第 26卷 第 2期 电 子 科 技 大 学 学 报 Vol. 26 No. 2
1997年 4月 Journal of UEST of China Apr. 1997

关于限制性手写体汉字的一种识别方法

傅 彦*
(电子科技大学计算机系 成都 610054)

【摘要】 成功地提出了关于限制性手写体汉字的一种识别方法,并在 486 DX/66 PC机上进行了模拟实验,取得了一定的效果。对于经预处理后的手写体汉字,采用结构识别法、统计决策法和模糊数学法相结合的一种方法。另外,提出了对相似文字的识别方法。

关 键 词 特征点; 特征向量; 模糊关系
中图分类号 TP391.4

由于我国汉字数目十分庞大,汉字种类极其繁多,汉字字型结构极其繁杂,汉字中存在着很多相似文字,加之人为书写汉字的随意性,到目前为止,尽管有关限制性手写体汉字识别方法取得了很大的进展,但仍然存在着许多尚需进行研究和解决的问题,为此,本文提出了一种识别方法。

1 输入汉字的分离

汉字种类繁多,字体千姿百态,按其结构类型可分为: 单字体、内外型、上下型、左右型。下面就经预处理后的汉字进行讨论。

将汉字根作为不可分割的原子,字根间的构造信息作为连结字根的运算符,汉字可用字根与结构信息组成的字符串来表示。由于运算符是二目运算,所以此字符串可用一个二叉树来表示。

1.1 汉字的结构文法表示

由于一棵二叉树可以用一个文法来表示,因此,可以把一个复杂的模式用类似于语言的结构方法来描述,即用模式的句法结构表示法来描述。此时,模式是用一个句子的形式来表示。

定义一个不可分割的字根和单字体作为基元(原始模式),原子表示其中间元,连结字根间的关系为运算符,则一个汉字可用下面的结构文法来描述[1]

$$G=\langle V_N, V_T, P, S\rangle$$

其中 $V_N=\{$〈原子〉,〈运算符〉$\}$

$V_T=\{$〈字根〉,〈单字体〉$\}$

$S=$〈汉字〉

$P=\{$〈汉字〉::= 〈单字体〉| 〈原子〉〈运算符〉〈原子〉
〈原子〉::= 〈字根〉| 〈原子〉〈运算符〉〈原子〉
〈运算符〉::= $L\text{-}R$| $U\text{-}D$| $I\text{-}B\}$

$L\text{-}R$ 为左右结构,$U\text{-}D$ 为上下结构,$I\text{-}B$ 为内外结构。

① 1995年 12月 9日收稿,1996年 9月 3日修改定稿
* 女 33岁 硕士 讲师

1.2 汉字结构在机器中的表示

由于汉字语法树是一棵二叉树,在计算机中,语法树中的任一结点可用三个域表示为:

Lchild	Data	Rchild

其中　Lchild与 Rchild分别表示节点与左右子节点的链接指针; Data表示节点有子节点时,存放左右两个子节点间关系的运算符的编码,无节点时存放字根 (或单字体)的编码。

1.3 输入汉字的分离原则和分离测试

从汉字本身的结构特征,搜索汉字中的分割线依次是上下扫描、左右扫描、变方向扫描,为此,汉字的分离为: 左右分割、上下分割、内外分割[2]。下面仅以内外分割为例。

由于内外型汉字种类繁多,因此,在扫描时采用变换方向的扫描,即沿着从左向右移动,从上向下扫描。令 $X_1(i)=\sum_{j=1}^{N} f(i,j)$,求出 i_1, i_2,满足如下条件,使得

$$X_1(i_1)=\min_{d_1\leqslant i\leqslant d_2} X_1(i) \qquad X_1(i_2)=\min_{N-d_2\leqslant i\leqslant N-d_1} X_1(i)$$

然后再沿着从上向下移动,从左向右扫描,令 $Y_1(j)=\sum_{i=i_1}^{i_2} f(i,j)$,此时如能找到 j_1、j_2,使得

$$Y_1(j_1)=\min_{d_1\leqslant j\leqslant d_2} Y_1(j)=0 \qquad Y_1(j_2)=\min_{N-d_2\leqslant j\leqslant N-d_1} Y_1(j)=0$$

