国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第19卷第1期 1997年2月 Vol. 19 No. 1

一种联机手写汉字识别方法*

姚丹霖 殷建平

(国防科技大学计算机系 长沙 410073)

摘 要 本文给出了一种基于动态汉字基元获取笔段有序序列的联机手写汉字识别方法。该方法对汉字书写笔顺无任何限制，对手写汉字变形有很好的适应能力。经测试，基于本方法研制的联机手写汉字识别系统的识别率为97.1%.

关键词 手写汉字，汉字基元，闭包，模式识别

分类号 TP391.4

An Approach to Recognize On-line Handwritten Chinese Characters

Yao Danlin Yin Jianping

(Department of Computer Science, NUDT, Changsha, 410073)

Abstract This paper introduces an approach to recognizing on-line handwritten Chinese characters, which gets ordered sequence of line segments based on the Dynamic Chinese Character Base-units. This approach has no restriction on writing order, and is adaptable to variants of handwritten Chinese character. As a result of testing, the recognizing rate of a system based on this approach is 97.1%.

Key words handwritten Chinese character, Chinese character base-unit, closure, pattern recognition

1 动态汉字基元

所谓动态汉字基元是指一个或多个笔段构成的最大集合，在此集合内任何一个笔段与同一集合内其他所有笔段具有直接或间接的交接关系。一个动态汉字基元具有两个属性：动态汉字基元中所有笔段构成的集合、动态汉字基元中所有笔段占据的矩形区域。

* 1996年7月22日收稿

设一个汉字具有 n 个笔段、m 个动态汉字基元，用 CC_i $(1\leq i\leq m)$ 表示第 i 个动态汉字基元，用 CS_i $(1\leq i\leq m)$ 表示第 i 个动态汉字基元中笔段的集合，用 CR_i $(1\leq i\leq m)$ 表示第 i 个动态汉字基元所占据的矩形区域，S_i $(1\leq i\leq n)$ 表示第 i 个笔段，$R(S_a, S_b)$ 表示两个笔段 S_a 和 S_b 之间的交接关系，则有：

$$CC_i = < CS_i,\ CR_i >$$

其中 $CS_i = \{S_{i1}, S_{i2}, \cdots, S_{ik}\}$ $\quad 1\leq i\leq m$ 且 $1\leq k\leq n$

满足：$\forall S_a(S_a\in CS_i \wedge CS_i - \{S_a\} \neq \varnothing \rightarrow \exists S_b \ S_b\in CS_i - \{S_a\} \wedge R(S_a, S_b) \wedge P(CS_i - \{S_a, S_b\}, \{S_a, S_b\}))$

并且：$\forall S_b\ (\exists S_a S_a\in CS_i \wedge R(S_a, S_b) \rightarrow S_b\in CS_i)$

其中 $P(A, B) \equiv A \neq \approx \rightarrow \exists S_a \exists S_b (S_a\in A \wedge S_b\in B \wedge R(S_a, S_b) \wedge P(A - \{S_a\}, B + \{S_a\}))$

在一个汉字的若干个动态汉字基元中，如果存在这样的两个动态汉字基元 CC_i 和 CC_j，其中 CC_i 的矩形区域 CR_i 有足够的部分包含在 CC_j 的矩形区域 CR_j 之内，则应将 CC_i 和 CC_j 进一步合并为广义动态汉字基元。参见图1给出的示例。

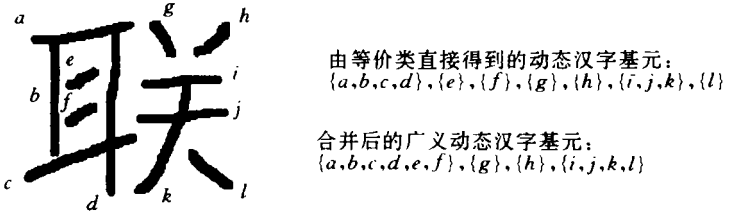

图1　动态汉字基元的合并

基于占据的矩形区域，可以定义任意两个动态汉字基元之间的空间位置关系，其空间位置关系共有八种，参见表1.

将动态汉字基元和部件相比，二者具有如下两个相同的性质：

[1] 内部结构稳定，因而它们包含的笔段可以稳定排序；

④在构成一个汉字时，遵循从上到下、从左到右的汉字结构原则，因而它们之间也可以稳定排序。

但动态汉字基元和部件又有着本质的区别：前者是一个动态的概念，对于同一个汉字的不同手写字，所包含的动态汉字基元有可能不尽相同，但它可以实时提取；后者是一个静态的概念，无法实时提取，只能在识别过程中通过回溯匹配而得到。

