第13卷 第1期
1987 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 13, No. 1
Jan., 1987

计算机实时手写中文字自动识别

叶 培 建
（北京控制工程研究所）

摘 要

本文介绍用于联机手写中文字自动识别的新方法与新算法．由于下述各点的实现，手写文字时可以减少许多限制，增加书写自由．①笔划的抽取经由两次分段实现：首先连续采样，将输入笔划变换成线段组合，再对线段的长度进行比较，删去相对不重要的成份．②用笔划校正技术将不应分离的笔划重新组合成规范笔划，或者将不应联写的复合笔划重新分解成基本笔划．③用非完全匹配技术使失真字可以识别．④用混序笔划重排算法可使一个混序笔划输入的字重新排列笔顺．⑤笔划位置和长度作为进一步特征，可区别模糊字．

一、引 言

用计算机进行文字处理的主要内容之一是文字的自动识别．目前，文字自动识别作为模式识别的一个分支，其理论与方法已日趋成熟．与对数字、西文及日文等文字的识别研究相比，中文字的自动识别研究起步较晚． 经过日本、美国、中国等国科学工作者的努力，近二十年来在印刷体，限制性手写体的脱机识别，以及在手写体的联机识别等方面都有了许多成就．在许多方法与算法的支持下，识别率及识别速度都不断提高．但由于中文字本身量大、字形复杂、印刷载体的质量和形式又多变、手写的习惯差异很大等因素，中文字自动识别领域还有大量工作要做．

限制性手写体的脱机识别困难较大，自由手写体的脱机识别困难就更大了．但是，限制性手写体的联机识别已有较大的进展．为了减少限制，增加书写自由，本文着重对如何提高书写自由度提出了一些新的方法与算法．

二、系统组成及中文字识别过程

整个实验识别系统由一个数字化板、一个接口、一台终端及一台小型 PDP-11/60 计算机组成．

数字化板可将笔在其表面的运动轨迹读出．笔每放下和提起一次构成一个“笔迹”．笔的轨迹由一串点 $\{P_i\}$ 构成，每点由其 x 和 y 坐标表达：$P_i = (x_i, y_i)$．最小可识别的字的尺寸由数字化板的分辨率决定．接口可以改变数据输送方式(串行、并行)和输送速度．该识别系统采用并行输出，最快可达 80 对数据/秒．终端具有图象功能，因而既可作

本文于 1985 年 2 月 14 日收到．

控制用，又可显示输入及输出字.

中文字的识别与其它文字识别一样[3]，是一个特性抽取和比较的过程.

1. 中文字识别的一些定义

(1) 基本笔划集 $\{e\}$. 中文字中不可再分割的笔划元素称为基本笔划，它们的集合为基本笔划集 $\{e\}=\{e_i\}$，在本实验中，$i=1, 2, \cdots, 18$.

(2) 组合笔划集 $\{\gamma\}$. 由几个基本笔划构成的笔划称为组合笔划，它们的集合为组合笔划集 $\{\gamma\}=\{\gamma_i\}$，在本实验中，$i=1, 2, \cdots, 6$.

(3) 允许笔划集 $\{E\}$. 手写过程中允许出现的笔划称为允许笔划，它们的集合由基本笔划集和组合笔划集构成，记为 $\{E\}=\{e\}+\{\gamma\}$.

(4) 中文字的笔划串. 一个标准的中文字由一个有序的基本笔划串构成:

$$^{r}R=(e'_1, \cdots, e'_j, \cdots),\ e'_j\in\{e\}.$$

一个被识别字由一个非序的允许笔划串构成:

$$X=E'_1, \cdots, E'_j, \cdots\ E'_j\in\{E\}.$$

在约定情况下，一个被识别字可由一有序的允许笔划串构成:

$$X=(E'_1, \cdots, E'_j, \cdots),\ E'_j\in\{E\}.$$

2. 手写中文字自动识别过程

手写中文字自动识别过程可分为预处理、笔划识别及字识别三部份. 下面两节将着重介绍笔划识别及字识别的详细内容.

