清华联机手写汉字识别系统—中文人机接口的智能化

杨泽红　　夏莹

清华大学计算机系人工智能教研组（100084）

摘要：本文介绍我们在联机手写汉字识别研究中所取得的成果以及系统实现中各个环节所采用的主要算法思想。在特征选择时，提出了“笔段作为联机手写汉字识别的主要特征”的方案；在特征的抽取过程和匹配过程中，利用模糊原理思想，提出了一系列适应手写汉字不稳定性和特征非确定性的模糊算法并在系统中得到最终实现。系统的良好性能证明了模糊原理在汉字识别研究中的特殊作用。

On-Line Recognition System for Handwritten Chinese Characters ----Intelligent Interface between Man and Machine

Zehong Yang, Ying Xia

Computer Department, Tsinghua University 100084, Beijing, China

Abstract: This paper gives the introduction to our research achievements on the recognition for on-line handwritten Chinese characters and the main algorithm in our recognition system. The stroke segments of a handwritten Chinese character are selected as main features for recognition. And the fuzzy theory is utilized in actual feature extracting and matching. It is proved that the fuzzy theory is effective in handwritten Chinese characters recognition.

前言

汉字的输入一直是中文信息处理系统的瓶颈问题，虽然各种各样的编码输入方案已经在实际应用中发挥了很大的作用。但是，从汉字信息处理系统和人们的使用要求来看，编码输入没有从根本上达到人们对汉字输入的期望。从使用者的角度看，一种不需要额外的学习和训练，不需要记忆额外的信息，只要象正常的书写或阅读就能完成汉字输入的方法是最理想的。

随着计算机技术、人工智能技术、模式识别以及相关学科的发展，人们提出了汉字的另一种输入方式：汉字的识别输入。严格上说，汉字的识别根据实现原理的不同又分为两

类：汉字字形识别和汉字语音识别。汉字的识别输人的目的是：使用者不需要额外的学习和训练，只要会写会念，就能实现汉字的输人。目前，汉字的识别输人成为中文信息系统实现人机接口智能化的有效途径。

识别系统实现方案

本文将对我们近期完成的联机手写汉字识别系统做全面的介绍。图1是联机手写汉字识别系统的框图，下面介绍系统中各功能的实现方案。

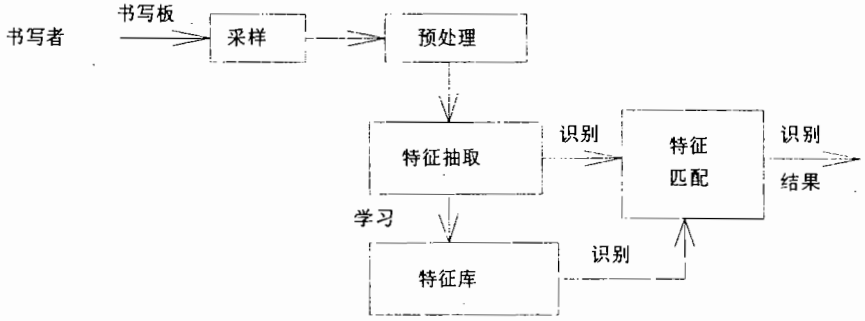

图1　　联机手写汉字识别系统框图

1. 书写轨迹信息的采集

汉字书写轨迹的采集是联机汉字识别的基础，可以有多种方式。一般情况下，汉字的书写是在一块图形输人板（简称书写板）上进行，轨迹信息是通过串行或并行通讯口进人计算机。所以，书写板的好坏将直接影响书写速度、书写效果乃至识别率。在我们的系统中，目前连接了三种书写板，均采用串行通讯口进行通讯。系统的采样模块主要完成如下功能：

（1）初始化通讯口：设置数据传送波特率、奇偶校验、数据位、停止位等通讯协议。

（2）初始化书写板：通常的书写板都有一系列控制命令，通过控制命令来初始化书写板，设置坐标数据的格式，使之适合系统要求。

（3）设置中断向量：由于采样是实时进行的，所以应该采用中断通讯方式，为此，需要设置相应的中断向量。

（4）采样处理：实时读取来自书写板的书写轨迹信息，并根据设定的数据格式进行数据转换，成为系统所需的简单的数据表示形式。这也是中断处理程序要做的工作。

2. 预处理

来自书写板的手写轨迹信息，不能直接用于识别，因为它包含有各种干扰和噪声。在联机手写汉字识别中，干扰、噪声主要来源于人手的抖动、笔的速度变化、书写板的量化噪声、感应噪声等。去除这些干扰、噪声的处理就是联机手写汉字识别系统的预处理任 .

