第27卷 第4期 湖南大学学报（自然科学版） Vol. 27, No. 4
2000年8月 Journal of Hunan University(Natural Sciences Edition) Aug. 2000

文章编号:1000-2472(2000)04-0101-05

基于笔划和笔顺的汉字识别算法

陈治平，林亚平，李军义
（湖南大学计算机科学系，湖南 长沙 410082）

摘 要：以笔划为基元结合笔划的顺序来表示汉字的结构信息，在此基础上提出了一种手写汉字识别的匹配算法，对于结构类似的汉字，该算法可以通过特征关系予以识别，从而提高汉字的识别率。

关键词：汉字识别；手写汉字；冒泡排序算法
中图分类号：TP391.43 **文献标识码**：A

A Recognition Algorithm of Chinese Character Based on Stroke Segment and Order

CHEN Zhi-ping, LIN Ya-ping, LI Jun-yi
(Dept of Computer Science, Hunan Univ, Changsha 410082, China)

Abstract: This paper takes stroke segment as pattern primitive to represent the structure information of Chinese character. On the basis, a new algorithm to recognize handwritten Chinese character is proposed. The characters with the similar structure information can be recognized by the characteristic relation so that the recognition performance is improved.

Key words: character recognition; handwritten character; bob sort method

汉字识别是通过输入装置将汉字的点阵图形输入到计算机，由计算机在汉字字符集合中识别出与之相匹配的汉字的过程. 汉字识别可以根据汉字的产生方式分为印刷体汉字识别、联机手写体汉字识别和脱机手写体汉字识别，其中最为困难的是非特定人手写汉字的识别. 非特定人手写汉字的识别的主要困难和问题表现在以下几点：

1)基本笔划变化，横不平、竖不直、直笔变弯等；

2)笔划模糊、不规范、该连的不连，不该连的却相连；

3)笔划发生倾斜、笔划长短、粗细、位置发生变化；

4)输入汉字存在增笔与减笔的现象.

总之，手写汉字字型的变化是最难以解决的问题，尤其是对于一些特征相似的汉字的识别结果难以令人满意. 本文在分析已有的手写汉字识别系统的基础上，基于笔划特征，

收稿日期：1999-10-10
作者简介：陈治平(1971－)，男，湖南长沙人，湖南大学讲师.

提出了一种新的手写汉字识别系统的方法. 该算法结合笔划及其笔划顺序对手写汉字进行识别,再利用标准汉字模板的特征关系对特征相似的汉字进行识别.

1 手写汉字笔划段的抽取

汉字可以用结构和数值两个特征描述.一般情况下,结构特征是对汉字笔划及其位置关系的描述,是区别不同汉字的最基本的信息.以结构特征为基础就可以消除手写汉字中的笔划与笔划之间的位置发生变化、笔划的粗细变化以及笔划长短发生变化等问题,结合汉字的数值特征,可使汉字的识别率得到很大的提高.

从汉字的结构特征来看,汉字是由基本笔划组成的二维图形,不同汉字其构成的笔划类型的数目是不同的,正确而快速地抽取笔划是汉字识别技术中关键的一环.通过对手写汉字点阵图的分析,可提取手写汉字的基本笔划段:一、丨、丿和丶四种. 这四种基本笔划段形成四种基本笔划,其他笔划可以通过笔划的相似性或由此四种基本笔划组合来形成.这样,由这四种基本笔划可合成手写汉字的各种笔划. 一个笔划段 S_{se} 可以用笔划的起点 (x_s,y_s) 和终点 (x_e,y_e) 两个坐标表示为

$$S_{se}=\{(x_s,y_s),(x_e,y_e)\},\text{其中 } x_s\leqslant x_e \tag{1}$$

整个汉字可表示为所有笔划段的集合:

$$C=\{S_i\mid i=1,2,\cdots,n\} \tag{2}$$

2 基于笔顺的汉字识别

汉字的笔划段集合只是给出了组成该汉字的笔划段,而没有规定其笔划的顺序.传统基于笔划特征的汉字识别方法一般只对汉字的第一个笔划段进行处理,或对相同笔划进行统计,通过比较各基本笔划的数目来进行识别. 因此对于汉字笔划数目是一致的汉字,如"土"、"士"、"工"、"干"等,其笔划集合都为两横一竖,这些汉字就无法区分开来.