2　笔段有序序列的获取

为了得到手写汉字的一维笔段有序序列，应对得到的手写汉字的所有（广义）动态

汉字基元及其所包含的笔段分别进行排序。

表1　动态汉字基元间的空间位置关系

关　系	关系名称	示　例
1	a 左 b 右	a　b
2	a 左上 b 右下	a 　b
3	a 上 b 下	a b
4	a 右上 b 左下	a b
5	a 右 b 左	b　a
6	a 右下 b 左上	b 　a
7	a 下 b 上	b a
8	a 左下 b 右上	b a

首先，基于笔段及其相互关系，定义笔段优先函数如下：

$$f_{SP}:\ S\times S\times R_s\rightarrow N$$

其中 S 为笔段类型集合，R_s 为笔段关系集合，N 为优先数集合

根据此优先函数，对每一个（广义）动态汉字基元，计算其所包含的各个笔段的优先数，按笔段优先数从大到小进行排序，即得到笔段有序的（广义）动态汉字基元。

其次，基于动态汉字基元的空间位置关系，定义（广义）动态汉字基元的优先函数如下：

$$f_{CP}:\ R_c\rightarrow N$$

其中 R_c 为动态汉字基元空间关系集合，N 为优先数集合。

根据此优先函数以及（广义）动态汉字基元的空间位置关系，可计算出手写汉字中各（广义）动态汉字基元的优先数，按优先数从大到小进行排序，即得到构成该手写汉字的（广义）动态汉字基元有序序列，对此有序序列实施一次扫描，即可得到构成一个汉字的一维笔段有序序列。

3　识别算法

设得到的手写汉字的一维笔段有序序列为：　$WS:\ S_1S_2\cdots S_j\cdots S_m$

而标准库中第 i 字的一维笔段有序序列为：　$NS_i:\ S_{i1}S_{i2}\cdots S_{ik}\cdots S_{in}$

识别过程就是计算 WS 和 NS_i 之间的相似度　$P\ (WS,\ NS_i)$

考虑到书写汉字时可能出现的笔段增加或减少，因而识别算法应能处理笔段的插入和删除。设当前参与匹配的笔段分别为 WS 的 S_j 和 NS_i 的 S_{ik}，则笔段的匹配、插入和删除操作如下：

[1] 匹配：当 S_j 和 S_{ik} 为相同类型的笔段或相邻类型的笔段时，则两笔段是可匹配的，其相似度为：

若 S_j 和 S_{ik} 为相同类型笔段，则 $p\ (S_j,\ S_{ik}) = 1$；

若 S_j 和 S_{ik} 为相邻类型笔段，则 $p\ (S_j,\ S_{ik}) = 0.5$；

④插入：当 S_j 和 S_{ik} 不可匹配、且 $m-j<n-k$ 时，表明 WS 中丢失了笔段，匹配过程中出现插入操作。亦即用一空笔段和 S_{ik} 匹配，其相似度为：

$$p\ (Null,\ S_{ik}) = 0.3$$

㈣删除：当 S_j 和 S_{ik} 不可匹配、且 $m-j>n-k$ 时，表明 WS 中出现了多余的笔段，匹配过程中出现删除操作。亦即用一个空笔段和 S_j 匹配，其相似度为：

$$p\ (S_j,\ Null) = 0.4$$

当匹配过程结束时，WS 和 NS_i 之间的总相似度为：

$$P_i = \sum_{k=1}^{n_i} p_{ik}$$

其中 n_i 是匹配过程中所进行的匹配、插入和删除操作的总次数。对 P_i 进行规范化，得到平均相似度为：

$$MP_i = P_i / n_i$$

当 MP_i 大于给定的阈值（例如0.7）时，则标准库中第 i 字是当前手写汉字的一个候选字。当有多个候选字时，各候选字按 MP_i 值的大小依次排序，具有最大平均相似度的候选字即为识别结果。

4 结 论

从前面的叙述可以看出，本方法着重于手写汉字的动态结构特征，提取得到的手写汉字的最终特征是一个一维的笔段有序序列，而匹配识别算法则是基于通常的串匹配的无回溯匹配算法，其空间和时间复杂度均为 $\mathbf{O}\ (n)$。

本方法对手写汉字的书写限制极其宽松，对手写字的大小、笔顺无任何约束，对连笔亦能很好的处理，对汉字部件的变形具有较强的适应能力。

特别地，由于最终特征是有序的一维特征，从而本方法天生具备极强的知识获取和知识更新能力，因此，采用本方法所设计的联机手写汉字识别系统，可逐步地适应特定用户的书写习惯，不断提高识别率，而这种使系统逐步适应用户以改善系统性能的方法，可能正是联机自然手写汉字识别系统能够真正达到高识别率的唯一途径。

参 考 文 献

1 汪庆宝，张征. 以笔划结构分析为基础的一种限制性手写汉字识别方法. 电子学报，1987，15（3）：118～121

2 Chen B，Lee H J. Recognition of handwritten Chinese characters via short line segments. Pattern Recogn.，1992，25（5）：543～552

3 张 中. 汉字识别技术. 北京：清华大学出版社，1992：69～73

（责任编辑 张 静）