45 79-82

联机手写体汉字识别方法的研究*

TP391.4

金楚屯　赵学军　蔡　雷　潘保昌
（重庆大学人工视觉实验室，重庆，400044；第一作者 35 岁，男，讲师，博士生）

摘　要　提出了一种基于笔划的一级分类、笔划特征二级分类的新方法来实现联机手写体汉字的识别。该方法对笔划变形的容忍度大、计算简单。通过两级分类，对国标二级汉字的识别率为 98% 以上，不要求笔顺，识别时间短。

关键词　汉字识别；笔划分类；笔划特征 / 距离准则
中国图书资料分类法分类号　TP391.4

模式识别　手写体

0　引　　言

目前，联机手写体汉字识别，在模式识别研究领域中，已取得一定的成功，联机手写体汉字识别系统在笔输入计算机方面已有较成功的应用。与传统的汉字编码输入方法相比较，联机汉字输入简单直观，容易学习，因此越来越多的有识之士投入此领域的研究，并使其完善，推广应用。

笔者提出了一种联机手写体汉字识别方法，它是基于笔划识别的一级分类，笔划特征的二级分类来完成整字识别的。在笔划识别中，采用了笔者在论文[1]中所提出的笔划识别方法，来实现笔划的正确识别。在文中，根据汉字的特征，定义了汉字笔划间的 7 种特征量，这些特征量易于抽取，有利于二级分类。本文在最后识别汉字时采用了最小绝对距离和最小相对距离准则相配合的关系，把距离最小的模式作为识别结果。

1　笔划识别

文献[1]中详细讨论了笔划的识别问题，把构成汉字的基本单元分成 7 种笔划（第一种分类），或者 9 种笔划（第二、第三种分类）或者 11 种笔划（第四种分类），笔者仍趋向于简单明了的第一种笔划分类，如表 1，这种分类有利于笔划的正确识别。

表 1　笔划代码分类表

笔划名称	横（提）	竖	撇	捺（点）	顺笔划	逆笔划	混合笔
笔划代码	1	2	3	4	5	6	7

表中，前 4 种笔划，称为单向笔划或基本笔划，后 3 种笔划称为变向笔划或者复合笔划，后者反映了在同一笔划中笔向的改变方向。顺笔划代表顺时针改变笔向，逆笔划代表逆时针

* 收文日期　1997-06-09

改变笔向，混合笔划代表同一笔划中，既有顺时针又有逆时针的变向。

2 笔划间特征量的定义及识别

汉字的一个较大特色，就是任何一个汉字的每一笔划间存在着一个相对的位置关系，这些关系表现了一个汉字的外形结构。显然正确地选取那些便于抽取的笔划间的特征信息，有利于整字的识别，以下先通过举例，来提出笔划间特征的定义。

通过笔输入装置，输入汉字“中”，经笔划识别构成“中”字的笔划代码是“2512”，在汉字标准字典中，可能会出现重码汉字，或者把“中”字识别成“4512”，即把第一笔划竖识别成捺。因此，笔者得利用笔划的特征信息，再进行下一级分类，找出识别结果。

在表 2 中，定义了 7 种笔划特征量，这些特征量的选取便于从输入点坐标中抽取出来。在笔者识别出当前输入笔划的同时，就可以确定它与以前 n 个笔划的位置关系。当 $n=1$ 时，只确定它与前一个笔划的位置关系，当 $n=2$ 时，确定与前二个笔划的位置关系，最大时，是确定与以前所有笔划的位置关系，显然当 n 越大，所反映的各个笔划位置关系就越多，复杂性大，n 选取较小，容易丢失相关的特征信息。通过对大量汉字的分析研究[2] 发现，构成一个汉字的基本单元是笔划，由若干笔划构成一个汉字中的结构单元，由若干结构单元，结合成一个汉字，75% 以上的汉字的结构单元都是由少于 5 种笔划构成的，且相邻笔划的相关关系最大。例如“中”的笔划识别码为“2512”，如取 $n=1$ 则笔划特征码为“146”.

