联机手写体汉字联想识别系统

唐降龙 舒文豪 刘家锋 李铁才

(哈尔滨工业大学631教研室)

摘要 本文提出一种联机手写体汉字联想识别系统. 在单字识别的基础上，有分词信息的前提下，利用汉语词的前后约束及字的特征信息对识别结果进行后处理. 从而提高了联机识别系统的识别率.

关键词: 信息处理，汉字识别，识别系统

1 前言

随着汉字信息处理的飞速发展，对汉字输入的要求正趋向方便化和省力化. 由于汉字的特定结构，造成我们没有使用打字机的习惯，人们更多的期待汉字识别技术的发展，期待出现限制较小、识别率很高的汉字识别设备.

联机手写体汉字识别系统是汉字识别领域中首先达到实用的设备. 国内已出现的产品指标大致为: 识别字数: 6763个; 识别率: 一般在90%以上，字写的愈规范，识别率愈高; 识别速度: 基本满足人的书写要求，国外(主要是日本)也有多种同类产品，其指标除硬件指标外，均不如国内产品. 目前这些产品都有一个共同的缺点，就是书写受较多的限制，不允许过多的连笔. 因此，"无笔顺限制，可以允许习惯性连笔"的系统是联机手写体汉字识别的发展方向. 这一阶段的研究突破会大大地扩大联机识别系统的应用范围，使联机识别系统走向一个新阶段.

本文提出一种联想识别方法. 在有分词信息的情况下，利用汉语文本中汉字上下文的词约束关系，采用词联想方法对识别结果进行后处理. 通过词的信息、单个汉字本身的特征信息进行再识别. 从而提高识别率，降低书写汉字的限制.

2 系统组成

识别系统由计算机与HGD-9200-1型手写图文输入装置组成(图1).

2.1 HGD-9200-1型手写图文输入板工作原理及性能简述

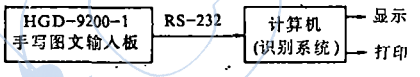

图1 系统组成

输入板是由我们自行设计和制造的. 该产品中国专利申请号为89102095. 它由四边形书写板(1)、导电笔(9)、笔信号采样电路(5)、双积分A/D转换电路(6)、数据输出电路(7)、驱动电路(10)和单片计算机(8)组成(见图2). 四边形书写板(1)的四边上均匀布置了间隔排列一条直线的点状电极(2)，并通过二极管对各点状电极进行隔离，形成 X 和 Y 电极，如图3所示. 当 X 电极(3)接高电平，Y 电极(4)接低电平，则左、右两边二极管导通; 上、下两边的二极管截止，于是书写板(1)的右边各点状电极均为高电平，左边各点状电极均为低电平，即在书写板(1)的表面形成均匀的 X 方向梯度电场. 反之，Y 电极接高电平，X 电极接低电平，在书写板

收稿日期: 1991—01—14，该课题为国家"七五"攻关项目.

(1)的表面形成均匀的 Y 方向梯度电场. 当导电笔(9)与书写板相接触时，书写板(1)的电场电压通过导电笔(9)传导到笔信号采样电路(5). 为了区别对 X 坐标与 Y 坐标采样，单片计算机(8)控制驱动电路(10)，按规定的周期改变 X，Y 电极的电平. 若采样时间与 X，Y 电平变化时间对应，就可以实时地把 X，Y 坐标采样结果按规定的周期送双积分 A／D 转换电路(6)进行变换. 然后经单片计算机(8)将数据处理成规定的数据格式，最后由数据输出电路(7)输出.

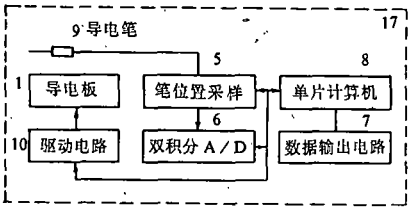

图 2 输入板组成

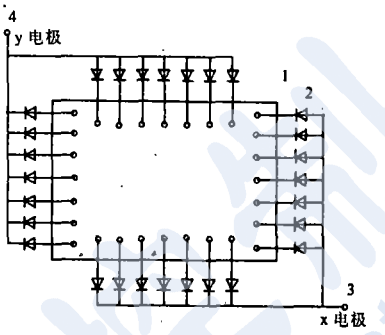

图 3 书写板

书写板书写面积：$115\times80mm^2$

分辨率：水平方向(X 轴)$<0.1mm$，垂直方向(Y 轴)$<0.08mm$

数据传送方式：通过 RS-232 串行口输出

输出速率：9600Bit／s

数据格式：二进制方式，每一个坐标点占 5 个字节.

2.2 单字识别系统软件

系统识别软件流程图如图 4 所示.

2.2.1 抽取识别基元

汉字是由笔划组成的，而笔划则由一些笔段组合而成. 书写汉字时由于一些习惯性连笔而造成汉字笔划数特征很不稳定. 但笔段数是相对稳定的. 因此，我们取一、丨、丿、㇏四种基本笔段作为识别的基元，将 X-Y 坐标平面分成 5 个方向码区(见图 5). 对输入汉字的数据序列求导数，使数据序列转换成方向码序列. 方向码序列虽然包涵了笔段特征，但也夹杂有多余信息，多余信息主要是由书写不规则和连笔造成的，因此，在抽取笔段前对方向码序列进行以下处理.

