中 文 信 息 学 报
第5卷 第4期 JOURNAL OF CHINESE INFORMATION PROCESSING vol.5 No.4

一个手写印刷体汉字识别实验系统

陈 玲 陈学德 郑 重
（沈阳工业大学计算机学院）
青 木 由 直
（北海道大学工学部）

【摘要】 本文在充分考察了手写汉字和中国大汉字集特点的基础上，提出了一组用于手写印刷体汉字识别的分类特征，它们是长笔划分布类型、各类笔划的数目、交叉点数目和折点数目。利用这组特征进行匹配就可直接识别出GB2312—80汉字集中的绝大部分汉字，再通过一个基于知识的推理过程即可进一步识别出已被分成类组的少数剩余汉字，这种将统计分类与基于知识的推理识别相结合的两级识别方法具有较高的效率。一个适应性较强的汉字笔划和特征点抽取方法也被设计，它是SLSA方法的改进，与机器学习功能相配合，大大提高了特征抽取的正确率。我们根据上述思想建立了一个手写印刷体汉字识别实验系统，并获得了较好的实验结果。

一、引 言

汉字与其它文字相比，具有两个显著的特点。其一是结构复杂；其二是字的种类多，而且相似字多。特别是手写汉字，同一个字也会因书写者不同而产生明显的差异。这就使手写汉字的机器识别成为一项很难的研究课题。尽管日本学者已经对手写汉字识别做了大量工作，并且已经进入实用阶段[1]，但是由于中国所用的汉字集是日本汉字集的两倍到三倍（如表1所示），所以研究适合于中国大汉字集的手写汉字识别方法仍然具有十分重要的意义。

表1 汉字集及其大小

汉字集	大小(日本)	大小(中国)
初等教育集	881	2500
日常使用集	1850	3500
国家标准集	2669	6763
字典集	$1\sim1.5\times10^4$	$1.6\sim4.5\times10^4$

当待识的汉字集较大时，特征选择是至关重要的，所用的特征必须充分反映汉字的本质

本文1991年2月3日收到

特征，并且必须是稳定的；对于手写汉字，所用的特征抽取方法对不同书写者造成的差异必须具有较强的适应能力。本文总结了我们对这些问题的研究和实践，提出了一组用于手写印刷体汉字识别的分类特征，并用机器学习的方法增强了一个笔划抽取方法的适应能力，最后报告了我们的实验结果。

二、特 征 选 择

汉字是由不同数目的、不同类型和不同长度的笔划按某些特定规则组成的平面几何图形。因此，汉字字形具有笔划特征、几何特征和拓扑特征。

笔划特征：笔划是构成汉字的最小基元。一个汉字所包含的笔画种类及其数目都是汉字的本质特征之一，称之为笔划特征。汉字笔划类型及其描述如表2所示。

表2　汉字笔划类型及其描述

名称	代号	形　　状	方向
横	H	一	→
竖	S	丨	↓
撇	P	丿丿/	↙ ↗
捺	N	㇏丶	↘
折	Z	㇇㇋㇌㇆㇉……	↻
		㇄㇙㇜㇛㇟㇈㇂……	↺

几何特征：组成汉字的各笔划之间、各部件（具有一定关系的相对不变的一组笔划）之间以及笔划与部件之间都有着比较稳定的相对位置关系；各笔划之间也有着相对长度比和倾角。这些相对关系都是汉字的另一种固有特征，称之为几何特征。

拓扑特征：笔划之间可能形成各种特征点。每条笔划都有“端点”，折笔划上有“折点”，两条笔划相接形成“歧点”，两条笔划相交形成“交点”。称这些点为特征点，如图1所示[2]。它们反映了汉字的拓扑结构，是汉字的重要特征，称之为拓扑特征。