表 2 笔划特征代码分类表

笔划关系	首首相连	首尾相连	尾首相连	尾尾相连	笔划相切	笔划相交	笔划相隔
特征代码	1	2	3	4	5	6	7

当汉字的笔划数较大时，如 n 选择较大，特征码的复杂程度增大，给识别带来困难。笔者注意汉字识别的一个特点，笔划数越多的汉字，只利用笔划代码信息，也比较容易识别，这是因为在笔划代码中已经包含了大量的汉字信息，甚至可以不用特征码，所以 n 的选取与笔划数 m 有较大的关系，通过大量的试验，总结出如表 3 所给出的关系。

表 3 笔划位置关系表

笔划数 m	$m\leqslant 3$	$3<m\leqslant 7$	$7<m\leqslant 12$	$12<m$
n 值	$n=\{\}$	$n=2$	$n=1$	$n=12$
位置关系	前所有笔划	前 2 个笔划	前 1 个笔划	只前 12 个笔划

在构成标准汉字字典外除有笔划代码外，还有相应的特征代码。

3 整字识别中的距离准则

汉字笔划码和特征码确定之后，与预先存储在字典中的标准笔划码和特征码匹配，找出最相似的汉字，获得识别结果，下面分三步介绍识别过程。

第一，进行笔划代码粗分类。假定所输入的汉字笔划代码是 $\{p_i\}$，字典中标准笔划代码是 $\{q_i\}$，$i=1\cdots m$，m 为笔划数，笔者可以计算输入笔划代码和标准笔划代码间的距离。在笔者的研究中采用绝对距离和相对距离准则来反映输入汉字与标准汉字之间的关系。

绝对距离：$d_{ab}=1/m*\sum(1-a_i)$，　　$0\leqslant a_i\leqslant 1$

相对距离：$d_{xd}=1/m*\sum(1-\max(a_i))$，　　$0\leqslant \max(a_i)\leqslant 1$

其中，α_i 反映了第 i 个输入笔划 p_i 与第 i 个标准笔划 q_i 的关系，$\max(\alpha_i)$ 反映了第 i 个输入笔划 p_i 与其余标准笔划中相似程度 α 最大的笔划间的关系。经过大量分析研究，笔者把七种笔划间的关系，列于表 4 所示。相同笔划的相似程度最大 $\alpha = 1$，反之不相似的笔划 $\alpha = 0$. 在笔者以后的研究中，论述了通过神经网络自学习方法，针对不同人书写习惯，来修正表 4 中 α 的值，使其联机汉字识别系统适用于不同人的书写风格。

表 4　输入笔划与标准笔划关系表

α 值		笔划代码						
		1	2	3	4	5	6	7
笔	1	1	0	0.2	0.7	0.25	0	0
	2	0	1	0.7	0.7	0.1	0.1	0
划	3	0.2	0.7	1	0	0.1	0	0
	4	0.7	0.7	0	1	0	0.1	0
代	5	0.25	0.1	0.1	0	1	0	0
	6	0	0.1	0	0.1	0	1	0
码	7	0	0	0	0	0	0	1

绝对距离是人们在联机识别中通常所选取的距离准则，在笔者的研究中，采用两种距离相结合的原则，与标准字典中的汉字相匹配，更有利于寻找汉字。假定选取 4 个判别阈值，D_{ab}，D_{rel}，D_{ab0}，D_{rel0}，如有：

1) $d_{ab} < D_{ab}$　　绝对距离小于某阈值

2) $d_{rel} < D_{rel}$　　或相对距离小于某阈值

3) $d_{ab} < D_{ab0}$ 且 $d_{rel} < D_{rel0}$　　或绝对距离和相对距离分别小于某阈值成立，则在标准字典中，找出了一组与输入笔划有最小绝对距离和最小相对距离的汉字。