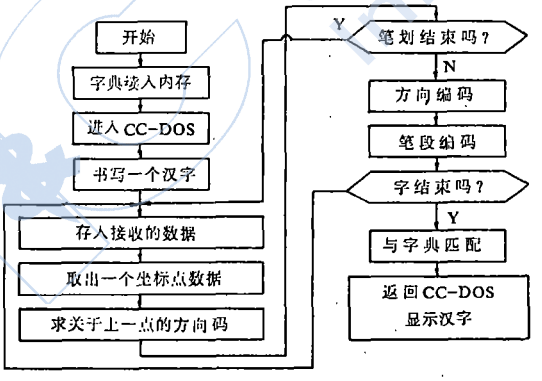

图 4 系统识别软件流程图

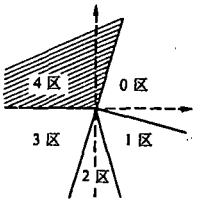

图 5 X-Y 坐标平面的 5 个方向码区

(1) 去掉首尾方向码；

(2) 移去由噪声所产生的方向码;

(3) 方向码归正;

(4) 去掉4，5区方向码.

经过上述四个子程序处理，就可从方向码序列中得到一、丨、丿、㇏四种基本笔段.

2.2.2 偏旁部首粗分类

对于一个能够识别6763个汉字的系统来说，如果用一个统一的算法来达到一次识别成功是不可能实现的，一般通过几次分类使识别范围缩小. 因此，一个好的分类方法不但可以提高识别速度，而且由于需要细分的汉字范围小，从而也就提高了书写的自由度.

通过分析汉字的位置结构不难发现具有稳定结构的汉字占6763个字中的80%以上，而剩余的20%汉字，可以人为地找出其固定结构作为分类依据. 基于上述汉字分类思想，通过偏旁部首粗分类，该系统通过这一分类使每类汉字不超过200个.

2.2.3 笔段向量、笔段位置细分类

经过粗分类使每类汉字数不超过200个，通过待识汉字的笔段向量特征及笔段间的相对位置特征，与字典特征匹配得出识别结果.

3 词组联想识别

前面我们讨论了单字识别方法，对一些相似的字就要求书写规正，如工与土相似，但构成词后，工人、工作、工厂与土地、土壤则相差很远. 如果我们对单字识别后再对词组识别，必然会大大提高识别率.

以双字词组为例说明词组联想识别方法. 设：字典模板特征与待识前字特征匹配距离小于 T 的集合为 FR；字典模板特征与待识后字特征匹配距离小于 T 的集合为 BR；T 为匹配距离阈值；双字联想词组全集为 W；前字联想词组集合为 FL；后字联想词组集合为 BL.

对于 $fr_i \in FR$，至少存在一个 $br_j \in BR$，使得 $fr_i\ br_j \in W$. 对于 $br_i \in BR$，至少有一个 $fr_j \in FR$，使得 $fr_j\ br_i \in W$.

双字识别有四种可能结果，我们可以有三种处理方法.

① 双字均正确识别

$fr_i \in FR$，$br_j \in BR$，因为 $fr_i\ br_j \in W$，所以 $fr_i \in FR \wedge FL$，$Br_j \in BR \wedge BL$

双字匹配距离 $DW1 = DC(fr_i) + DC(br_j)$.

DC 为单字匹配距离，取 $DW1$ 最小的双字为识别结果.

② 前字正识、后字误识

$$fr_i \in FR,\ br_j \notin BR,\ br_j \in BL$$

设：br_j 的特征向量为 $\bar{\theta}_{bj}$，BR 的特征向量为 $\bar{\theta}_b$，则 br_j 与 BR 的相似度为:

$$DR(br_j) = D(\bar{\theta}_{bj},\ \bar{\theta}_b)$$

双字匹配距离为

$$DW2 = DC(fr_i) + DR(br_j)$$

根据 $DR(br_j)$ 来认为 br_j 是 BR 中的一个字，取 $DW2$ 最小的双字为识别结果.

③ 前字误识、后字正识

$fr_i \in FR$，$br_j \in BR$，$fr_i \in FL$，同前一种情况类似，根据 $DR(fr_i)$ 对 fr_i 进行再识别取

$DW3=DR(fr_i)+DC(br_j)$最小的双字为识别结果.

④ 双字均误识

由于没有正确的词组信息，无法进行词组联想.

由于无法预先知道属于哪一种情况，必须同时求得 $DW1$，$DW2$，$DW3$，其中距离最小者为词组联想词组识别结果.

综上所述，词组联想识别方法利用了识别单字过程与词组联想过程的不相关性，综合考虑字的信息和词组的约束信息，很好地弥补了单字识别过程中的误识.

4 实验结果与讨论

(1) 经过 10 人次分别书写 1000 字文章实验，识别率大于 97%.

(2) 增大了书写自由度.

(3) 由于需要大量的词组作识别依据，因此需要较大的存储空间.

参 考 文 献

1 唐降龙等. 联机手写体准行书识别中的笔段抽取及粗分类. 哈工大70周年校庆优秀论文集，1990. 6

ON-LINE ASSOCIATING RECOGNITION SYSTEM FOR HANDWRITTEN CHINESE CHARACTERS

TANG Xianglong SHU Wenhao LIU Jiafeng LI Tiecai

(Harbin Institute of Technology)

Abstract

An on-line associating recognition system for handwritten Chinese characters is described in this paper. On the basis of single character recognition and phrase information, the recognizing result is postprocessed by using relations between characters and phrase. Recognizing rate of this system is higher than others of a kind.

Keywords: information processing, Chinese character recognition, recognition system