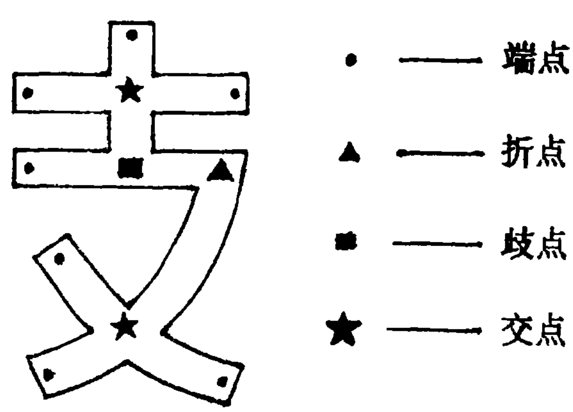

图 1　汉字的特征点

分析上述特征，我们发现它们具有下列性质：

1．稳定性

上述各种汉字字形特征，有些比较稳定，有些则不稳定。我们认为，对于手写汉字而言，笔划特征中的笔划种类及其数目，几何特征中笔画和部件之间的相对位置关系，以及拓扑特征中的交点和折点都是比较稳定的汉字字形特征。

2．重要性

各种特征在不同的汉字中的重要性是不同的。换句话说，不同的汉字各有其重要的特征。如对“王”字，三横的相对长短不影响识别，但对于“土”和“士”来说，两横的相对长短至关重要；而对“由”、“甲”和”申”来说，其重要的区别特征则交点的数目和中间长竖的上下出头情况。这类区别特征可以表示成知识。

对于手写汉字识别，识别特征无疑要选用稳定的特征；并且，预分类特征必须具有普遍的重要性，而细分特征则应是对应类组内待识汉字的重要区别特征。根据这个原则，我们选择了汉字的长笔划分布类型、每类笔划的数目、交点数目和折点数目构成的特征值作为预分类特征，再使用关于类组内汉字之间差异的知识做进一步的识别。

三、统计分类与基于知识的识别推理

为了对大汉字集进行分类，我们所用的特征值V如下：

$$V=(v_1,v_2,v_3v_4,v_5,v_6,v_7,v_8)$$

其中：

(1) v_1是长笔划分布类型。它在某种程度上反映了汉字的基本几何结构。设一个汉字模式由$n\times n$矩阵$f(i,j)$表示，

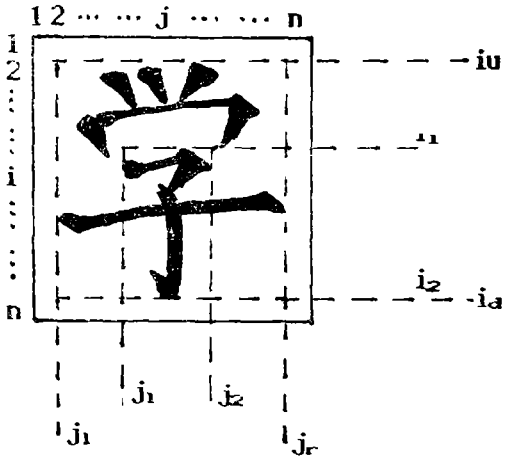

图2　汉字模式及区域

$$f(i,j)=\begin{cases}1,\text{当象元}(i,j)\text{为黑点},\\0,\text{当象元}(i,j)\text{为白点},\end{cases}\quad i,j=1,2,\cdots,n;$$

去掉全白点边缘，汉字模式的实际边界纵向上从i_u到i_d，横向上从j_l到j_r。又设任一条笔划t的轨迹所点的最小矩形区域是从i_1到i_2和从j_1到j_2。如图2所示。我们定义t是横向长笔划，当

$$(j_2-j_l)/(j_r-j_l)\geqslant l_0;$$

t是纵向长笔划，当

$$(i_2-i_1)/(i_d-t_u)\geqslant l_0;$$

t既是横向长笔划也是纵向长笔划，当

$$[(j_2-j_l)/(j_r-j_l)\geqslant l_0]\wedge[(i_2-i_1)/(i_d-i_u)\geqslant l_0];$$

t不是长笔划，当

$$[(j_2-j_l)/(j_r-j_l)<l_0]\wedge[(i_2-i_1)/(i_d-i_u)<l_0];$$

这里，l_0是一个特定的常量。

基于上述定义，规定

$$v_1=\begin{cases}0,\text{当无任何长笔划存在（排除了包围型和半包围型汉字）；}\\1,\text{当有横向长笔划，但无纵向长笔划（排除了左右型汉字）；}\\2,\text{当无横向长笔划，但有纵向长笔划（排除了上下型汉字）；}\\3,\text{当既有横向长笔划，又有纵向长笔划（排除了上下型和左右型汉字）。}\end{cases}$$