务。预处理功能应该包括字符分割、平滑、去噪声（飞点）、空间采样、规范化等。下面介绍我们系统中所使用的预处理方法。

2.1 字符分割

字符分割是指区分哪些笔划是同一个汉字的笔划，关键在于找出一个字的书写结束。我们在书写板上设置固定的结束码区，每写完一个字，就用笔触结束码区，系统把结束码之前的笔划当作同一个汉字的。这种方式简单、稳妥，但比较费时。不过，只要养成习惯，这是一种比较好的方法之一。

2.2 去噪声（飞点）

联机手写汉字识别系统的噪声主要来源于手的抖动和感应噪声（或变形噪声）。首先，手写汉字由于人手的抖动而产生的不稳定主要在于落笔和抬笔阶段。所以我们在笔划的轨迹信息中，删除最前和最后的几个坐标点，以消除落笔和抬笔时人手抖动所带来的噪声。其次，对于书写过程中的孤立噪声点，我们规定一个噪声阈值D_{max}，用公式(2.1)求出任一点与相邻点间的距离。

$$\begin{cases} \Delta D_- = \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2} \\ \Delta D_+ = \sqrt{(X_{i+1} - X_i)^2 + (Y_{i+1} - Y_i)^2} \end{cases} \qquad (2.1)$$

当一个点(Xi,Yi)与相邻两点(Xi-1,Yi-1)、(Xi+1,Yi+1)之间的距离大于或等于噪声阈值时，认为该点是噪声。

对于非孤立点噪声，在联机手写汉字识别系统中很少出现（除非书写板有问题），所以系统中不予考虑。

2.3 离散坐标点的连续化处理

离散坐标点的连续化处理，是我们提出的“拐点的模板匹配检测法”所要求的。由于人的书写速度不定的，同一笔划的书写也不是均匀的，所以，从书写板接收到的离散坐标点是不均匀的。而用模板匹配法检测拐点时，要求轨迹坐标点是均匀的，否则，测出的拐点是没有意义的。所以，我们需要把接收到的离散坐标点进行均匀化处理，得到一条连续的离散坐标点笔划段。（如图2）

2.4 规格化处理

对于每个手写的汉字，其大小是不一样的。规格化处理就是通过一定的比例变化，使手写的汉字具有统一的大小规格，从而使特征的匹配运算变得简单、一致。我们采用公式(2.2)做规格化处理，它使任意大小和形状的手写汉字按比例充满整个S×S点阵，成为方块状，这与宋体汉字的形状相一致，弥补了手写汉字的形状上的差异。

（其中（X1,Y1),(X2,Y2),...,(Xn,Yn)是手写信息的N个采样点坐标）

$$\begin{cases} X'_i = (X_i - X_{min}) \times S / (X_{max} - X_{min}) \\ Y'_i = (Y_i - Y_{min}) \times S / (Y_{max} - Y_{min}) \end{cases} \quad (2.2)$$

式中，i=1,2,...,N

Xmax=MAX(X1,X2,...,Xn)，Xmin=MIN(X1,X2,...,Xn)

Ymax=MAX(Y1,Y2,...,Yn),Ymin=MIN(Y1,Y2,...,Yn)