三、笔 划 识 别

由数字化板输出的笔迹信号经规范化，平滑处理及滤波后即可进行分段. 分段的目的是用几个直线矢量段来表示输入的笔迹而又不失其原特性. 分段的技术有多种，如直线距离法、平均角度法、连续采样法等. 考虑到本系统既要适用于中文笔迹分段，又要适用于西文笔迹分段，因此采用连续采样法分段. 为了进一步消除书写形成的失真，整个分段过程分成二个部份.

1. 分段 1 和连续采样分段法

分段 1 的目的是把输入的一个笔迹以数目最少而又保持笔迹基本特性的几个直线段来表示. 它是用连续采样分段法完成的.

(1) 两门限采样法. 对一笔迹的轨迹点逐一观察，去掉多余的信息点，仅保留反映笔迹特性的点. 具体做法如下:

观察一个笔迹 T

$$T=(P_1, \cdots, P_k, P_{k+1}, P_{k+2}, \cdots, P_K),$$

对 k 从 1 到 $K-2$ 的点列出 $l=l_k=|P_{k+1}-P_k|$，$\alpha=\alpha_k=(\overline{P_kP_{k+1}}, \overline{P_{k+1}P_{k+2}})$. 如果 $\alpha<\alpha_s$，且 $l>l_s$，则点 P_{k+1} 保留. 再以 P_{k+1} 为 P_k，继续算下去. 否则删去 P_{k+1}，以 P_{k+2} 为 P_{k+1}，P_{k+3} 为 P_{k+2}，继续算下去. 这儿 l_s 和 α_s 分别为系统规定的分段长度门限值和角度门限值. 在这个采样过程中，每一循环可以删去一个或者更多的多余信息点.

(2) 最大角度采样法. 对一笔迹的全部点逐一观察，逐次研究相邻点之间的夹角，并

删去具有最大夹角的三点中的中间点．在采样过程中每个循环仅能删去一个多余的信息点，但利用这个方法可以控制一笔迹的最终剩余信息点数目．

(3) 连续采样法．连续采样法是交替应用上述两种采样方式，首先用两门限采样法，一个循环删去大部分多余信息点，接着规定一个笔迹的最终点数目 ND，应用最大角度采样法，并多次重复，一直到笔迹还剩下 ND 点．再重复一次两门限采样法，最后获得一条具有不大于 ND 点的新笔迹． 考虑到中文字本身笔划的特性，ND 取五便足以表示 $\{E\}$ 中全部笔划．

2. 分段 2 (补充分段)

补充分段的目的在于删去残留在已分段笔迹中的失真．在分段 1 中，由于门限值 l_s 依赖于笔迹总长，它不能取得过大，否则会在分段过程中删去可能的有效笔迹段．而 l_s 取较小值则会保留不该保留的笔迹段．补充分段的作用就为了删去这些多余的笔迹段．在补充分段过程中作为参考的比较长度不是笔迹总长，而是把全部段中具有第二长长度的那段作为参考段长度，如果一笔迹仅有两段，则取长的那段为参考段．

3. 笔划特性

为了进行笔划识别，必须给笔划选择一些能反映其特性的参数．可供选择的参数有笔划的段数、段的特性(长度及方向)、两段间的夹角、笔划的重心和笔划的长度．

(1) 笔划段数为 K．

(2) 段特征一般是一矢量，记为 $S(k)$，

$$S(k) = [l(k),\ O(k)].$$

$l(k)$ 为 $S(k)$ 的长度；$O(k)$ 为 $S(k)$ 的方向，它是平面八方向之一．

(3) 两段间夹角 $\theta(k)$ 定义为 $S(k)$ 按顺时针方向转到 $S(k+1)$ 时所形成的角度．

(4) 笔划重心 g 由组成该笔划的全部点的几何平均值决定．

(5) 笔划长度．中文字中，一般多段笔划字中的段长度不起改变字义的作用，而笔划较少的字中笔划的长度往往引起字义的变化，所以对笔划长度 λ 作如下定义：

$$\lambda = l(1),\ \text{如果}\ K = 1;$$

$$\lambda = 0,\quad \text{如果}\ K > 1,$$

4. 笔划识别

首先根据中文字笔划的特性，依据“树”的理论建立一棵 “笔划识别树”．该树的全部树顶由$\{E\}$中元素构成．当被识别的笔划来到时，根据已抽取的特性，令其在树上沿着合适的枝攀登．如它到达一个顶，该顶就是被识别的结果，如果它无法到达一个顶，那么该笔划被判为非 $[E]$ 元素，识别过程停止．

