微机应用丛书

计算机文字识别技术

胡家忠 著

高等出版社

微机应用丛书

计算机文字识别技术

胡家忠 著

高等教育出版社

内 容 简 介

本书论述了计算机识别英文、数字、汉字的原理和方法。分别讨论了关于文字识别的预处理技术，英文字母及阿拉伯数字的识别，联机手写体汉字的识别，印刷体汉字的识别，手写印刷体汉字的识别，以及表格处理等。书中还对人工智能在文字识别中的应用作了初步的探讨。

这是一本系统地介绍有关计算机文字识别原理和方法的专门书籍。书中内容反映了最新的科研成果，可以作为计算机应用和信息处理等有关专业的研究生及大学高年级学生的参考书，也可供从事文字识别和信息处理领域的研究人员参考。

计算机文字识别技术
胡家忠 著
丛书主编 李桂青

气象出版社
（北京西郊白石桥路46号）
武汉市委印刷厂印刷
气象出版社发行 全国各地新华书店经售

* * *

开本：787×1092 1/16 印张：10 字数：240千字
1994年7月第1版 1994年7月第1次印刷
印数：0001—2000
ISBN 7-5029-1685-7/TP·0044
定价：9.60元

前　言

近二十年来，计算机制造及应用技术取得了日新月异的进展，渗透到各个技术领域。人类正在步入一个技术迅猛发展的新时期。这个新时期的一个重要标志就是计算机在信息处理领域的广泛应用。在日常的生活和工作中，存在着大量的信息处理问题，绝大部分信息是以语言文字作为媒介传播、交换和记载的。因此，随着计算机技术的推广应用，人与计算机的交道愈多，文字语言进入计算机的要求就愈迫切。模式识别技术的发展，极大地改善了人与计算机之间的交互关系，文字识别是模式识别的一个重要分支，是迄今为止被人们研究得较为充分的一个领域。随着文字识别技术的进步，使得记录在纸上(或介质上)的文字得以高速自动地输入计算机，使计算机更有效地进行信息处理成为可能，并为计算机进行知识处理奠定基础。

文字识别是新一代智能计算接口的重要组成部分。它涉及到计算机数字图象处理、模式识别、人工智能、模糊数学、组合数学、信息论，自然语言理解等学科。近二十年来，国内外对数字及英文符号、日本假名及汉字、中国汉字进行了广泛深入的研究，提出了许多行之有效的识别方法，应用软件也越来越丰富，并且有实用装置出现。

著者长期从事计算机文字识别技术的研究。书中介绍的各种识别方法多数经过了著者的实验验证，是著者及著者所在的研究室工作的总结。同时，书中还大量收集了国内外同行的研究成果。

著者认为，应当特别强调文字识别中的预处理工作，因为预处理的好坏直接关系到识别方法的难易程度，直接影响识别结果的正确率。其次，应当根据识别对象(例如是数字、英文、符号还是汉字？是印刷体还是手写体？是联机的还是脱机的？)的不同特点，采取不同的方法进行处理，才能达到识别快而准的目的。

为了叙述方便，本书对不同类别的文字采取分章编写的方法。著者强调面向实际的工程观点，而没有过多地追求理论上的完备性。但并不是说这些方法是孤立存在的，只适用于该类文字。实际上，方法之间是相互渗透的，各种方法有其自身的长处及短处，只有在实际的研究中细细体会，才能融汇贯通，应用起来才会得心应手。

本书可以作为模式识别、人工智能、计算机应用及信息处理等有关专业的研究生，大学高年级学生的参考书，也可供从事文字识别及信息处理方面研究人员参考。

本书在出版工作中，得到了《微机应用丛书》主编李桂青教授的大力支持，黄俊杰教授审阅了全书的内容，杨晓非、陈良华、黄铁军、韵湘、陈国等同志仔细阅读了全书，提出了许多宝贵意见，并帮助做了大量的程序验证及参考图表的编制工作。在此，谨表示深深的谢意。