如 j_1、j_2存在,则可判断为能进行内外分割,此时,i_1、i_2、j_1、j_2为其对应的分割线。

1.4 语法树的形成和汉字组装

汉字语法树是在字根分离和字根识别过程中建立的,字根的分离搜索出结构类型,而字根的识别得到字根的编码信息。由于汉字具有层次嵌套的形式,为此,在进行字根分离时,可采用递归手段。其分离结果可由二叉树表示,该二叉树的枝节点即为子模式的连结枢带,叶子为其子模式,将叶子通过枝节点连结起来,则为原模式。为此,其汉字组装就是遍历语法树,在设计时采用中序遍历。

2 标准模式的描述

标准模式实际上是在识别过程中所使用的样本,其建立是汉字识别中至关重要的一环。

对于任意一个字根,可按如下方式描述: 用向量表示模式的各笔划的特征并存入计算机中,每一条笔划以其始点、终点、屈折点、交点来描述,同时给每一个特征点附加上方向。

设用向量 $R(k,M_R,N)$表示具有 M_k 条笔划,每条笔划有 N 个等距离的特征点的候补分类 k 的等点近似模式为

$$R(k,M_k,N)=\{(r_{11},r_{12},\cdots,r_{1N}),(r_{21},r_{22},\cdots,r_{2N}),\cdots(r_{M_{k1}},r_{M_{k2}},\cdots,r_{M_kN})\}$$

其中　$r_{mn}=(x_{mn},y_{mn})$表示第 m 条笔划的第 n 个笔点的 x,y 坐标。但以上表示有一个不足,对于笔划长度短的和长的,无法正确而准确的加以表达,为此进行如下修正

$$R(k,M_k,N_{M_k})=\{r_{11},r_{12},\cdots,r_{1N_1}),(r_{21},r_{22},\cdots,r_{2N_2})\cdots(r_{M_k1},r_{M_k2},\cdots r_{M_kN_{M_k}})\}$$

其中　$r_{ij}=(r_{ij},y_{ij},a_{ij},d_{ij},h_{ij})$; x_{ij}, y_{ij}分别是向量 r_{ij}的坐标位置; a_{ij}表示 r_{ij}的几何学形状信息; d_{ij}表示 r_{ij}的方向; h_{ij}表示 r_{ij}的外围构造信息。

3 标准字典的建立和分类

由于遍历语法树仅得到的一些编码信息,因此,须建立汉字点阵和编码相对应的汉字字典。

一个标准字典可按如下方式建立: {(字根 i,特征向量 i,偏码 i)}

当识别字典很大时,为了减少字根的匹配次数,达到高效快速的识别,可采用聚类识别法对字典进行分类。聚类识别法大致分三类:合并聚类法、修正聚类法、模糊聚类法。根据模糊聚类法的基本思想[3],首先建立 X 上的模糊相容关系阵 $R=(a_{ij})$,其中

$$a_{ji}=a_{ij}=\begin{cases}1 & i=j\\ 1-\dfrac{1}{\min(m_i,m_j)^*}\sum_{a=0}^{m_i}\left(\min_{c=0,1,\cdots m_j}\left(\sum_{b=0}^{n}\dfrac{\|r_{ab}-r_{cb}\|}{\|r_{ab}+r_{cb}\|}\right)\right) & i\neq j\end{cases}$$

从离散数学的观点,如要进行分类(划分),必须将 R 改造成模糊等价关系矩阵 Q[4]。

根据定理,设 $X=\{\bar{R}_1,\bar{R}_2,\cdots,\bar{R}_M\}$ 是需要聚类的全体样本集,其中,$\bar{R}_i$ 对应于字根 i 的特征向量,如 R 是 X 集合上的模糊相容关系阵,则 $Q=R^{M-1}$ 是一模糊等价关系阵。其次再根据规定的 λ 值,考察 Q 的截矩阵 Q_λ 的各元素,以此来作聚类。当 λ 值选得愈大($0<\lambda<1$),分成的类别数愈多,当 $\lambda=1$ 时,各样本必会自成一类,而 $\lambda=0$,所有样本则必会归成一类。为此可预先选定初值 λ,再根据得到的分类结果和希望数目调整 λ 使其符合要求。分类后的新字典建立如下:

