教计算机认字的技术
——汉字识别

● 刘　群（中国科学院计算技术研究所）

对于中国人来说，使用计算机一个主要的障碍就是汉字的输入问题。目前这个问题已经在一定程度上得到了解决。从总体上看，汉字输入技术可以分为三类：键盘输入、语音输入和汉字识别。

所谓汉字识别，简单的说，就是将汉字从图像形式变为计算机内部编码的过程。

从学科分类上说，汉字识别属于模式识别学科领域，属于超多类模式集合的分类问题，是文字识别领域中最为困难的领域之一，涉及模式识别、统计学、图像处理、模糊数学、信息论、人工智能、人工神经元网络、语言学等多门学科。

汉字识别实际上是一类技术的总称。从输入方式上分，汉字识别可以分为联机（on-line）输入和脱机（off-line）输入两种方式。联机输入方式又称为笔输入方式，脱机输入方式主要是扫描输入方式。从识别的对象上分，汉字识别又可分为手写体汉字识别和印刷体汉字识别。由于笔输入方式所输入的只能是手写体汉字，而印刷体汉字只能通过扫描输入，所以汉字识别技术可以分为三大类：印刷体汉字识别，联机手写体汉字识别，脱机手写体汉字识别。

历史与现状

国外的西文文字识别研究始于50年代，我国的汉字识别研究起步较晚，大约从70年代末才开始汉字识别技术的研究。但我国的汉字识别研究发展非常迅速。尤其是进入90年代以来，很多汉字识别软件进入了市场，一些国际大公司也凭借其雄厚的资金实力和软硬件优势在这方面开展研究开发和市场运作，更加剧汉字识别市场的竞争。

印刷体文字识别又称为OCR（光学字符识别），所以印刷体汉字识别软件也简称中文OCR软件。目前我国市场上已经有了多个商品化的中文OCR产品，其性能指标已达到很高的水平（识别率一般都达到了98%以上），完全达到了实用的水平。

衡量一个中文OCR软件的技术指标一般包括以下几个方面：

● 识别率：现有系统的识别率一般都在98%以上，对于印刷质量好的文本甚至可以达到99%以上；另外还要考虑系统的拒识率和误识率，拒识率和误识率都是越低越好，但这二者又是互相矛盾的；

● 识别速度：现有系统一般都在20-30字/秒左右（Pentium级微机）；

● 识别字数：现在的系统一般都能识别国标一级和二级汉字；

● 多字体（包括简繁体）、多字号的识别能力；

● 中英文混排的汉字识别能力；

● 版面分析和还原能力；

● 系统鲁棒性，即对于不同的字体字号以及印刷质量的适应能力。鲁棒性较好的系统，其识别率不至于随着印刷质量的下降而过于迅速地下降。

对于一个实用的印刷体汉字识别系统来说，版面分析与恢复的能力、多字体多字号混排和中英文混排的识别能力也是非常重要的。我国的研究工作者在这几方面进行了很多研究，使系统的能力都有了较大的改进，但离真正好用的目标还有一定的距离。

对于脱机手写体汉字识别来说，由于自由书写的随意性太大，识别过于困难，人们把精力主要集中在手写印刷体（即书写工整的楷书手写体）汉字识别的研究上。目前我国有的实验系统对于比较工整的非特定人手写汉字文本，识别率已达到95%以上，技术上已经达到了初步实用化的条件。

笔输入方式可能是对普通用户最自然的一种汉字输入方式了:不用学习，不用记任何编码，发音不准也没有关系，只要你会写汉字，就可以将汉字输入计算机。因而，联机手写体汉字识别具有巨大的市场潜力。

目前,我国市场上的汉字笔输入产品有近十种之多,市场占有率较高的有中科院自动化所汉王公司的汉王笔、台湾蒙恬科技的蒙恬笔、摩托罗拉公司的慧笔、清华文通公司的文通笔和清华紫光公司的紫光笔等等。

基本原理

不同类型的汉字识别技术原理不尽相同。

印刷体汉字识别的输入是一个页面的完整的二维图像,这个图像可以是二值的黑白图像,也可以是多值的灰度图像。

一个印刷体汉字识别系统,其识别过程一般包括以下几个阶段:

1.预处理:由于印刷质量不好,或扫描输入时处理不当,输入的图像往往存在一些问题,如倾斜、噪声(如无效的黑点)等。预处理包括对输入的图像进行倾斜校正,去除噪声等操作。另外,为了使后面的识别操作更加有效,还要对输入图像进行一些函数变换;

2.版面分析:将图像中的标题、正文、公式、图像、表格、花边等区域分割开,并分别进行不同的处理,对于文本段落要确定文本的排版顺序(横排还是直排),如果需要版面恢复的话，还要确定文本的字体和字号,对于表格要调用专门的表格处理程序确定表格线和每一格的内容;

3.文字切分:将大段的文本切分为单个的汉字,一般是先切分为行(对于直排来说是列),再切分为字;

4.文字识别:识别出单个的汉字,对于汉字和英文字母分别调用不同的识别程序;

5.后处理:利用语言学和其他一些经验知识对识别结果进行校正。例如某一个字识别得到的候选字有"大"和"太",如果后面一个字识别的结果是"阳",那么这个字是"太"的可能性就非常大。

汉字识别的方法大致可分为句法模式识别(又称结构模式识别)、统计模式识别和人工神经元网络等几类。句法模式识别方法直接利用汉字的结构,首先识别出构成汉字的最小基元(笔划或笔段),利用这些基元以及基元之间的关系识别出更大的结构(如字根或偏旁部首),最后再识别出整个汉字。这种方法的优点是反映了汉字的自然结构，不受字体变化的影响,区分相似字比较简单。但在印刷体汉字识别中,由于得到的文本图像都存在不同程度的干扰和变形,很难准确地得到构成一个汉字的所有基元并准确判断它们之间的关系,系统的抗干扰能力较差,导致实际的识别率往往很低。目前,纯粹的句法模式识别方法在印刷体汉字识别中几乎不再有人采用。

统计模式识别方法是将待识别的汉字点阵图像作为一个整体，根据这个整体进行一系列的统计运算,得到关于这个汉字的一组特征,再将这些特征与字典中已知的汉字特征进行比较,得到最为相似的汉字。统计模式识别方法的关键在于统计特征的选取。科学家研究了很多种统计特征用于汉字识别,这些特征往往各有不同的优缺点,实际应用中一般要采用多种特征进行综合考虑。统计模式识别的特点是算法比较简单,容易实现,抗干扰能力强,缺点是细分能力较弱,区分相似字的能力较差。

人工神经元网络(简称 ANN)是一种模拟人脑神经元细胞结构的一种网络。它由大量的"神经元"组成,每个神经元都执行比较简单的操作,但神经元之间的联接非常复杂,神经元网络作为一个整体可以实现非常复杂的功能。人工神经元网络在一定程度上模拟了人脑的结构，与传统的冯·诺依曼结构计算机有很大的不同,具有自组织、自适应、非线性等很好的特点。但由于汉字识别问题的复杂性,如果将 ANN 直接应用于汉字识别,会导致 ANN 规模过大,时间和空间的消耗都难以接受。

以上几种方法都有各自不同的优缺点,现在很多研究者都开始把这些方法互相结合,取长补短,取得了很好的效果,这也是今后汉字识别研究的一个趋势。

脱机手写体汉字识别的输入和印刷体汉字识别一样,也是一个二维图像,识别的基本原理和过程也大致相同,但所处理问题的侧重点却有较大不同。脱机手写体汉字识别一般不用处理很复杂的版面,但汉字的切分和单个汉字的识别却比印刷体汉字识别复杂得多。特别是自然书写的汉字,由于书写随意性太大,识别非常困难。

联机手写汉字输入的输入设备是一支笔和一块手写板。手写板可以感知到笔的运动轨迹,并将这些轨迹数据传递给计算机。现在有些手写(下接 19 页)

让您获得多个不同用途的账号

您或许会问，既然我能够在网页上浏览，那自己肯定有一个E-mail账号，那又何必多此一举呢？您之所以“多此一举”，是因为有了转信邮箱的话，除了上述理由之外，您实际上又多了一个E-mail账号，您可以用它来处理种种“另类”邮件，比如说那可以是您的交友联络邮箱，你可以拥有几个账号用于不同的用途。