5. 笔划识别校正

到目前为止所考虑的笔划是由输入的一个笔迹演变而来的．但由于多种原因 (笔板接触不好、抖动、人为原因等)，在写字过程中把不应分离的笔划写成几个分开的笔迹，或者把几个应分离的笔划联写成了一个笔迹．为了解决这个问题，要在初步识别的基础上，对所得到的笔划进行校正，以便把所有的输入笔迹识别成数目最为合理的基本笔划．

主要的校正措施有：(1) 把不应分离的笔迹联结起来．这由连续笔迹之间的距离和

这些笔迹相应笔划的可联性来决定．（2）把属于组合笔划的笔划按规则分解成基本笔划．

6. 笔划和字的描述

一个字的全部笔迹经上述处理后，可由一串笔划来表示：

$$A\triangleq({}^{A}N,\ {}^{1}T^{A},\cdots,\ {}^{n}T^{A},\cdots,\ {}^{A_N}T^{A}).$$

其中，${}^{A}N$ 为字 A 的笔划数目；${}^{n}T^{A}=[{}^{A}C(n),\ {}^{A}g(n),\ {}^{A}\lambda(n)]$．${}^{A}C(n)$ 为笔划 ${}^{n}T^{A}$ 的笔划级，即一基本笔划的代码；${}^{A}g(n)$ 为笔划 ${}^{n}T^{A}$ 的重心；${}^{A}\lambda(n)$ 为笔划 ${}^{n}T^{A}$ 的长度．

四、字 识 别

1. 字识别的基本过程

利用一个四重组，可以把标准字集 $\{R\}$ 中的一个字 ${}^{r}R$ 和被识别字 X 分别表达成：

$$X=({}^{x}N,\ {}^{x}C,\ {}^{x}G,\ {}^{x}L),$$
$${}^{x}C=[{}^{x}C(1),\cdots,{}^{x}C(n),\ \cdots,\ {}^{x}C({}^{x}N)],$$
$${}^{x}G=[{}^{x}g(1),\cdots,\ {}^{x}g(n),\cdots,\ {}^{x}g({}^{x}N)],$$
$${}^{x}L=[{}^{x}\lambda(1),\ \cdots,\ {}^{x}\lambda(n),\ \cdots,\ {}^{x}\lambda({}^{x}N)],$$
$${}^{r}R=({}^{r}N,\ {}^{r}C,\ {}^{r}G,\ {}^{r}L),$$
$${}^{r}C=[{}^{r}C(1),\ \cdots,\ {}^{r}C(n),\ \cdots,\ {}^{r}C({}^{r}N)],$$
$${}^{r}G=[{}^{r}g(1),\ \cdots,\ {}^{r}g(n),\ \cdots,\ {}^{r}g({}^{r}N)],$$
$${}^{r}L=[{}^{r}\lambda(1),\ \cdots,\ {}^{r}\lambda(n),\ \cdots,\ {}^{r}\lambda({}^{r}N)].$$

X 的识别过程将是 X 和 ${}^{r}R$ 两个四重组的比较过程．X 将被识别为最为接近的那个 ${}^{r}R$．在手写文字识别中，为了解决笔划的失真及书写顺序混乱的问题，本文介绍两种算法，它们允许两个字的笔划不完全匹配，并能重排混序书写的字的笔划顺序．