第二，进行笔划特征码的细分类。根据笔划代码粗分类，把要识别的汉字压缩到几十个、或几个汉字的范围，再利用所获取的笔划特征码来找出识别结果。这里匹配的原则仍采用上述绝对距离和相对距离的概念来处理。

绝对距离：$d_{ab} = 1/m' * \sum(1 - \alpha_i')$，　　$0 \leqslant \alpha_i' \leqslant 1$

相对距离：$d_{rel} = 1/m' * \sum(1 - \max(\alpha_i'))$　　$0 \leqslant \max(\alpha_i') \leqslant 1$

其中，α_i'，$\max(\alpha_i')$ 的函义与 α_i，$\max(\alpha_i)$ 是相同的，笔者把 7 种特征码之间的相似程度 α_i' 列于表 5 所示，同样，相同的特征码相似程度最大，为 1，反之不相似特征码为 0. 在笔者以后的研究中，也论述了用神经网络自学习方法，来修正表 5 中 α_i' 的值。

表 5　特征代码关系表

α' 值		笔划代码						
		1	2	3	4	5	6	7
特	1	1	0	0	0	0.5	0.5	0.5
	2	0	1	0	0	0.5	0.5	0.5
征	3	0	0	1	0	0.5	0.5	0.5
	4	0	0	0	1	0.5	0.5	0.5
代	5	0.5	0.5	0.5	0.5	1	0.6	0.2
	6	0.5	0.5	0.5	0.5	0.6	1	0
码	7	0.5	0.5	0.5	0.5	0.2	0	1

选取 4 个判别阈值，D_{ds}、D_{rd}、D_{dm0}、D_{rd0}，若有：

1) $d_{ds} < D_{ds}$

2) $d_{rd} < D_{rd}$

3) $d_{ds} < D_{dm0}$ 且 $d_{rd} < D_{rd0}$

成立，则在标准字典中，又找出了一组与笔划特征码有最小绝对距离和最小相对距离的汉字，当判别阈值取得较小时，可以把对应的这组汉字数压缩到很少，甚至只有一个汉字，从而达到识别汉字的目的。

第三，当汉字的笔划代码之间及笔划特征码之间都很相似的情况下，几种距离都很小，例如"土"和"士"，这时，可采用人机对话方式由书写者来选择，或其它笔划特征来识别。

4 结束语

笔者采用了 7 种笔划作为组成汉字的基本元素，能比较容易地处理那些不规范的笔划，对笔划的变形也容易接收，采用绝对距离和相对距离相结合的匹配方法，能克服汉字书写笔顺的要求，对个别错误输入笔划或错误识别笔划，通过特征码的细分类进行补充。笔者提出的方法，在 386 计算机上用 C 语言实现，建立了国标二级汉字的特征字典，当书写者比较规范地书写汉字时，经大量的录入试验，最后总的识别率可以达到 98% 以上。

参 考 文 献

1 余楚中，赵学军，彭静. 联机手写体汉字识别中的笔划分类及笔划识别. 重庆大学学报，1998，21(2) 2：131～134

2 贾永康. 识别联机手写体汉字的多级分类方法. 信号处理，1995，(12)：32～34

3 张中. 汉字识别技术. 北京：清华大学出版社，1992，63～76

4 胡家忠. 计算机文字识别技术. 北京：气象出版社，52～67

A Study of On-line Handwriting Chinese Character Recognition

Yu Chuchong Zhao Xuejun Cai Lei Pan Baochang

(Laboratory of Artificial Vision, Chongqing University)

ABSTRACT Based on the first classifying of stroke and the second class of stroke characteristics a new method of completing on-line handwriting Chinese character recognition is presented. There are several advantages with on-line handwriting chinese character recognition such as simplicity, high degree of tolerance to stroke variance, and recognition of up to 98%.

KEYWORDS Chinese character recognition; stroke classifying; stroke characteristics / stroke combination; distance Rule