由于文字识别工作正在日新月异地向前发展，各种识别方法层出不穷，尚在不断的完善过程中，再加上著作学识浅薄，因而无论在内容选择还是在文字的编排上，一定存在着不少的缺点，甚至是错误，敬请广大读者批评指正。

目　录

第一章　绪论

我们在这里所研究的文字识别技术，是指用计算机自动、高速地辨识写在纸（或介质）上的数字、英文符号或汉字，在学科上属于模式识别和人工智能的范畴。它是新一代智能计算机的智能接口的一个重要组成部分。研究文字识别方法，涉及计算机数字图象处理，模式识别、模糊数学、组合数学、信息论、自然语言理解等学科，是介于基础研究与应用研究之间的一门综合性的学科。

1.1　研究计算机识别文字的目的及意义

大家知道，人们之间的思想交流是通过语言和文字进行的。社会发展到今天，已把人类带入信息时代。随着计算机技术的发展及计算机的普遍使用，人们已不再停留在用自己的耳朵和眼睛去直接获得这些信息，并用手将信息记录在纸上，而是使用计算机代替人们的简单、重复的劳动，将语言及文字高速自动地输入计算机，用计算机对它们进行编辑和整理，保存在磁盘、磁带或其它介质上，可随时以各种方式（例如打印机输出、通过电话线进行通讯、通过显示器输出到荧光屏上等）满足人们的不同需要。因此，研究计算机识别文字的目的就是解决文字信息高速、自动输入计算机的问题，使计算机能方便地进行信息加工处理。在以下领域中具有广泛的前途。

（一）在信息处理领域中使用文字识别技术可以大大提高计算机的使用效率，克服人与机器的矛盾。

随着计算机的发展，计算机进行信息处理的速度越来越高。与此相适应的输出装置的速度也大幅度提高，例如激光印字机每秒钟可以输出1000个印刷符号。然而，作为计算机的输入手段却没有多大的改观，仍然停留在依靠人用手指敲击键盘，使计算机在大部分时间里处于闲置状态。计算机的性能越好，人与机器矛盾就越突出。因此，输入的低速度已成为计算机系统提高使用效率的瓶颈，解决这一问题的出路就在于计算机自动识别文字。

在计算机产业刚刚兴起的时候，也许从事计算机录入工作的职业还是一种时髦的职业，人们对它抱有某种神秘感。随着社会的进步，人们文化水平的提高，人们将会越来越觉得这是一项枯燥乏味的工作。再加上这一职业并不稳定，随着年龄的增长，将会失去这一职业，所以解决计算机自动录入的问题刻不容缓。

（二）文字自动识别是智能计算机智能接口的重要组成部分。所谓智能计算机就是用计算机代替人类的部分脑力劳动，视觉是智能计算机接受外界信息的重要手段，它使计算机能阅读文字，看懂图形，理解文章。因而随着资料文献、报表的增加，对文字识别的需求越来越大。

（三）文字自动识别是办公室自动化、新闻出版、机器翻译中最为理想的输入方法。

（四）文字识别后将庞大的黑白点阵图象压缩成机器内部编码，压缩量在100倍以上，对提高通讯容量及速度是大有好处的。

(五)随着笔记本式计算机的发展,联机手写文字的识别将变得更为重要。它可以成为编辑人员直接使用的编辑板,也可以作为记者高速传递消息的重要手段。预计将会有广阔的应用前景。

1.2 研究方法

人能识字,这是众所周知的事情,但是人是通过怎样的途径识字的呢?却还是一个谜。它涉及到心理学和生理学的领域,这些都不是计算机识别的研究对象。计算机识别文字时,是把每个文字作为几何图形来进行研究的。

我们知道,印刷或写在纸上的文字图形,经扫描器输入到计算机时是一幅 $m\times n$ 的黑白点阵图象,共有 $2^{m\times n}$ 种不同的状态,仅就最简单的阿拉伯数字来说,设 $m=n=9$ 就有 2^{81} 种状态,虽然状态数是有限的,却是一个天文数字。当今世界上的存贮器最大容量才 2^{36}。这说明,我们不可能通过计算机来表述文字的每一种黑白点阵组合状态,因而无法验证计算机所识别的字(那怕是最简单的数字 0~9)是完全正确的。也就是说,计算机还不能用明确、有限的步骤达到对每一个字的正确识别。因此,计算机识别文字的目标是:用尽可能少的步骤,在现实可能的存储容量的范围内,用尽可能短的时间达到接近于人的识字水平。