基于这种特点,本文提出一种新的方法,通过利用汉字的笔划段结合笔划的顺序(简称笔顺)来产生整个汉字的笔划结构特征. 利用标准字库的汉字笔划及笔顺形成标准模板,手写汉字通过笔划段的抽取后对各笔划段利用同样的笔顺形成规则产生对应于手写汉字的笔顺,比较手写汉字的笔顺与标准模板中的笔顺的相似度,选取相似度最大的模板汉字作为识别结果.

对应于汉字的笔顺我们可以通过已识别的汉字笔划段集合来形成汉字的笔顺. 判断两笔划段的先后顺序我们可以根据笔划段的交叉与不交叉的关系来进行判别.

1)交叉的笔划

设笔划段 $S_1=\{(X_{11},Y_{11}),(X_{12},Y_{12})\}(X_{11}\leqslant X_{12})$ 与笔划段 $S_2=\{(X_{21},Y_{21}),(X_{22},Y_{22})\}(X_{21}\leqslant X_{22})$,有无交叉可以根据 S_1 的两端点是否在线段 S_2 的异侧以及 S_2 的两端点是否在线段 S_1 的异侧来进行判定. 判定两点 $(x_1,y_1),(x_2,y_2)$ 是否在某一直线 $y+kx+b=0$ 的两侧判别如下:

$$(y_1+kx_1+b)(y_2+kx_2+b)>0 \tag{3}$$

利用式(3)可以得出判别 S_1 的两端点是否在线段 S_2 的异侧以及 S_2 的两端是否在线段 S_1 的异侧，若 S_1 的两端点在线段 S_2 的异侧且 S_2 的两端点在线段 S_1 的异侧则说明两线段交叉，具体表达式如下：

$$[(Y_{12}-Y_{11})(X_{21}-X_{11})-(Y_{21}-Y_{11})(X_{12}-X_{11})]\times$$
$$[(Y_{12}-Y_{11})(X_{22}-X_{11})-(Y_{22}-Y_{11})(X_{12}-X_{11})]<0$$
$$[(Y_{22}-Y_{21})(X_{11}-X_{21})-(Y_{11}-Y_{21})(X_{11}-X_{21})]\times$$
$$[(Y_{22}-Y_{21})(X_{12}-X_{21})-(Y_{12}-Y_{21})(X_{22}-X_{21})]<0 \qquad (X_{12}>X_{11}) \qquad (4)$$
$$(X_{11}-X_{21})\cdot(X_{22}-X_{11})<0 \qquad (X_{12}=X_{11}) \qquad (5)$$

当满足(4)或(5)的不等式时则说明笔划 S_1 与笔划 S_2 是处于相交的情况，则按笔划的横、竖、撇、捺的次序产生相应的笔顺.

2)没有交叉的笔划

相应笔划段 $S_1=\{(X_{11},Y_{11}),(X_{12},Y_{12})\}$ 与笔划段 $S_2=\{(X_{21},Y_{21}),(X_{22},Y_{22})\}$，$X_{11}+X_{12}-X_{21}-X_{22}\leqslant 0$ 在没有交叉的情况下其笔顺先后的判断可以依据笔划的中点位置进行判定，其判定公式为

$$-1<\frac{Y_{22}+Y_{21}-Y_{12}-Y_{11}}{X_{22}+X_{21}-X_{12}-X_{11}}\leqslant 1 \qquad (X_{11}+X_{12}-X_{21}-X_{22}\neq 0)$$
$$Y_{22}+Y_{21}-Y_{12}-Y_{11}<0 \qquad (X_{11}+X_{12}-X_{21}-X_{22}=0) \qquad (6)$$

若满足不等式条件则可认为笔划段 S_1 的笔顺要先于笔划段 S_2 的笔顺.