（2）v_2,v_3,v_4,v_5,v_6分别是横、竖、撇、捺和折五种基本笔划的统计数目。它们反映了一个汉字的基本笔划的组成信息。

（3）v_7,v_8分别是交叉点和折点的统计数目。它们在一定程度上反映了汉字字形的拓扑结构。

特征值V具有足够的分类能力。利用V在分类字典中进行匹配就可直接识别出GB2312—80汉字集中97%以上的汉字；剩余的少数汉字也被分类成几十个较小的类组，最大的一组是（半、本、未、末），也只包含四个字。

为了在小类组内继续识别，我们采用了关于类组内汉字之间差异的知识，利用知识的过程表示法[3]来表示这些知识。针对每个待识类组都有一个与之对应的知识过程，它将利用待识汉字的已知事实（特征抽取的结果之一）精确地识别出该汉字是该组中的哪个汉字。例如，针对类组（半，本，未，末）的知识过程为：

```
PROCEDURE KPexample(Code);
VAR
  t·i1,t·i2,t·j1,t·j2,t·i,t·j: t表示任意一条笔划，它纵向上从i1到i2、横向上从j1到j2，
      t·i=(t·i2-t·i1)/2, t·j=(t·j2-t·j1)/2;
  h1: 在上方的横笔划;
  h2: 在下方的横笔划;
  P: 撇笔划;
  n: 捺笔划;
BEGIN
  IF (h2·i1+h2·i)<n·i1
    THEN IF (h1·j2-h1·j1)≤(h2·j2-h2·j1)
         THEN Code:='97B2'{返回"未"的代码}
         ELSE Code:='92A7'{返回"末"的代码}
    ELSE IF (h2·i1+h2·i)>P·i1
         THEN Code:='895D'{返回"本"的代码}
         ELSE Code:='88E9'{返回"半"的代码}
END;
```

四、特征抽取与机器学习

分类所用的特征值V中分量v_2，v_3，v_4，v_5，v_6必须通过追踪抽取笔划才能得到；而v_1则

可由笔划所占的矩形区域求出；v_7，v_8也可由笔划之间的关系导出或在追踪抽取笔划的过程顺便求出。推理识别所用的事实多数是关键性笔划的区域。因此，这里的特征抽取问题主要是笔划抽取问题。

为了抽取笔划，传统的方法是先利用一定的细化算法对汉字原始点阵进行细化处理。这不但会抹掉有用的笔形信息，而且还会产生失真，以致最终的笔划抽取结果有很大误差，并且还要付出时间代价。近年来出现了不用细化而直接抽取笔划的趋势[4][5]。文献[4]曾介绍了一个将汉字图形看作由线状图形组成，而利用直线段的中心点进行逼近的不经细化直接提取笔划的SLSA法。该方法具有快速、不受笔划粗细影响等优点，但对笔划的弯曲变形还需利用对笔划的误差校正技术另行处理。我们对SLSA法进行了如下扩充和改进：

(1) 使之在抽取笔划的同时，抽取出交叉点和折点的数目，即同时获得特征值V中的v_7和v_8；

(2) 扩充了机器学习功能，使得SLSA实现程序中用到的一些控制和校正参数可以通过样本学习过程进行调整，提高了笔划分离和模糊笔划确认的正确率。

机器学习能够增强系统对识别对象的适应能力[6]。我们这里对机器学习功能的实现过程如下：

第1步[学习进程]

先用特征抽取程序以予置的标准参数对学习样本S抽取特征值V′，并要求用户告之该样本是什么汉字，这样就取得了特征值V′与对应汉字代码的对应关系（V′,C），将其加入学习字典；重复这个过程直至停止学习，设这时已对n个样本进行了学习，即学习字典中有n项。（实际上，用于对6763个国标集汉字进行预分类的标准字典就是这样建立的。）