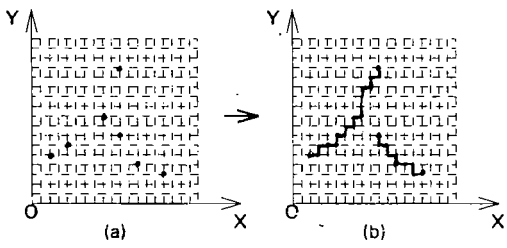

图2 离散坐标点的连续化处理

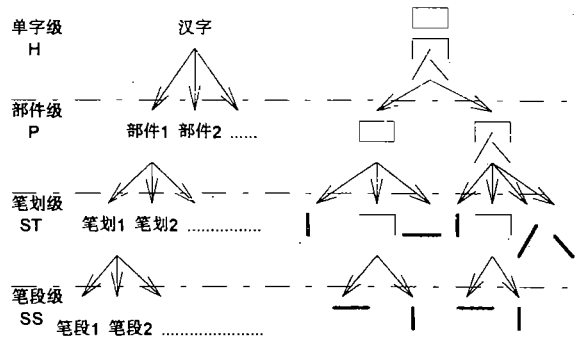

图3 汉字层次结构分析图

3. 特征选择和特征抽取

特征的选择和抽取是识别系统的核心之一，经过大量的分析和实验，我们提出了“汉字笔段作为联机识别主要特征”的识别方案并在特征的抽取过程中使用了快速有效的模板匹配检测法。

3.1 特征选择

对手写汉字的简单分析认为，汉字由偏旁部首（简称部件）按一定的结构关系组成，部件的划分各不相同，综合认为大约有几十到几百种独立的部件。部件又是由笔划按一定的结构关系组成，笔划还可分解成为一系列相关的笔段。(如图3)

理论分析以及大量的统计实验均表明，在汉字的联机书写过程中，笔段比笔划更稳定。在许多连笔书写方式下，笔划变动很大，但笔段特征得到了很好的保持，我们称之为汉字笔段的稳定性。同时，汉字笔段的另一特点是简单性。从笔段的端点，我们可以得到笔段的方向、长度等结构信息，所以，端点坐标就是汉字笔段的全面而简单的描述。而且，端点坐标特征对于相似字的识别是非常关键的（如“人”与“入”）。基于汉字笔段的两大特点，我们提出把笔段作为联机手写汉字识别的主要特征。

3.2 笔段特征的抽取

汉字笔段抽取的关键在于拐点的检测。笔划中的拐点是由于笔划书写过程中书写方向的变化而产生的，根据书写笔划的方向和方向的变化来检测拐点自然就成为通常使用的一种方法。这种检测方法有严密的理论支持，但在实际应用中显得比较复杂，计算难度大。

为了能简单有效地抽取笔段，我们提出了一种新的拐点检测方法——模板匹配法。模板匹配法求拐点的原理是：以平面上点的坐标为基元，利用平面上点、线之间的几何关系，用线段的比值近似表示角度，从而得到曲线上各点相对于一定曲线段的端点的弯曲度的近似表示，这样，用简单的加、减、乘、除运算替代了复杂的方向导数和导数变化的求解。

设平面上一条曲线段$A\hat{C}B$，C是曲线段$A\hat{C}B$的中点（即$A\hat{C}B$拉成直线段后，直线段ACB的中点）。为了避免复杂的方向导数的求解运算，我们用线段之间的比值来表示方向的变化。设A、B、C点的坐标为（X_A，Y_A）、（X_B，Y_B）、（X_C，Y_C）。用直线连接A、B两端点，求出直线段AB的中点D（X_D，Y_D）。连接C、D两点，用下列公式近似表示C点相对于A、B两端点的方向变化（这里简称为C点的曲度$|\hat{C}|$）：