3)汉字笔划的顺序规则为：

①不交叉的笔划按公式(6)的规则产生(若两笔划只是相连则认为是不交叉的笔划)；

②交叉的笔划按公式(4)或公式(5)形成；

③使用冒泡法来产生整个汉字的笔顺.

具体步骤为：从剩余笔划集合中选取第一个笔划做为备选笔划与其他笔划相比较，若为相交的关系则按②进行判定其先后，若没有相交则按①进行判定. 若判定的结果备选笔划优先，则选取下一个还没有比较的笔划进行比较，否则，将比较笔划作为新的备选笔划继续进行笔划的比较过程，直到比较完最后一个笔划为止. 将备选笔划作为当前笔划的笔划，并从剩余笔划集合中删除该笔划，继续下一个笔顺的笔划的挑选过程，直到集合中只剩下一个笔划为止. 如汉字"生"根据其特征笔顺为：丿、一、一、丨、一.

这样，通过笔划的选取过程形成了整个汉字的笔顺过程. 令汉字的笔顺为 O，则

$$O=(S_1,S_2,\cdots,S_N) \qquad N \text{ 为汉字的笔划数目}$$

4)笔顺特征类似的汉字识别

手写汉字的笔顺与标准汉字的笔顺模板相比较，其笔顺一致的汉字匹配有可能得到多个结果. 在这种情况下，利用汉字的笔顺特征难以判定到底是哪一个汉字. 因此，我们必须借助于别的方法来进行判别. 由于汉字既包含结构特征又包含数值特征，同时对于笔顺特征相似的汉字其结构特征或数值特征相对较稳定，这样我们可以利用汉字的这些稳定特征来参与相同笔顺特征的汉字识别过程.

描述汉字的稳定特征为

$$T=\{(i,j,r)\,|\,r\in R\} \qquad (7)$$

其中，(i,j,r)代表笔划 S_i 与笔划 S_j 存在着 r 的关系；R 为两个笔划段之间的关系集合，如：大于、小于、相交、不相交、相离等.

这样，我们就可以将笔顺特征一致的汉字识别出来. 如"土"、"士"、"工"、"干"的识别我们首先利用笔顺的方法将"干"识别出来，而"土"、"士"、"工"则必须建立相应的特征关系如下

$$T_{\pm}=\{(1,2,\text{相交}),(1,3,\text{小于})\}$$
$$T_{\pm}=\{(1,2,\text{相交}),(1,3,\text{大于})\}$$
$$T_{\perp}=\{(1,2,\text{不相交})\}$$

利用汉字识别模板中的特征关系检测识别汉字是否满足其特征关系，若满足，则可以认为备选汉字即为识别的汉字；若不满足，则进行下一个相同笔顺结构的汉字的比较过程. 这样，通过汉字的特征关系就可以识别出笔顺相同的汉字，从而提高汉字的识别率.

3 笔划抽取失误汉字的识别

由于手写汉字的不规范，基本笔划可能发生变化，如横笔不平、竖笔不直、笔划发生倾斜等，这样可能导致汉字的匹配效果，因此必须对手写汉字的笔划与标准笔划进行某种加权平均后来进行比较. 我们可以根据待识别的汉字某笔划与识别模板的相应笔划之间的相似程度建立比较相似矩阵，矩阵的维数与汉字的笔划数目一致，而其相似程度我们可以根据已抽取笔划段 S_i 的位置来得出与匹配笔划段 S_j 的相似度 m_{ij} 如下

$$m_{ij}=\frac{(S_i,S_j)}{\|S_i\|\cdot\|S_j\|}=\cos(\alpha) \tag{8}$$

式中 (S_i,S_j) 为向量 S_i，S_j 之间的内积，$\|S_i\|$，$\|S_j\|$ 分别为向量的模，α 为向量 S_i，S_j 在 n 维空间的夹角. 当此两向量夹角为 0 时，$m_{ij}=\cos(\alpha)=1$，则两向量完全相似.