第2步[总结过程]

对学习字典中的每一特征值V_i，

$$V_i' = (v_{i1}', v_{i2}', v_{i3}' v_{i4}', v_{i5}', v_{i6}', v_{i7}', v_{i8}'),$$

与其对应的标准特征值V_i（V_i可用V_i'对应的C_i在标准字典中查到），

$$V_i = (v_{i1}, v_{i2}, v_{i3}, v_{i4}, v_{i5}, v_{i6}, v_{i7}, v_{i8}),$$

求距离D_i，

$$D_i = (d_{i1}, d_{i2}, d_{i3}, d_{i4}, d_{i5}, d_{i6}, d_{i7}, d_{i8}),\ i = 1, 2, \cdots, n$$

其中：$d_{ij} = v_{ij}' - v_{ij}$，$j = 1, 2, \cdots, 8$。

对n个学习样本进行特征抽取的距离（总偏差）为D，

$$D = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8),$$

其中：$d_j = \sum_{i=1}^{n} d_{ij}$，$j = 1, 2, \cdots, n$。

D说明了特征抽取的偏差信息。例如，如果$d_2 = 0$，则说明横笔划的抽取基本正确；如果$d_2 < 0$，则说明将某些横笔划错误地确定成了其它种类的笔划，此时往往同时有$d_4 > 0$（说明可能将横笔划错认为撇笔划，亦即学习样本中的横常常向右上斜）或者$d_5 > 0$（说明可能将横笔划错认为了捺笔划，亦即学习样本中的横笔划常常向右下斜）；而$|d_2|$的大小则说明了偏差的程度。

利用D就可对特征抽取程序的调用参数进行调整，增强了对当前识别对象的适应性。

五、实验和结果

基于上述思想，我们建立了一个手写印刷体汉字识别实验系统。该系统使用ETSON GT—3000V Scanner作为输入设备，用PASCAL语言和8086/8088汇编语言在华立B16机上实现。目前，已对GB2312—80汉字集建立了分类字典和知识过程，并在国标集中随机抽取500字，每字书写6个模式，共3000个样本模式进行了实验。实验结果表明，特征抽取的正确率为92.7%；在特征抽取正确的条件下，分类和推理识别的正确率为100%；所以，总的识别正确率为92.7%。

参考文献

[1] 张炘中，中国汉字识别技术综论，第三届全国汉字及汉语语音识别学术会议论文集，1989、9

[2] 张炘中，闾昌德，汉字识别的特征点法及其一种应用，中文信息学报，Vol.1，No.3 (1987)，pp13—19

[3] 杨祥金，蔡庆生，人工智能，科学技术出版社重庆分社，重庆，1988，pp35—44

[4] 顾新理，汪燮华，线状图形的SLSA法提取，中文信息学报，Vol. 1，No. 3 (1987)，pp59—68

[5] H. Ogava, K. Taniguchi, Thinning and Stroke Segmentation for Handwritten Chinese Character Recognition, Pattern Recognition, Vol. 15: No. 4 (1982)

[6] 涂序彦，人工智能及其应用，电子工业出版社，北京，1988，pp206—227

An Experiment System for Recognition of Handprinted Chinese Character

Chen Ling, Chen Xuede, Zhengzhong

(College of Computer, Shenyang Polytechnic University)

Yoshinao Aoki

(Faculty of Engineering, Hokkaido University, Japan)

Abstract

This paper describes a study and experiment on handprinted Chinese character recognition. Firstly, it presents a group of stable and essential features of handprinted Chinese character which are the distribution type of long strokes, the type and the numbers of strokes and the numbers of intersections and corners. The features express very well some information on the structural type, the basic elements and the topologic critical points of Chinese characters so that they have great capacity and excellent suitability for recognizing handprinted Chinese characters. Secondly, it recommends a two-stage recognition scheme which integrates the statistical classification with the knowledge-based inference. Thirdly, it improves a method for the extraction of strokes and feature points of handprinted Chinese characters by a machine learning mechanism so that the accuracy of the feature extraction is raised. Last, it gives the experiment results.