为了达到上述目标,需要对被识别的对象文字的变形施加某些限制(对于手写字符来说,往往不能完全接受这些限制,因而导至识别率下降)。另一方面,我们在研究识别方法时,往往取一定数量的书写样张作为研究对象,即使样张收集了很多很多,但对 $2^{n\times n}$ 种状态而言,却仍然是一个极小极小的数字。对样张识别率的测试,只能称为对某范围内的学习样张的积累识别率。即使识别率很高,也不能说成是对任意的未学习文字的识率别高,而只能通过对学习样张和未学习样张的测试及比较,当未学习样张的识别率接近于学习样张的识别率,并经多方面的测试都比较稳定时,才能大致衡量出识别方法的优劣。应当说,这是一个相当模糊的不准确的概念,但却是我们在实际进行研究时应当遵循的准则。实际上,我们在评价某种方法时,往往是以累积识别率,识别速度以及计算机的开销作为标准来综合进行评价的。也就是说,是从工程应用的观点来研究每一种识别方法的。我们的目标是力求又快又准地解决实际的识别问题,而不追求数学上的严格证明。

1.3 计算机文字识别技术的发展概况

利用机器识别文字符号,可以说从 1929 年陶舍克利用光学模板匹配识别开始。当时,他使用了 10 块模板对应 10 个数字,依次把待识别的数字投影到这 10 块模板上,当模板透过的光达到最小时(数字遮挡了模板的透光部分),数字就被识别成这块模板上的数字。电子计算机于 1946 年问世,从什么时候开始用计算机识别文字,就说不太清楚了。大约在 50 年代末 60 年代初,就已经出现了关于利用计算机识别数字及英文符号的研究论文。随后,日本对汉字识别进行了研究,大约从 70 年代开始,相继对印刷体汉字识别、手写印刷体汉字识别及在线手写汉字识别进行了研究。1980 年进行了印刷体汉字识别的公开表演,1981 年 5 月在日本第 56 届商业展览会上,富士通研究实验室进行了手写印刷体汉字识别的公开表演。1984 年日本研制成多体印刷汉字识别装置,识别率为 99.98%,识别速度大于 100 字/秒,代

表了印刷体汉字识别的最好水平。最近几年出现了手写印刷体汉字识别装置，识别率可达90%，识别速度为5～40字/秒，笔顺可变，笔划数不变的联机手写楷书汉字识别装置已有产品出售，正在研究具有一定规则的手写行书识别装置。

我国的汉字识别研究比日本晚了大约10年，1988年后才有初步实用的印刷体识别系统问世，距今不过五年时间，联机手写汉字识别系统初步可用，手写印刷体汉字识别系统虽然鉴定了几个系统，但还达不到实用化的要求，主要是识别率偏低，估计至少需要提高10个百分点，才能付诸实用。我国研制的汉字识别系统多为软件开发，不需要专用的硬件装置，所以价格便宜，适合我国国情。

从文字识别技术来看，主要集中在特征抽取及匹配两个方面，这一直是OCR技术的两大支柱，基于松弛匹配一类的匹配方法虽然较好地解决了手写字符的变形问题，但匹配速度慢，对字形相近的文字的区分能力差，需要辅之以结构判别才有可能解决识别问题。从目前情况来看，虽然提出了不少的特征提取及匹配方法，在一定的程度上能够识别受到一定限制的印刷的或书写的文字，但决不能说文字识别的问题已经解决，就印刷体文字识别系统来说，目前只能做到对印刷质量好的文件达到较高的识别率；对印刷质量稍差或纸的质量不好的印刷品，由于识别率低，达不到实用化的要求；联机手写体识别装置也达不到比较自由书写、符合人们平时书写习惯的程度，至于手写印刷体识别系统，差距就更大了。

