文章编号:1009-038X(2000)04-0400-04

联机手写体汉字识别系统中汉字的输入及其特征的提取

章颖芳, 戴月明
(无锡轻工大学信息与控制工程学院,江苏无锡 214036)

摘 要:汉字识别是模式识别领域最富挑战性、又极具应用前景的研究课题之一,而联机汉字识别是近期需求十分迫切的技术.字量大、字形复杂多变、笔顺没有一定规范、笔划数目变动等多种因素,是联机汉字识别的主要困难.笔划相对易于提取是联机识别的优点.针对联机手写体汉字的特点,提出笔划轨迹点方向量化的方法,提取笔划和计算笔划间的连接关系,形成输入样本字的特征.

关键词:联机手写汉字识别;笔划;笔划元;笔划轨迹点方向;特征提取
中图分类号:TP391.1 **文献标识码**:A

Inputting and Abstracting in Online Chinese Character Recognition System

ZHANG Ying-fang, DAI Yue-ming
(School of Information & Control Engineering, Wuxi University of Light Industry, Wuxi 214036)

Abstract: Chinese Character Recognition (CCR) is known as one of the most challenging and useful technique in pattern recognition. Further more, On-line Chinese Character Recognition (OLCCR) is one of the great demand in many practicable applications. Abundant surveys for this topic have been presented. However, the recognition of handwritten Chinese characters is still difficult mainly due to the large number of categories, the complexities of characters, similarity among different categories, and the wide variability among writers. With regard to the characteristics of on-line CCR, the features including strokes and relations among them were extracted. Frame representation is developed for the description and storage of the on-line CCR's knowledge, which includes constructing and writing rules.

Key words: on-line handwritten Chinese character recognition; strokes; segments; direction of segments; characteristics abstracting

---

联机手写汉字识别(也称"在线识别"、"实时识别")是人工实时地把汉字输入计算机的方法.这种方法在操作上与键盘输入不同,使用者只需在板上按正常方式书写,无需额外的学习和培训,真正做

收稿日期:1999-12-03;修订日期:2000-04-12.
作者简介:章颖芳(1972-),女,浙江缙云人,工学硕士,讲师.

到了“会写汉字就会输入”.这对普及计算机应用，促进办公自动化非常有利.本文就汉字的输入及其特征的提取作一些探讨.

1　电子书写板

典型的联机手写汉字识别原理框图示于图 1，可见，一台 PC 机配上电子书写板和联机识别软件就能构成一个联机识别装置.从这个意义上讲，联机汉字识别的关键有两个：一是电子书写板；二是联机识别算法.电子书写板作为字形输入装置，和笔一起直接与人打交道，它的性能好坏直接影响到书写的速度和识别精度，关系到系统的总体性能.

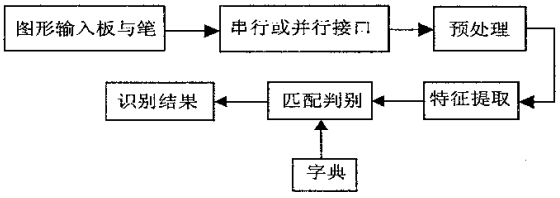


图 1　联机手写汉字识别原理框图

Fig.1　Frame of online hand-written Chinese character recognition

电子书写板又称平板式图形数字化仪，按其工作原理可分为电磁感应型、磁致伸缩型、压敏型、静电耦合型和电阻型等几种.其中电磁感应型的精度较高，性能较好.压敏型价格低廉，易于推广普及.作者在实验中采用压敏电子书写板，型号为 WT-92.当人用笔在板上书写时，它能把笔相对于板的坐标位置检测出来，形成文字笔划轨迹上各点的 X，Y 坐标序列信号，并不断输入到计算机，完成模数转换和采样量化.采样到的数据以开关流模式发往主机.开关流模式特别适用于对连续线段进行数字化处理，所以对汉字笔划输入非常有效.

2　预处理

采样收集的数字信号含有各种干扰和噪声，不能直接用于识别.产生这些干扰的原因主要有书写的随意性、人手抖动、书写的速度变化、书写板的量化噪声、感应噪声等[3].因此在对联机手写汉字进行识别前，必须对输入信号进行预处理，包括字符分割、平滑、去噪声、空间重采样和规范化等.

2.1　字符分割

指区分哪些笔划属于同一个汉字.作者采用的是单字输入方式，因而选用简单的结束码方式分割一个整字.这种方式费点时间，但简单、可靠.