2. 两笔划差的算法

设有两笔划 ${}^{a}T$ 和 ${}^{b}T$，并且 ${}^{a}T,\ {}^{b}T\in\{e\}$．${}^{a}T$ 的笔划级为 $C(a)$，${}^{b}T$ 的笔划级为 $C(b)$．

$${}^{a}T=\{K(a),\ {}^{a}S(1),\cdots,\ {}^{a}S(k),\cdots,\ {}^{a}S[K(a)]\},$$
$${}^{b}T=\{K(b),\ {}^{b}S(1),\cdots,\ {}^{b}S(k),\cdots,\ {}^{b}S[K(b)]\}.$$

$K(a)$，$K(b)$ 为 ${}^{a}T$，${}^{b}T$ 的段数[假定 $K(b)\geqslant K(a)$]．${}^{a}S(k)$ 和 ${}^{b}S(k)$ 隐合各段的长度和方向：

$${}^{a}S(k)=[{}^{a}l(k),\ {}^{a}O(k)],$$
$${}^{b}S(k)=[{}^{b}l(k),\ {}^{b}O(k)].$$

两笔划差是笔划段数差(记为 $dn({}^{a}T,\ {}^{b}T)$) 与笔划各段方向差(记为 $do({}^{a}T,\ {}^{b}T)$)之和．

$$dn({}^{a}T,\ {}^{b}T)=|[K(b)-\beta_b\times 0.5]-[K(a)-\beta_a\times 0.5]|.$$

β_a，β_b 取值为 1 或 0，这由 ${}^{a}T$，${}^{b}T$ 的性质决定．如果笔划包含有很短的一段，β 值取 1，否则取 0．

$$do({}^{a}T,\ {}^{b}T) = \underset{K(b)-K(a)+1}{\mathrm{Min}} \left[\frac{\sum_{h=1}^{K(a)} {}^{a}O(k) - {}^{b}O(k+m-1)}{K(a)} \right]$$

其中 $m = 1,\ 2,\ \cdots,\ K(b) - K(a) + 1$.

考虑到这两项差在两笔划差中所起的作用不一样,在笔划段数相近时,希望 do 起主导作用,而在笔划段相差较大时,希望 dn 起主导作用,因而引入权 ω, 使最后获得两笔划的差为:

$$DC({}^{a}T,\ {}^{b}T) = [dn({}^{a}T,\ {}^{b}T)]^2 \cdot \omega_1 + do({}^{a}T,\ {}^{b}T) \cdot \omega_2.$$

ω_1 和 ω_2 为权,本系统中, ω_1 和 ω_2 分别取 1 和 0.8.

3. 混序笔划重排顺序算法

两个字进行比较时, 其笔划书写顺序是非常重要的. 假定一被识别字的各笔划书写顺序是任意的,则它与参考字比较之前, 应首先将其笔划顺序按与之比较的参考字重排. 重排顺序算法的基本思想如下:

(1) 被识别字笔划顺序按参考字笔划顺序重排的必要性和可能性: 如果两字之间的最小笔划差之和小于一个规定值,说明这两字有较高的笔划相似程度,有可能也有必要重排,否则重排是不合理的.

(2) 把一个字的全部笔划按笔划级分成若干分集. 对于两字中相应的分集, 研究如何把它们的元素一一配对, 使它们之间的位置差之和最小. 具有最小位置差之和的配对决定了被识别字中各笔划应占有的顺序位置.

4. 字识别

字识别的过程就是把 X 和参考字集 $\{R\}$ 中的每个字 ${}^{r}R$ 进行比较. 考虑到必要和优先性,以下述的要求和顺序进行比较:

(1) 笔划数目差 $DN(X, {}^{r}R)$,

$$DN(X,\ {}^{r}R) = |{}^{r}N - {}^{X}N|.$$

(2) 笔划级差 $DT(X, {}^{r}R)$,

$$DT(X,\ {}^{r}R) = \sum_{n=1}^{N} DC({}^{n}T^{X},\ {}^{n}T^{r}).$$

(3) 笔划位置差 $DP(X, {}^{r}R)$,

$$DP(X,\ {}^{r}R) = \sum_{n=1}^{N} |{}^{X}g(n) - {}^{r}g(n)|$$

(4) 笔划综合差 $DG(X, {}^{r}R)$,

$$DG(X,\ {}^{r}R) = \frac{DT(X, {}^{r}R)}{\overline{DT}} + \frac{DG(X,\ {}^{r}R)}{\overline{DG}}.$$

这里 $\overline{DT}$ 和 $\overline{DG}$ 是由系统特性决定的两个归化参数.