在人们的日常生活中，在机关事务处理、工业、商业交往中，需要识别文字的数量如同天文数字那么大，可以说，到目前为止，利用机器识别的文字量只占需识别的文字量的极小极小的百分比。大约到了80年代初，随着个人计算机的出现，CCD平板扫描器的商品化，文字识别技术才得到蓬勃发展。在美国和日本，为了处理数量极大的邮件，邮局采用邮检用光学文字识别装置。尽管有相当多的邮件由于辨认不清被拒识后由人工加以分拣，但这种设备在经济上还是合算的。美国的税务机关和车辆登记所已经用文字识别装置读取公众用印刷体手写的数字。当前在使用中的一些系统虽然比人读得快，但远没有人读得准，虽然自动识别在许多商业应用中是有价值的，但它和人的识别能力相比差别还很大。

当前的文字自动识别水平远远不能满足自动文字识别最大的潜在的应用所提出的要求，这就是使一般办公室的打字员的工作及出版社的检字排版员的工作(阅读文件并打印出来供人而不是供机器阅读)局部自动化，另一个重要应用是传真电报，这里的对象也是人而不是机器。如果能自动识别出文字，以编码的方式传送文字而不传送文字点阵显然要快得多，还可大大减少对通讯带宽的要求，特别在我国，就更有意义了。

从目前的文字识别技术水平来看，与实际的需求之间的确存在很大距离，可以说，在文字识别领域需要发现一些关键的计算方法，至少现在还没有完全掌握这些方法，另一方面，文字识别必须充分地运用人识字的知识，即字→词→句的理解，从这个角度来说，文字识别技术正期待着人工智能在自然语言理解方面的进步。

1.4 本书的构成

本书力图系统地介绍计算机文字识别的原理和方法，反映最新的科研成果，探索文字识别技术发展的途径。基于这种目的，本书由以下三部分构成：

(一)总论、基础(1～3章)

本章说明了研究计算机的识别文字的目的及意义。在研究方法中强调说明了计算机识别文字的目标及如何达到这些目标。接着通过计算机文字识别技术的发展说明计算机文字识别技术正在一步一步地走向实用化。提出了进一步发展的新课题。

继本章之后，在第二章中说明了计算机文字识别的一般原理和方法。指出提取特征及怎样组合这些特征是文字识别的根本任务。分别介绍了特征匹配方法及结构分析方法，说明将这两种方法有机地结合起来是文字识别技术发展的趋势。

在第三章，介绍了各种预处理方法。它们是计算机文字识别技术普遍采用的方法，是文字识别技术的基础。

(二)处理方法(4～10 章)

这是本书的核心部分。针对各种不同识别对象的特点，介绍了一些具有代表性的识别方法。

首先，根据数字及英文符号的字形特点，在第四章重点介绍了按凸凹分布的特征提取及识别方法。在第五章中，较全面地分析了汉字字形特点及其统计分布规律，为在第 6～8 章中讨论汉字特征提取及识别奠定基础。接着在第六章中介绍了在线手写汉字的特点及其相应的识别策略，重点介绍了以笔划代表点为特征的从正规楷书→笔顺可变→笔顺笔划可变→行书的动态匹配识别方法。提出了研制符合人们书写习惯的在线手写汉字识别系统的具体途径。这可能对笔式计算机在我国的推广应用具有现实意义。在第七章中，讨论了印刷体汉字识别系统的一些具有代表性的方法，这些方法大多比较成熟。目前，我国已在几个较好的识别系统的基础上集成出一个印刷体汉字识别系统，这就是世界上第一个不用人们再校对的印刷体汉字识别系统。在第八章中，系统地介绍了在手写印刷体文字识别中具有代表性的识别方法，这是文字识别领域中最为困难的一个领域。目前只能说取得了部分的成功。相信通过一些具体方法的介绍能对从事这些研究工作的读者有所启发。在第九章中，对人工智能在汉字识别，特别是在手写汉字识别中的运用作了初步探讨，指出汉字识别系统的最后成功正期待着自然语言理解的进步。在第十章中，对表格处理方面的问题作了说明，实际上，它是各种识别技术，特别是文字识别技术的综合运用。