$$|\hat{C}| = \frac{|CD|}{|AD|} \qquad (3.1)$$

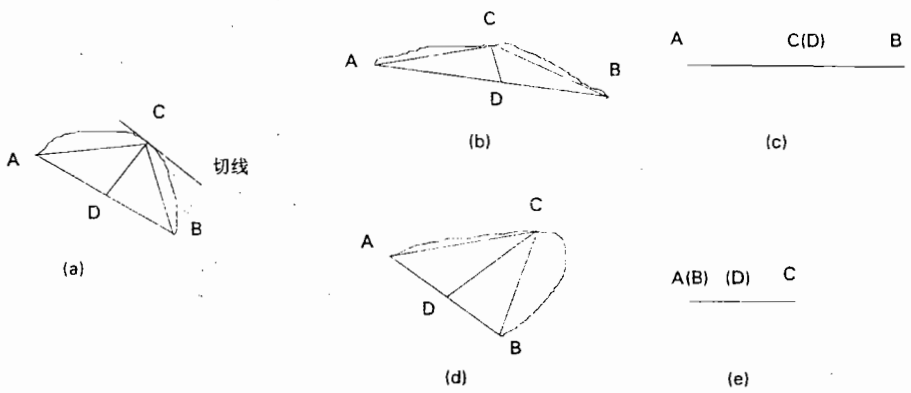

图4　曲线段中点曲度示意图

4. 特征匹配

通过前面的分析，我们把汉字的笔段特征作为识别的主要依据。在系统实现中，考虑到搜索空间以及汉字的端点拐点数的相对稳定性，把汉字的端点拐点数作为辅助特征。另外，利用汉字的笔段坐标的平均值（简称汉字笔段的重心）来描述汉字外形的相对稳定性。得到了这些特征后，从识别效率及实用效果考虑，我们提出了特征的“模糊化”和“模糊匹配”思想。

对于汉字的识别来说，并非所有笔段都起着同等重要的作用。不同的汉字可能具有某些相同的笔段，所以，并不是汉字的所有笔段对改汉字的识别都有意义。因此我们只要抓住对汉字识别起关键作用的笔段，就抓住了汉字识别的关键。但是，对每个汉字都要找出

对其识别起关键作用的所有笔段（简称关键笔段集），是很困难的，对汉字识别而言也是没有必要的（因为手写汉字本身的不稳定性）。为了降低特征匹配的复杂性，同时容忍某些次要笔段的一定程度的变形（连笔、断笔及笔顺变化等），我们提出汉字的模糊关键笔段集作为汉字笔段特征的近似描述（模糊关键笔段集实际上是一个汉字的不完全的近似的关键笔段集）。这样，在笔段特征的匹配运算中，只选择模糊关键笔段集中的笔段，而忽略模糊关键笔段集之外的次要笔段，从而，通过较少的关键性笔段信息表达出汉字的主体结构。在系统中取得良好的效果。

在特征的匹配运算中，考虑到特征的不稳定性以及每个特征对汉字识别的重要程度不同，各个特征都存在一个表示该特征重要性和可信度的模糊因子q，实现特征的模糊匹配。模糊因子q的确定，最有效最直接的办法就是实验统计。从大量的实验中找出满意的可以比较准确体现特征的贡献和可信度的因子。除了模糊因子在一定程度上解决笔段特征的非确定性以外，系统实现中还提出模糊匹配空间概念以解决一定笔划书写范围内笔顺的不确定性问题。

系统测试

我们从不同年龄、不同性别、不同职业的100份手写测试样张中任取10套进行测试，测试结果比较满意:

测试范围: 6763个国标一二级简体汉字;

识 别 率: 一选最高96.2%，最低80.5%，一选平均85.5%

十选最高99.1%，最低91.3%，十选平均95.0%

识别速度: 0.25秒/字（486微机）

内存空间: 小于30KB

扩展内存: 约350KB

参考文献

[1] Nouboud, F.; Plamondon, R.," A structural approach to on-line character recognition: system design and applications", International Journal of Pattern Recognition and Artificial Intelligence Vol: 5 Iss: 1-2 p. 311-35, June 1991, Singapore.

[2] 夏莹等，“联机识别自由书写汉字的方法和系统”，中文电脑国际会议'94，新加坡，1994

[3] 唐降龙等，“基于知识的联机手写汉字部件抽取”，模式识别与人工智能会议，1991 P292-295

[4] 曲维光等，“联机手写体汉字识别系统中部件识别的引人及其实现”，模式识别与人工智能，1991 P296-299

[5] 杨泽红、夏莹等，“联机手写汉字基于关键笔段特征的模糊匹配识别方法”，第五届全国汉字及汉语语音识别学术会议论文集，1994.9 成都

[6] 林颉、顾小凤，“联机汉字识别中汉字的表述及字典的建立”，第五届全国汉字及汉语语音识别学术会议论文集，1994.9 成都