利用相似度的概念可得两汉字的相似度矩阵 M 为

$$M=\begin{bmatrix} m_{11} & \cdots & m_{1N} \\ \vdots & & \vdots \\ m_{N1} & \cdots & m_{NN} \end{bmatrix} \tag{9}$$

由于笔划比较的近似度只代表备选汉字与识别汉字之间的一种相互程度，因此汉字的相似度矩阵 M 应为一对称矩阵，其相似度矩阵 M 可变

$$M=\begin{bmatrix} m_{11} & \cdots & m_{1N} \\ \vdots & & \vdots \\ m_{N1} & \cdots & m_{NN} \end{bmatrix} \tag{10}$$

在考虑手写汉字的笔划没有丢失的情况下，该笔划的比较只与笔划的笔顺相同的进行比较，因此当 $i\neq j$ 时，其相似度值为 0，而 $i=j$ 时，其相似度值为 m_{ii}，对应相似度矩阵 M 可进一步简化为如下的对角矩阵形式

$$M=\begin{bmatrix} m_{11} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & m_{NN} \end{bmatrix} \tag{11}$$

因此汉字的比较就可以通过下面的相似关系的运算获取手写汉字与被匹配的汉字之间的相似程度 R

$$R=\sum_{i,j=1}^{N} m_{ij} \tag{12}$$

$R=N$ 时,即所有笔划都完全匹配(对应于前面所说的匹配方式实际上为相似的程度匹配的一种特例),也就是说,其手写汉字的笔划与笔顺完全一致,则认为匹配成功,所选取的汉字即为所要识别的汉字. 这种情况一般来说仅对应于采用与模板对象字体相同的印刷体汉字的识别而言会产生这样的效果,而手写汉字的识别其相似程度一般会要小于 N. 当匹配的结果在 $R\neq N$ 的前提下,选取最大的 R 作为所要识别的结果.

4　结果及结论

利用前面的方法从已有的汉字点阵库中,抽取其笔划,并按笔顺的产生规则,形成标准汉字笔顺模板,选取汉字笔顺模板中的相同笔顺中的汉字,手工输入其特征矩阵,建立特征关系. 在对印刷体汉字的识别试验中,其识别效果在印刷质量比较好的情况下其识别率在 95 %以上;在对相同笔顺的汉字的识别中,由于汉字的相同笔顺字比较多,为了取得区分相同笔顺字的实际经验,我们对 400 个高频字中具有相同笔顺字的字对,逐字进行判别试验,统计结果表明,多数字由于其特征关系比较容易确定,判别准确率在 90 %以上. 另外一些相同笔顺字由于其特征基本相似(如"厅"与"万"等)难以区分,这表明了本方法的局限性;这样的一些字可以采用前后相关的方法进行校正处理.

由于建立标准模板库采用了标准汉字的结构,同时匹配过程只有其笔段的倾斜度参与其匹配过程,而不涉及笔划的长短,因此,识别时受试者只要其输入笔划基本正确,不存在连笔,且没有增笔与减笔的情况下,其书写的字体与大小无关,则识别结果就能够得到保证,因而是一种汉字识别尤其是特征相似的汉字识别的好方法.

参考文献:

[1] 陈友斌,丁晓青,吴佑寿. 一种新的用于手写汉字识别的非线性规一化方法[J]. 模式识别与人工智能,1998,11(3):310－317.

[2] 崔怀林. 基于笔划特征的手写汉字分类与识别字典的构造方法[J]. 模式识别与人工智能,1998,11(2):228－232.

[3] 杨　凤,朱志祥,李志舜. 手写体汉字联机识别系统的研究与开发[J]. 模式识别与人工智能,1999,12(1):121－124.

[4] 曾棋荣. 手写汉字识别的研究[D]. 北京:清华大学,1989.