让您显示网虫风采

我们常常会碰到一些刚刚上网的朋友，当他看到我们的名片时，都要问一句“您的E-mail地址怎么这样简单”？那时我们就会很得意地回答，就是这样简单好记，当然也很好用。当听到经常泡网的朋友告诉我，他的E-mail地址是XXX@126.com，这时，我也会充满自信地说，“哦，我在126.com也都申请了一个，我的地址是YYY，域名跟您一样。”几乎所有的“网虫”都起码拥有一个免费转信站账号，也可以这么说，使用免费转信站账号已成为“网虫”的习惯。

您了解转信站的特点与优点了吗？您也想拥有一个转信邮箱了吗？现在就到http://www.126.com去申请一个属于您的转信邮箱吧，这个集电子邮件、自动回信、转信功能于一体的免费电子邮件服务，是能真正显示“网虫”身份与个性的电子邮件地址。

（责任编辑：朱婷婷）

（上接12页）板可以和显示屏幕做在一起，用户的感觉就好像在屏幕上直接书写，非常直观。还有的手写板采用了压感技术，能够感知用户运笔的轻重，不仅接收的信息更为丰富，而且用户的感觉更加接近于使用真实的笔。

与脱机汉字识别不同，联机手写汉字识别的输入不是一个单纯的二维图像，而是一个随时间变化的坐标序列。由于增加了一个时间维信息，因此可以获得许多对于汉字识别十分重要的信息，如笔的起落、笔划、笔顺、笔速等。由于联机手写汉字识别可以比较容易地获得汉字基元信息，因此比较适合于采用句法模式识别的方法。汉字识别的基元选取一般有笔划和笔段两种。笔划大家都很熟悉，根据划分的粗细不同分为几种到几十种，而笔段是指构成笔划的小直线段，一般根据其方向可分为四类或八类，是比笔划更小的构成单位。这两种汉字基元都有系统采用。

手写体汉字识别的困难主要在于用户书写的随意性，如笔顺变化、连笔、偏旁部首之间距离过大、偏旁部首的比例不合适、字写得不够端正（有一定倾斜）、笔划的变形（如一竖的末尾多了一钩）、笔划的缺失等等。评价一个联机手写汉字识别系统好坏的主要依据就是系统的鲁棒性，也就是系统适应用户书写随意性的能力。现有的系统在这方面各有千秋，有的比较能适应笔顺的变化，有的则对连笔的处理较好。但总的来说，对于自然书写的汉字识别率都还不是太高。这也是联机手写体汉字识别今后主要的发展方向。

应用展望

随着我国信息化建设的迅速发展，汉字识别技术获得了越来越广泛的应用。

我国的《四库全书》是迄今为止人类历史上最大的百科全书，它是清代乾隆年间历经十年编纂而成，收书三万六千余册，七万九千余卷，约八亿字，包容历史、文学、哲学、社会科学、自然科学等各方面内容。最近，我国已将《四库全书》的全部内容录入计算机，制成原文及全文检索版。如此巨大的工作量，仅仅几百人，在短短的三年内就全部完成。如果没有汉字识别技术的帮助是不可想像的。

现在市场上出现了很多中文的个人数字助理（PDA）产品，也就是一台手掌大小的计算机，由于可以随身携带，为人们的生活带来了极大的方便，这些产品几乎无一例外都采用了汉字识别技术。如果你今后收到任何一张名片，只要用一个小型扫描仪一扫，名片上的姓名、单位、电话等数据就进入了你的PDA中，以后你只要用笔在你的PDA上写出你要查的人名字，就可以马上查到此人的各种数据，这是何等的方便。

目前，我国已经启动了数字图书馆工程，一旦实现了数字图书馆，我们就可以安坐家中，查遍各大图书馆的所有资料。但如何实现将浩如烟海的图书资料输入计算机并提供各种形式的检索功能是数字图书馆所面临的重大问题之一，而汉字识别正是解决这一问题的最有力的手段。

汉字识别技术的发展，大大缩短了普通中国人和计算机之间的距离，对于弘扬古老的中华文化，加速我国的信息产业发展，必将起着越来越重要的推动作用。

（责任编辑：朱婷婷）