2.2　平滑

人书写时的速度不是均匀的，尤其是在书写“折”、“勾”或者起笔、落笔时，因而笔划坐标与时间 t 的关系不是线性的.为了消除这些影响，在提取特征进行识别之前，必须对采集到的输入信号进行平滑处理.

设采样到的数值化笔划坐标为：$(x_1,y_1),(x_2,y_2),\cdots,(x_t,y_t),(x_n,y_n)$，平滑处理就是[1]

$$x_{st}=f(x_{st-1},x_t)=\Delta S_1 x_{st-1}+\Delta S_2 x_t \quad (1)$$

$$y_{st}=f(y_{st-1},y_t)=\Delta S_1 y_{st-1}+\Delta S_2 y_t \quad (2)$$

其中 (x_t,y_t) 是平滑后笔划在 t 时刻的坐标；(x_{st},y_{st}) 和 (x_{st-1},y_{st-1}) 分别是平滑后前笔划在时刻 t 和 $t-1$ 时的坐标；ΔS_1 和 ΔS_2 是加权系数，且 $\Delta S_1+\Delta S_2=1$.

由(1)和(2)式可知，平滑前各点坐标值和该点平滑前的坐标值以及前一点平滑后的坐标值有关.选择不同的 ΔS_1，ΔS_2 值，得到不同的平滑效果.若 $\Delta S_1=0$，$\Delta S_2=1$ 时，则不进行平滑处理；ΔS_1 越大，平滑以后前后各点相关关系越大.

2.3　去噪声和空间重采样

由于书写板有硬件噪声，且采样精度较高，使得坐标序列中含有大量冗余点，需要做噪声剔除和重采样处理，使采样点较均匀，并压缩数据量[2].

设噪声阈值为 L_M，采样阈值为 L_m，记

$$[(x_{st}-x_{st-1})^2+(y_{st}-y_{st-1})^2]^{1/2}=D \quad (3)$$

当 (x_{st},y_{st}) 和上一点 (x_{st-1},y_{st-1}) 的直线距离 $D\geq L_M$ 时，认为 (x_{st},y_{st}) 是噪声，剔除之；当 $D\leq L_m$ 时，认为 (x_{st},y_{st}) 是分辨率较高造成的冗余点，也剔除.只有当 D 在 $L_m\sim L_M$ 范围内才保留它，并取该点平滑前的值，否则就去掉.故得下式：

$$x=x_{st-1} \quad (D\leq L_m \text{或} D\geq L_M) \quad (4)$$

$$x=x_t \quad (L_m<D<L_M) \quad (5)$$

$$y=y_{st-1} \quad (D\leq L_m \text{或} D\geq L_M) \quad (6)$$

$$y=y_t \quad (L_m<D<L_M) \quad (7)$$

L_M，L_m 值由实验确定，它们对处理的效果影响较大.在本实验中，x、$y\in[1,1\ 000]$ 情况下，$L_m=14$，$L_M=260$.

2.4　规范化

这是一个尺度缩放的过程，用于统一汉字尺寸和纠正位置的变形.最简单也是最常用的方法是外接框规范化[4].在 X，Y 轴方向分别进行线性变换，使汉字位于中心并且正好填满方框.作者采用的也是这种简单有效的规范化方法，规范化后的汉字为 32×32.

3 联机手写汉字特征的提取

汉字特征包括笔划和关系两部分,关系是根据笔划间的位置计算出来的.

3.1 有关定义

定义1 笔划是指一次书写,即从落笔到抬笔之间所形成的轨迹;笔划中具有单一方向的直线段,叫做笔划元;只含有一个笔划元的笔划叫简单笔划,含有多个首尾相接笔划元的笔划叫复合笔划.

定义2 笔划上点的方向是指该点的切线方向.

定义3 笔划元的类型按其笔划方向分横(horizon)、竖(vertical)、撇(slash)、捺(backslash)4种,量化表示为1,2,4,8.设(b_x,b_y),(e_x,e_y)是笔划元的始、终点坐标,且$\tan\alpha=\frac{e_y-b_y}{e_x-b_x}$,则当$\tan\alpha\in[-\frac{1}{5},\frac{2}{5}]$,笔划元为横;当$|\tan\alpha|\geq 5$时,笔划元为竖;当$\tan\alpha\in(\frac{2}{5},5)$,笔划元为撇;$\tan\alpha\in(-5,-\frac{2}{5})$时,笔划元为捺.