(三)文字识别输入设备及识别系统(11～12 章)

在第十一章中，对以 CCD 器件为核心的图象输入设备的原理作了介绍。在第十二章中，介绍了目前正在使用的一些识别系统，使读者对文字识别的硬、软件系统有一个清楚的了解。

第二章　文字识别的原理和方法

文字识别是模式识别的一个重要分支，是迄今为止在模式识别中研究得比较充分的一个领域。文字识别实际上就是解决文字的分类问题。一般通过特征判别及特征匹配的方法来进行处理。

特征判别是通过文字类(例如英文或汉字)的共同的规则(例如下面将要讲到的区域特征，四周边特征等)进行分类判别。它不需要利用各种文字的具体知识，根据特征抽取的程度(知识的使用程度)分阶段地用结构分析的办法完成字符的识别。因此，整个文字认识的历史，就是特征抽取的发展史。

匹配的方法则是根据各个文字的知识(称为字典)采取按形匹配的方法进行。按实现的技术途径不同又可分为两种：一种是直接利用输入的二维平面图象与字典中记忆的图象进行全域匹配；另一种是只抽出部分图象与字典进行匹配。然后根据各部分形状及其相对位置关系，与保存在字典中的知识进行对照，从而识别出每一个具体的文字。前一种匹配方法适合于象数字、英文符号一类的小字符集；后一种匹配方法适合于象汉字一类的大字符集。

匹配的方法一般用于规范化的印刷文字，特别是同一字体的印刷文字。结构分析方法多用于手写文字的识别。一般说来，匹配方法的程序编制简单，字典占据空间大，识别速度高；结构分析方法程序复杂，能够处理手写体文字的变形问题，具有区分近似文字的优点，但将其用于初始分类则有不稳定的缺点。所以，在手写体文字的识别中，往往将两种方法结合起来使用。

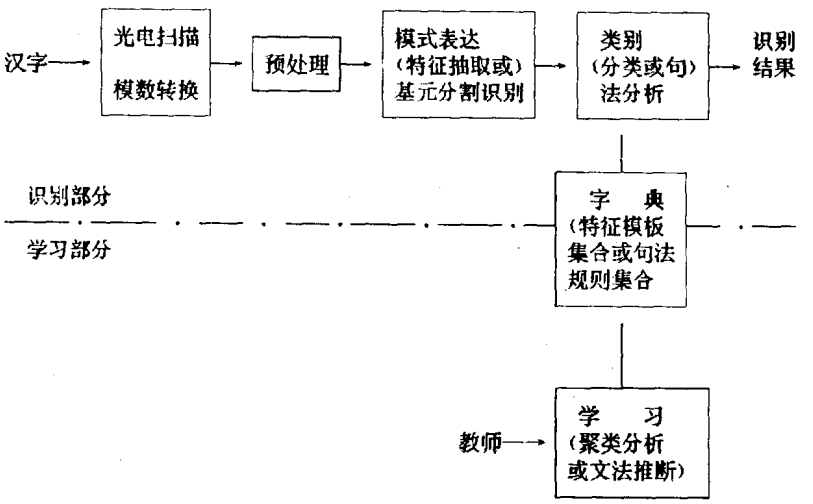


图 2.1　文字识别原理框图

2.1 文字识别的原理

文字识别的原理框图如图 2.1 所示。由扫描器扫描打印或写在纸上的文字，经模数转换成具有一定灰度值的数字采样信号送入计算机，预处理环节一般包括消除噪声，二值化，行字切分、平滑、规范化，进行线性或非线性变换等。经过预处理的文字成为规范化的二值点阵信息(如图其 2.2 所示)，其中"1"表示笔划部分，"0"表示背景部分。