{字根 i,特征向量 $\bar{R}_i$,偏码 i,链接指针}

4 输入模式描述

4.1 方向模式抽出

对于模式中的任一个黑点 $f(i,j)=1(1\leqslant i\leqslant M,1\leqslant j\leqslant N)$,令 $r=(i,j)$,$P_k(r)$ $(k=1,2,\cdots,8)$ 分别是黑点在其八个方向上的象素数(该方向上连续黑点的数目),八个方向的前进方向可由 r_1~r_8 来反映,下面求 $P_k(r)$。

首先置 $P_k(r)=0$,当 $f(r+r_k)=0$ 时,$P_k(r)=P_k(r)$;当 $f(r+f_k)=1$ 时,$P_k(r)=P_k(r)+1$,反复此过程,直到遇到白点或边界点为止,此时 $P_k(r)$ 为关于 $f(r)$ 在 1~8 方向的象素数,此象素数可确定该黑点基于 Q_k 方向的几何学的距离 $d_k(r)$,即

$$\begin{cases}d_1(r)=\sqrt{2}\,(P_1(r)+P_5(r))\\ d_2(r)=(P_2(r)+P_6(r))\\ d_3(r)=\sqrt{2}\,(P_3(r)+P_7(r))\\ d_4(r)=(P_4(r)+P_8(r))\end{cases}$$

令 $d_k(r)=\max_{1\leqslant m\leqslant 4}\{d_m(r)\}$,则可将原模式变成具有方向的模式,即令 $f(i,j)=k$,其中 k 是 $d_k(r)$ 中的 k。

4.2 特征点的抽出

定义连结数 Nc_4[5]

$$Nc_4=\begin{cases}0 & \text{孤立点}\\ 1 & \text{端点(始,终点)}\\ 2 & \text{内点}\\ 3 & \text{三分枝点}\\ 4 & \text{四分枝点}\end{cases}$$

该 Nc_4 可确定相应的特征点。但由此抽出的特征点并不能准确地反映输入模式,还需作如下改进。

设 $F=\{f_k\}(k=1,2,\cdots,k)$ 表示由上抽出的全部特征点集合,其中,$f_k=(x_k,y_k,s_R,Q_R,\{C_k\},H_R)$。$x_k,y_k$ 为特征点 f_k 的 I,J 坐标;s_k 为 f_k 的种类(端点、分枝点、交点,抽样点、PSI分割点);Q_k 为 f_k 的方向编码值;$\{C_k\}$ 为与 f_k 相连的特征点的番号集合;H_k 为 f_k 的外围构造信息。则 $F=\{f_k\}$

为描述输入模式的特征,每个 f_k 都为多元向量,而 F 也可以看成是由多个特征点 f_k 组成的特征向量。

5　笔划的抽出

5.1　候补点的选择

由于特征点包括其位置、方向、几何学形状等诸种信息,为此,首先对笔划 r_{mn},从 $f_1, f_2, \cdots, f_k$ 中找出凡与 r_{mn} 具有相同几何学形状信息点 $f_{k_1}, f_{k_2}, \cdots, f_{k_i}$ $(k_i < k)$;然后,考虑其夹角 θ_{mn} 和 a_{gl},其中 θ_{mn} 为向量 $\overrightarrow{r_{mn}r_{mn+1}}$ 与 i 轴所组成的角度,a_{gl} 为 $\overrightarrow{f_g f_j}$ $(f \in \{C_g\}, g \in \{k_1, k_2, \cdots, k_i\})$ 与 i 轴成的角度中最小者,若 $|\theta_{mn} - a_{gl}| < \theta_{mn}$ (θ_{mn} 为阈值),则 $\overrightarrow{r_{mn}r_{mn+1}}$ 与 $f_g f_l$ 是同方向,此时将 f_g 作为其候补点,设满足此条件的有 i_j 个,设为 f_d $(d = i_1, i_2, \cdots, i_j, i_j < k_i < k)$;最后,利用 Euclidian 距离求出 D_d^{mn}

$$D_d^{mn} = \| r_{mn} - f_d \| = \sqrt{(x_{mn} - x_d)^2 + (y_{mn} - y_d)^2} \qquad (d = i_1, i_2, \cdots, i_j)$$

