义的自然科学家如摩莱肖特所认为的那样，不是象他们从他们的观点和专业出发所必然主张的那样，即不是大厦本身。向后退时，我同唯物主义者完全一致；但是往前进时就不一致了。"一

(a)


本报讯　记者张贻复报道：上海市开始有步骤地大面积推行工贸双线承包和出口代理制。3月29日，上海市人民政府在有关动员大会上就此作了部署。这是上海市在全国率先实行外贸体制改革的重大措施，也是实施沿海发展战略，"两头在外，大进大出"的关键所在。

长期以来，工厂企业管生产，外贸单位管收购。在这种格局下，外贸出口创汇、收汇同工厂的经济效益没有多大关系，难以激发企业积极性，扩大出口创汇。工贸双线承包和出口代理制从根本上解决了工


本报讯记者张融复报
道上海市开始有步骤地大面
积推行工贸双线◇包和出口代
理制。，月29日，上海市人民
政府在有关动员大会上就此作
了部署。这是上海市在全国率
先实行外贸体制改革的重大措
施，也是实施沿海发展战略
"两头在◇大进大出"的关
键所在。
长期以来工厂企业管生
产，外贸单位管收购。在这种
格局下，外贸出口创汇、收汇
同工厂的经济效益没有多大关
系，难以激发企业积极性，扩
大出口创汇。工贸双线承包和
出口代理制从根本上解决了工

(b)

图5　印刷体汉字文本识别实例

(1) 当 $C \geqslant \beta$ (β 取 0.9)且只有一个候补文字,则判定为被识文字, 否则为(2).

(2) 据各候补文字的 J, Z, Q, D, B 依此和输入文字图形浮动匹配判别.

设 N_D, N_Z, N_Q, N_J, N_B 分别为候补文字的 D, Z, Q, J, B 点数, S_D, S_Z, S_Q, S_J, S_B 分别为它们的浮动匹配范围, D_Σ 是浮动匹配距离,则有

$$D_\beta = \min_{(i,j)\in S_\beta} [(i - x_{k\beta})^2 + (j - y_{k\beta})^2]^{\frac{1}{2}} / L_\beta,\ \beta \in (D, Z, Q, J, B),\ L_\beta \text{ 是常数.}$$

$$D_\Sigma = \sum_{i=1}^{N_D} D_{D_i} + \sum_{j=1}^{N_Z} D_{Z_j} + \sum_{k=1}^{N_Q} D_{Q_k} + \sum_{l=1}^{N_J} D_{J_l} + \sum_{m=1}^{N_B} D_{B_m}, \tag{14}$$

当
$$\begin{cases} D_{\Sigma_i} = \min(D_{\Sigma_\alpha}) < \rho_1, & \alpha = 1, 2, \cdots p(\text{候补文字数}), \\ D_{\Sigma_i} < D_{\Sigma_\alpha} - \rho_2, & \alpha \neq i,\ \rho_1, \rho_2: \text{常数}. \end{cases}$$

则对应 i 的候补文字为识别结果,否则为拒识.

系统识别汉字文本的实例示于图 5. 图 5 (a),(b)为原文本,下部是识别结果. ◇是拒识字.图 5(a) 是包括 5,6 号字的书籍一段;图 5(b) 是小 5 号字报纸一块,将原文本两栏自动连接后识别.

七、结 语

本系统是基于汉字识别特征点法的包括版面分析、文本识别、编改处理的一个完整的系统. 在普通微机上用软件实现了与先进世界水平类似的技术指标, 是一个实用的系统. 本系统已经初步商品化,能用于社会上普遍使用的书刊、文件自动识别输入,将在办公自动化、建立汉语语料库、书刊再版、机器翻译以及汉字图像文本高倍压缩存贮和传输中发挥作用. 今后要扩展应用范围,改善用户界面,提高系统指标.

参 考 文 献

[1] 张炘中,中文信息学报,1(1987),3: 1—7.
[2] Mori, K. & Masuda, I., *Proc. of 5th Intern. Conf. on Pottern Recognition*, 1980, 692—720.
[3] 桑源,日经エレクトロニクス,1981,279: 148—167.
[4] 目黑、梅日,信学論 (D), J67-D, 1984, 8: 908—915.
[5] 张炘中、阎昌德、刘秀英,中文信息学报,1(1987),3: 13—19.
[6] 秋山照雄,信学論 (D), J69-D, 1986, 1187—1195.