对于二值化点阵，按照识别方法的要求，抽取代表该字的特征，并与存贮在计算机中已知标准文字的特征进行匹配判别，找出字典特征集中与输入文字特征最接近的一个文字，这个字被认为是该字的识别结果。图 2.1 点划线的下部分是识别系统的学习部分。它的功能是自动生成计算机特征字典，学习根据已经准备好的多个字样，抽出代表该字的特征后，自动进行修改，按照字典的规定位置存放该特征。学习有两种：一种是在人的参与下进行，称为"有教师学习"；一种由机器自动进行，称为无教师学习。

图 2.2 文字的点阵图形

2.2 文字识别方法概述

2.2.1 统计决策方法

为了说明用统计决策方法识别文字的原理，下面举两个例子。

1. 定点采样方法

用(i,j)表示文字点阵的坐标，用$f(i,j)$表示该点的灰度值，如果 k 字的笔道通过(i,j)点，则$f(i,j)=1$；如果 k' 字的笔道不通过(i,j)点，则$f(i,j)=0$。这样，$f(i,j)$的值就可以用来区分 k、k' 两个字。我们所需要认识的数字有 10 个，英文符号有 52 个，汉字有成千上万个。一般来说，可以从输入的文字中，测得 N 个特征，这 N 个特征中，每个集合可以考虑成一个向量，称为特征向量。所谓分类问题就是将特征空间中每个可能的向量指定到一个模式类中去。

2. 相关法

上面所说的定点采样方法过于理想化了，只要输入文字稍有变形或移动，即使该点附近存在正确的特征点，结果因为判断仅仅依据一点来进行，常常会产生误判，甚至印刷质量好的文字，也会出现较多的误识。如果我们不是仅仅根据某几个采样点，而是将一个 $n\times n$ 的正规化文字点阵作为字典特征，也就是说，如果把字种 k 的笔划点阵集合 $w_k(i,j)$作为特征向量的一个集合，输入字的点阵为 $f(i,j)$，计算 $\sum_{i,j} f(i,j)\cdot w_k(i,j)$，这个值越大，说明一致性越好。基于这个思想，从如何决定 $w_k(i,j)$，一直到进行怎样的相关计算，开发了各种各样的统计决策的方法。

将上述分类思想予以抽象，从数学上来说，分类问题可以借助于“判别函数”来进行，设用 $w_1、w_2\cdots w_m$ 表示需要加以识别的 m 个模式类，并且令

$$X=\begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{vmatrix} \tag{2-1}$$

表示特征向量，其中 x_i 表示第 i 个特征的度量，用 $D_j(X)$表示与模式类 $w_j(j=1,2,\cdots m)$相联系的判别函数，那么如果特征向量 x 所表示的输入模式在 w_i 中，记为 $x\sim w_i$，则 $D_i(X)$的值必须最大，即对于所有的 $X\sim w_i$，

$$D_i(X)>D_j(X) \qquad i=1,2,\cdots m, i\neq j \tag{2-2}$$

与类 w_i,w_j 相联系的区域之间的边界，称为判决边界。由下述方程表示：

$$D_i(X)-D_j(X)=0 \tag{2-3}$$

$D_i(X)$可以选择满足式(2-2)的各种不同形式的判别函数，在文字识别中经常用到下节将要讲到的几种重要的判别函数。

3. 最小距离分类器

最小距离分类器是线性分类器。它以输入文字与一些参考向量或者特征空间中一些模型点之间的距离作为分类准则。

假定给出 m 个参考向量：$G_1,G_2\cdots G_m$，关于 $G_1,G_2,\cdots G_m$ 的最小距离分类方案是当 $|X-G_i|=\min$ 时，$X\sim w_i$。设 X 表示输入未知文字的特征向量 $X=(x_1,x_2\cdots x_m)$，G 表示字典中某一标准文字的向量 $G=(g_1,g_2,\cdots g_m)$。在模式识别中经常使用下述距离：