将其结果进行排序 $D_{j1}^{mn} \leqslant D_{j2}^{mn} \leqslant \cdots \leqslant D_{j1}^{mn}$ $(j_l = i_j)$,若 j_l 小于规定阈值 L,则 $f_{j1}, f_{j2}, \cdots, f_{jl}$ 都是候补点,否则,$f_{j1}, f_{j2}, \cdots, f_L$ 为其对应的候补点。

5.2　对应点的决定

设 (r_{mn}, r_{mn+1}) 为笔点 r_{mn} 与 r_{mn+1} 形成的笔点对,从 r_{mn} 朝 r_{mn+1} 作向量 $U_{mn} = (r_{mn+1} - r_{mn})$,如 (r_{mn}, r_{mn+1}) 存在候补点对 $(f_{mn}^{+}, f_{mn+1}^{-})$,此 $(f_{mn}^{+}, f_{mn+1}^{-})$ 满足如下条件:首先 f_{mn}^{+}, f_{mn+1}^{-} 是众多笔点对中具有最短路径的笔点对;其次,$r_{mn+1} - r_{mn}$ 与 $f_{mn+1}^{-} - f_{mn}^{+}$ 的类似度 r 最大,即满足 $r \geqslant \theta_R$。其 r 规定为

$$r(U_{mn}, Z_{mn}, l_{mn}) = \begin{cases} k(U_{mn}) \dfrac{(U_{mn}, Z_{mn})}{\| U_{mn} \| \| Z_{mn} \|} & \text{如 } 1 - \dfrac{l_{mn}}{\| Z_{mn} \|} \leqslant W \\ 0 & \text{否则} \end{cases}$$

其中　$Z_{mn} = f_{mn+1}^{-} - f_{mn}^{+}$,$l_{mn}$ 为 f_{mn}^{+} 与 f_{mn+1}^{-} 之间的文字线长度;k 是关于 U_{mn} 的权函数;W 是阈值。

以下将从 $\overrightarrow{r_{mn-1}r_{mn}}$ 与 $\overrightarrow{r_{mn}, r_{mn+1}}$ 来考察 r_{mn} 的对应点:

1) 若 $f_{mn}^{+} = f_{mn}^{-}$,则 $f_{mn} = f_{mn}^{+} = f_{mn}^{-}$ 作为 r_{mn} 的对应点;

2) 若 $f_{mn}^{+} \neq f_{mn}^{-}$,则将 (f_{mn}^{-}, f_{mn}^{+}) 作为其新的对应点的候补点,此时,若 f_{mn}^{-} 和 f_{mn}^{+} 是相连的,则抽出其文字线,并求 $\overrightarrow{r_{mn-1}r_{mn}}$ 和 $\overrightarrow{r_{mn}r_{mn+1}}$ 的类似度 r_-、r_+。若 $r_- = r_+$,则 $f_{mn} = f_{mn}^{+} = f_{mn}^{-}$;若 $r_- \geqslant r_{TH}$ (阈值),$r_- \geqslant r_+$,则表明 $\overrightarrow{r_{mn-1}r_{mn}}$ 与 $\overrightarrow{f_{mn}^{-} f_{mn}^{+}}$ 相接近,此时选 f_{mn}^{+} 作为 r_{mn} 的对应点 (f_{mn}^{-} 为 r_{mn-1} 的对应点)。若 $r_+ \geqslant r_{TH}$, $r_+ \geqslant r_-$,则表明 $\overrightarrow{r_{mn}r_{mn+1}}$ 与 $\overrightarrow{f_{mn}^{-} f_{mn}^{+}}$ 相接近,此时选 f_{mn}^{-} 作为 r_{mn} 的对应点 (f_{mn}^{+} 为 r_{mn+1} 的对应点)。