我们所提取的模式基元实际上是笔划元.选用笔划元而不是笔划作为模式基元,就是为了提取上的方便.同时用笔划元作基元,有利于统一处理书写时的连笔情况.例如:对于笔划"乙"可分解为"一、／、一、|"4种笔划元.

3.2 笔划的提取

联机条件下笔划提取通常是采用先求出笔划曲线中的角点(指方向发生转折的地方),然后用直线段连接相邻两点作为笔划元.主要有两种方法:一种是曲线的多边形拟合法[5],此方法用于汉字笔划提取上,分割过碎且转折点位置检测不准确.另一种方法是曲线跟踪角点检测法[6],缺点在于计算量太大,而且对于圆滑渐变的曲弧无能为力.但圆滑渐变曲弧在手写汉字中出现的频率相当高.作者在仔细研究了上述方法,分析各自优缺点之后,结合联机汉字笔划提取任务的特点和要求,设计了一种快速、准确、有效的笔划提取方案.

3.2.1 以书写笔划间的时间间隔进行初步分割

由于汉字都是一笔一笔写成,在联机条件下,允许常见连笔时,也很少见到一笔写成一个字的情况.因而在前一笔抬笔到后一笔落笔之间总有一定的等待时间.利用这段时间间隔,可方便地完成对联机笔划的初步提取.正常书写情况下,设一串连续采样点对应的发生时刻为:$t_0,t_1,t_2,\cdots,t_n$.连续测量其中相邻点的时间间隔:$\Delta t_k=t_k-t_{k-1}$,一旦发现$\Delta t_k>T_H$,则认为k点为两个笔划间的分割点,从此提出一个笔划,其中T_H为预先实验给定的门限,取$T_H=7$ ms.经过这样的初步提取处理后,一个整字被分解为若干一笔写成的笔划,如果是简单笔划,则此笔划的提取已完成,若是复合笔划,尚需进一步的分割.

3.2.2 用笔划上点的方向分解笔划元

对经上述分割得出的复合笔划,如"乙、了"等,根据书写的方向变化,可检测出转折点,获得笔划元的分解.笔划上某点的方向可据近邻几点的坐标求得.设k点的切线方向为degree_k,则

$$\mathrm{degree}_k=\frac{y_{k+i}-y_{k-j}}{x_{k+i}-x_{k-j}} \tag{8}$$

$$\mathrm{degree}_k=\mathrm{degree}_{i+1}\quad(k\leq i) \tag{9}$$

$$\mathrm{degree}_k=\mathrm{degree}_{n-j}\quad(k\geq n-j,\ n\text{为笔划中总的点数}) \tag{10}$$

式中i,j为实验参数,本文中取$i=4,j=3$.根据定义3,可得到点k的量化方向.当某点的方向与其前一点的方向不相同时,则认为方向发生转折,选取该点为分割点.该过程把绝大多数复合笔划分解成几个单一方向的笔划元.而无分割错误产生.对连笔同样能做到笔划元分解,如图2所示.

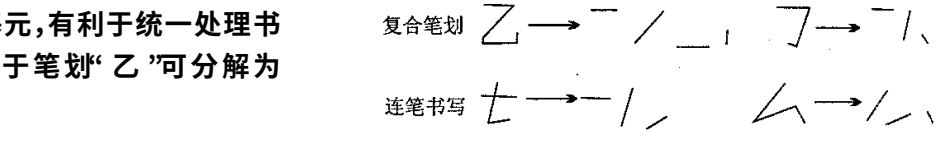


图2 笔划元分解

Fig.1 Releasing of strokes

3.2.3 复合笔划中短小笔划的拼接

经过以上两阶段的处理,汉字已完全分解为笔划元.但由于书写时的抖动、圆滑渐变笔划,或其它干扰等原因,造成复合笔划分解后中间有一部分短小笔划元,如"　　"等.此时根据短小笔划元与前后笔划元间的方向关系,使其淹没在前后笔划元中.经处理后,上述笔划变为"　　／一".

上述3个阶段,处理算法虽然简单,但是因为考虑了联机笔划提取的特点,每阶段各有针对性,所以执行起来十分有效,已经成功地应用于作者所在实验室的联机识别中.

3.3 连接关系的提取

通过笔划提取,一个汉字分解为若干单一方向的笔划元,笔划元之间的相互位置是比较稳定的,因此关系信息对于汉字识别也是至关重要的.现以4种笔划元为基础,分析它们在组成汉字时相互之间的位置关系.