(1)明考夫斯基距离

$$D(X,G)=\left[\sum_{i=1}^{m}|x_i-g_i|^q\right]^{1/q} \tag{2-4}$$

当 $q=1$ 时，为绝对值距离：

$$D(X,G)=\sum_{i=1}^{m}|x_i-g_i| \tag{2-5}$$

当 $q=2$ 时，为欧氏距离：

$$D(X,G)=\sqrt{\sum_{i=1}^{m}(x_i-g_i)^2} \tag{2-6}$$

(2)马氏距离

当 X,G 两个 m 维向量呈正态分布，且具有相同的协方差矩阵 $\sum$时，其马氏距离为：

$$D(X,G)=\left[(X-G)\sum{}^{-1}(X-G)^T\right]^2 \tag{2-7}$$

利用最小距离准则进行文字识别时，分别计算输入文字的特征向量 X 和字典文字向量 G_i 之间的距离，$D(X,G_1),D(X,G_2)\cdots D(X,G_m)$，求出其中最小的 $D(X,G_i)$。即可判定输入文字属于 w_i 类。

显然，一个最小距离分类器的性能依赖于适当地选择那些参考向量。

4. 最邻近分类

上述的最小距离分类准则可以这样来理解：既然两个向量 $X、G$ 的距离最小，说明向量 X 和 G 最接近。因此，我们完全可以通过度量两个向量 $X、G$ 的接近程度，这就是最邻近分类

准则。实际上它们都是两个向量之间的相似性的一种度量。在最邻近分类中，经常使用的是类似度 R。两个向量类似度定义为：

$$R(X,G)=\frac{(X,G)}{\|X\|\cdot\|G\|}=\cos\alpha \tag{2-8}$$

式中，分子为向量 X,G 之间的内积，分母中 $\|X\|$，$\|G\|$ 分别表示向量 X、G 的模，α 是向量 X、G 在 m 维空间的夹角(参见图 2.3)。将 m 维向量代入(2-8)式得到

$$R(X,G)=\frac{\sum_{i=1}^{m}x_i\cdot g_i}{\left[\sum_{i=1}^{m}x_i^2\cdot\sum_{i=1}^{m}g_i^2\right]^{1/2}} \tag{2-9}$$

显然，当 X,G 两个向量完全相同时，其夹角为 0，$R(X,G)=1$，它们的距离 $D(X,G)=0$。

除了类似度判别准则外，还有由类似度直接扩展成的复合类似度及方向复合类似度的判断法则，它们都是从类似度概念直接扩展而来的。我们将在印刷体文字识别及手写印刷体汉字识别中，结合具体的识别方法加以说明。

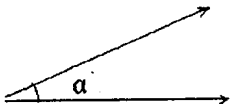


图 2.3 类似度的几何意义

2.2.2 句法方法

1. 什么是句法方法

大家知道，利用句法方法很容易地描述一个字的结构，按照句法方法，每个字是由它的各部分(称为子模式或模式基元)按照一定的顺序组合起来的。利用模式结构与语言之间的相似性，模式识别常以句法的方式进行，即由一组给定的句法规则来剖析模式的结构。(图 2.4 表示出“汉”字与句子的结构的相似性)。

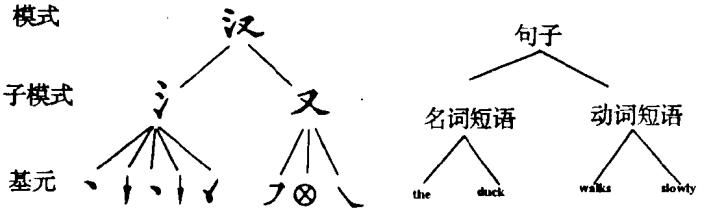


图 2.4 “汉”字结构与句子结构关系的对比

句法模式识别的方框图如图 2.5 所示。短划线上方为识别部分，下方为分析部分，其中识别部分由预处理、基元(包括基元和子模式之间关系)提取和句法(结构)分析组成。而分析部分包括基元选择及文法(或结构)推断两部分。