6　字根的识别

根据字典的构成方式,可按如下方式进行:1) 反复调用对应点决定子程序,直到输入字根与字典中的某个字根比较,其标准字根的全部笔点都能找到唯一对应点对,并记下该字根的编号,如 i,并记 $S = \{i\}$;2) 取出字根 i 的链接指针 j;3) 若 $j = \Lambda$,则转 5),否则转 4);4) 对字根 j 进行如同 1) 的匹配,若匹配成功,则将 j 加入 S 得 $S = \{i, \cdots, j\}$,同时 $j \to i$,转 2),否则将 j 的链接指针送 j,转 3);5) 若无法确定对应点,则打拒识标志结束,否则转 6);6) 若 $|S| = 1$,同时 $\sum_{m=1}^{M}\sum_{n=1}^{N_M} \| r_{mn} - f_{mn} \| \leqslant W$ (阈值),则将字根 i 作为识别结果,返回其对应编码,否则打拒识标志结束。若 $|S| > 1$,记下使 $\sum_{m=1}^{M}\sum_{n=1}^{N_M} \| r_{mn} - f_{mn}^{i} \| \leqslant W$ 的 i,转 7);7) 若 $S = \Lambda$,则打拒识标志,若 $|S| = 1$,则打出对应字根 i 作为识别结果,否则对任意 $i, i' \in S$,考察它们所对应的标准字根,i 和 i' 的标准模式为 $R(i, M, N)$,

$R'(i', M', N')(M \leqslant M')$,先求出 i 和 i' 的 C-笔划,P-笔划,N-笔划。下面考察输入模式与分类 i 和 i' 之间的一致度问题。

设分类 i 的 P-,N-笔划的条数分别为 n_P, n_N,相对于输入模式的 P-,N-笔划为 e_P, e_N,分类 i' 为 $n_{P'}, R_{N'}$,相对于输入模式的为 $e_{P'}, e_{N'}$,则与分类 i, i' 的一致度为

$$d(i) = e_P + (n_N - e_N) \qquad d(i') = e_{P'} + (n_{N'} - e_{N'})$$

若此时 1)$d(i) = d(i)$,则转 8);2) $d(i) > d(i')$,则输入模式为分类 i;3)$d(i) < d(i')$,则输入模式为分类 i'。

8) 考察 $d(U_{mn}, U_{m'n'}) = \sum_{i=0}^{1} \| r_{mn+i} - r_{m'n'+i} \|$,若 $H_{mn} = H_{m'n'}$,则定义 $T(U_{mn}, U_{m'n'})$ 为上式的值并返回 7);否则规定 $d(U_{mn}, U_{m'n'}) = * (\infty)$,并打出拒识标志。7)、8)可对相似文字进行识别,如 7)可对“目”和“日”识别,8)可对“甲”和“由”识别。

7 识别试验及考察

应用 LISP语言在 486DX/66机上进行模拟试验,对手工输入的经预处理后的 300多个汉字字根,建立相应的字典。在字典中,若不进行分类,则识别速度和结果准确率都有不同程度的增加和下降。对于相似文字,如“土”和“士”,通过附加上反映字根的第二特征,即笔划间的位置关系及外围构造信息,即可得到正确的识别结果。对如“日”和“目”等的相似字根,对“目”,在识别时,“日”与“目”的一切笔划都被抽出,同时,d 也满足小于规定的阀值,此时可根据一致度来进一步判断,以提高识别率。另外,对不存在与其大致相似的另外字根的字根,则在字根集中只能找到唯一的候补模式。

字根集的分类可提高识别速度,分类中建立模糊相容矩阵,由此求模糊等价矩阵,同时进行截值为“λ”的聚类。“λ”值是从 1→0(步长为 -0.01),当 λ=1时各字根自成一类,λ=0时全部字根成一类,λ值的最终确定是由分类的数目或分类中元素数目的多少而确定。

参 考 文 献

1 陈尚勤.模式识别理论及应用.成都:成都电讯工程学院出版社,1985

2 Fujimuea O. Structural patterns of chinese characters. New York: 1984

3 陈胎源.模糊数学.武昌:华中工学院出版社,1984

4 Quda R O, Hert P E. Decision in pattern recognition. Washington D C, 1987

5 Fu K S, Pattern recognition theory and application. New york: 1990

A Recognition Method Based Limited-handwritten Chinese Character

Fu Yan

(Dept. of Computers, UEST of China Chengdu 610054)

Abstract A microcomputer-based limited-handwritten chinese character recognition method is proposed in this paper. The simulation test is built on the basis of a 486 DX/66 microcomputer. The method is to be put into effort. The proposed algorithms for preprocessing of limited-hand-written chinese character is a method combining strctural pattern recognition and statistical pat-tern recognition and fuzzy mathematics. A recognition method about analog chinese character is proposed in this paper.

Key words characteristic point; characteristic vector; fuzzy relation

编辑 徐培红