(5) 笔划长度差 $DL(X,\ {}^{r}R)$,

$$DL(X,\ {}^{r}R) = \underset{N}{\mathrm{Max}}[\,|{}^{X}\lambda(n) - {}^{r}\lambda(n)|\,].$$

具体的比较过程要遵循一定的规则,以减少比较时间. 主要规则为"逐级选择",即每

一次比较仅对上一步中选的候选者起作用.

各步比较过程的具体条件主要包括门限值的选择和识别条件. 其中门限值 S_t,S_p 和 S_g 的选择主要考虑到每步既不要漏掉可能的候选者,又不至于有太多的候选者. 它们受字的大小、笔划数目的多少、笔划级等参数的影响. 识别条件为 $DN(X, {}^rR) = 0$, $DT(X, {}^rR) < S_t$, $DP(X, {}^rR) < S_p$, $DG(X, {}^rR) < S_g$, $DL(X, {}^rR)$ 为最小.

五、结　论

用本文所介绍的实验系统对几个人分别写的 500 字集（500 不同字）进行了初步测定. 在考虑书写顺序不完全匹配的情况下,正确识别率达 99%; 而当不考虑书写顺序时,正确识别率也可达 92%. 错误识别率与拒识率分别为 4%. 识别一个字的时间约为 1—2 秒. 执行程序占用 200 × 256 字节的存贮空间.

与其它已提出的一些方法相比[4,5],这个实验系统具有如下特点:

(1) 可用于不完全匹配的文字识别,使手写笔迹更加自由.

(2) 适用于混序书写识别,系统具有通用性(多人书写同一字, 很难做到笔划顺序完全一致).

(3) 分段技术合理,适应性强,允许有较大的书写噪声.

(4) 笔划校正技术在一定程度上解决了笔划分离和联写的问题.

(5) 笔划长度的引入,使一个模糊字集中的字得以辨别(如辨别士和土).

参　考　文　献

[1] Nakata, K., Problems in Chinese Character Recognition, Central Research Lab. Hitach, Tokoyo, Japan, First USA-JAPAN Computer Conf., 1972.

[2] Mori, Advance in Recognition of Chinese character, JAPAN, CH 1499-3/80/0000-0692500.75 ©1980, *IEEE*.

[3] Fu, K. S., Application of Partern Handling Methods to Chinese Language Processing, School of Electrical Engineering, Purdue University, U. S. A., 1979.

[4] Ye, P. J., Hugli, H., Pellandini, F., Techniques for on-line Handwritten Chinese Character Recognition with Reduced Writing Constraints, 7th Int. Conf. On Pattern Recognition, 1984.

[5] Ye, P. J., Hugli, H., Pellandini, F., On-line recognition of Handwritten Chinese character: Rearrangement of Stroke Sequence, First I. F. S. A Congress, 1985.

COMPUTER ON-LINE HANDWRITTEN CHINESE CHARACTER RECOGNITION

YE PEIJIAN

(Beijing Institute of Control Engineering)

ABSTRACT

In this paper, a few new techniques and algorithms for computer on-line handwritten Chinese character recognition are presented. More freedom in writing can be provided by the following points: (1) segmentation is based firstly on successive sampling and secondly on a method which eliminates, based on a local comparison, the undesired relative short segments of a strok; (2) by means of correction algorithm, strokes which should not be separated can be linked together, and strokes which should not be linked can be split into several basic strokes; (3) the distorted characters can be recognized by a non-exact matching process; (4) with reference to the stroke sequence order of a template character stored in the dictionary, the stroke sequence order of an input character can be rearranged; (5) use stroke position and stroke length as additional features to distinguish ambiguous characters.