在句法方法中，一个模式由一个句子表示。该句子属于一个文法所规定的语言，用一组模式基元和它们的组合关系来提供模式结构描述的语言，支配基元组合成模式的规则由所谓模式文法来确定。模式结构信息的另一种表示方法是利用关系图，在关系图中结点表示子模式，分枝表示子模式之间的关系。

文字的基元是笔划。一条直线可由它的起点、终点、线的长度和倾斜度来描述。类似地，一条曲线也可以用它的头、尾及曲率来描述。在选择了基元后，下一步是构造一个(或多

个)文法,以便生成一个(或多个)语言来描述正在研究的模式。

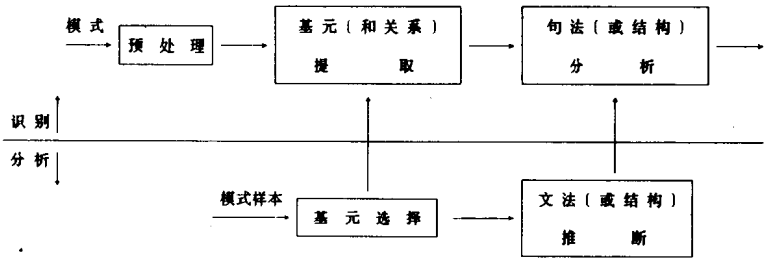
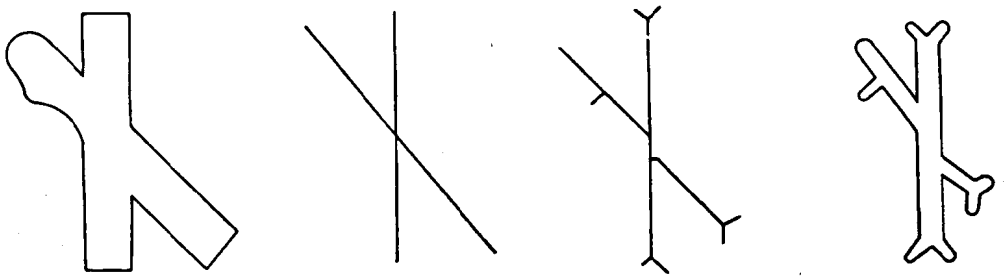


图 2.5 一个句法模式识别系统的方框图

2.误差校正句法分析

用句法方法识别文字的主要困难是笔划不易被计算机正确地提取。

如图 2.6*a* 所示,在用细线方法提取图 2.6*a* 的笔划及其相互关系时,理想的结果当然是图 2.6*b*,而实际上却常常出现图 2.6*c* 的情况,这样就与图 2.6*d* 的情况混淆了。实际上,模式的变形及噪声经常存在,模式的分割误差、基元及子模子的误识别,最终会导致句子被描述的该类文法所拒绝。因此,提出了误差校正的句法分析方法来解决实际的变形及噪声问题。



a.输入笔道; *b*.理想细线化结果; *c*.实际细线化结果; *d*另一种笔道

图 2.6 用细线化方法抽取笔道的例子

所谓误差校正程序就是将描述模式的文法进行扩展,使扩展后的文法不仅包含正确的文法,而且还包含可能的变形及误差。这就是说,对于每一个文字,计算机都保存有一个对这个字的文法描述;而输入待识文字,根据提取的笔划及其组合也有一个文法进行描述。现在的问题是:如果待识文字是字典中的某一个字,怎样从输入字的文法描述中,按规则进行变换,变换的结果就是字典中关于这个字的文法描述?就是说,对于两个符号串 $x,y\in\sum^*$ 定义一个变换 T,使得 $y\in T(x)$。在变换中,经常使用下述三个变换:

(1)替换误差变换 T_s

$$w_1aw_2|\xrightarrow{T_s}w_1bw_2 \quad \text{对所有 } a,b\in\sum,a\neq b$$

(2)删除误差变换 T_D

$$w_1aw_2|\xrightarrow{T_D}w_1w_2 \quad \text{对所有 } a\in\sum$$

(3)插入误差变换 T_I

$$w_1w_2|w_1aw_2 \quad \text{对所有 } a\in\sum$$