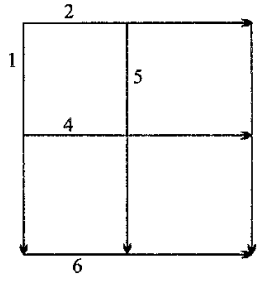
记笔划元的起点为头(head),终点为尾(tail),

头尾之间的非端点为中(middle).

两笔划元的连接：对于两笔划元 A 与 B，如果能在 A 上找到一点 C_a，在 B 上找到一点 C_b，使 C_a 到 C_b 的距离不大于给定常数 e，则称 A 与 B 连接，其中 e 是大于或等于 0 的实数. 如果两笔划元 A 与 B 连接，点 a 与 b 是两笔划元间的最近点，$a\in A$，$b\in B$，a 到 b 的距离为 d，$0\leq d\leq e$，从 a 向 A 笔划元的头或尾测距离，记下距离最短的那个点的标号为 P_a. 从 b 向 B 笔划元的头或尾测距离，记下距离最短的那个点的标号为 P_b(P_a，$P_b\in\{h,t,m\}$)，称 A 对 B 是 P_aP_b 关系.

从以上分析可知，用 R，R′表示 A 与 B 的关系，即 ARB，$BR'A$，则 R 与 R′是对偶的.

综上所述，连接可细分为相连、相抵和相交 3 种. 相连是笔划元的端点和端点之间的连接；相抵是笔划元的端点与非端点之间的连接；相交则是笔划元的非端点与非端点之间的连接. 连接关系共有 9 种，其中相连 4 种，相抵 4 种，相交 1 种，如图 3 所示，图中序号 1，2，…，5 表示笔划元，箭头表示书写方向.



关系名称			对偶关系		
1	hh	2	2	hh	1
2	th	3	3	ht	2
3	tt	6	6	tt	3
6	ht	1	1	th	6
5	hm	2	2	mh	5
5	tm	6	6	mt	5
4	mm	5	5	mm	4

图 3　两笔段之间的 9 种连接关系

Fig. 3　Nine connective relation between two segments

对于联机手写样本，笔划元之间各种关系的稳定性是不一致的. 复合笔划中后一笔划元对前笔划元的 ht 关系最为稳定，其次是 m m 关系，然后是其它几种关系. 本实验中采用适应性强、具有模糊概念、有利于匹配的几种关系类型. 作者在本系统中利用了 11 种，具体的标记及其含义如下：

ht：后一笔划元对前一笔划元的连接关系；

xx：交叉；

tt：端点对非端点的连接关系；

pp：非端点对端点的连接关系；

ll：端点与端点的连接关系；

lt：ll 和 tt 的模糊；

lp：ll 和 pp 的模糊；

lx：ll 和 xx 的模糊；

tx：tt 和 xx 的模糊；

px：pp 和 xx 的模糊；

an：ll、tt(或 pp)和 xx 的模糊.

其中模糊阈值是预先给定的实验常数. 本实验中，ll，tt，pp 之间的模糊阈值 LTP = 1；ll，tt，pp，xx 之间的模糊阈值 LPX = 2.

4　实验结果及其分析

通过对 26 个区 2 444 个字的特征的提取实验，获得 99.8% 的正确率.

笔划元提取发生错误的主要表现是：

1) 复合笔划中的两个小笔划元变形为一个笔划元：当组成两笔划元的点很少(均小于 6)时，在笔划提取的第二阶段，点的方向已经不是所属笔划元的切线方向.

2) 简单笔划由于书写原因被分解成几个笔划元.

参考文献

[1] 张忻中. 汉字识别技术[M]，北京：清华大学出版社，1991.

[2] 王绪龙. 汉字信息处理[M]，北京：国防工业出版社，1990.

[3] TAPPERT C C，SUEN C Y，WAKAHARA T. The State of Art in On-line Handwriting Recognition[J]. **IEEE T-PAMI**，1990(8)：12.

[4] ARAKAWAH. On-line Recognition of Handwriting Characters[J]. **PR**，1983，16(1)：9～16.

[5] WAKAHARA T. On-line Cursive Script Recognition Using LAT[C]. Proc 9th ICPR，1988.

[6] YHAP E F，GREANIAS E C. An On-line Chinese Character Recognition System[J]. **IBM J Res Dev**，1981，25(3)：187～195.

(责任编辑：秦和平)