文字识别技术

上 册

邮电部邮政科学技术研究所

文字识别技术

段贵昌编译

（下）

邮电部邮政科学技术研究所

前　言

模式识别（*Pattern Recognition*）是关于文字、图形、声音、物体等的自动处理和分类的总称，属于人工智能（*artificial intelligence*）范畴。这一新兴技术有着广阔的发展前景，受到各国普遍重视，十几年来无论在理论上还是在实践上都取得了很大进展。文字是简单的图形，是人们交流信息的重要工具，从技术上看文字图形的识别相对说比较简单，易于处理和分类。因此在模式识别中，文字识别的研究比较深入，取得的成果也大。

文字识别目前主要指的是光学文字识别（*OCR*），它识别的对象包括字母、数字和符号三种。文字识别技术的用途颇为广泛。它可用来阅读邮件地址，实现邮件的自动分拣，可用来识别支票汇票、贸易单式、管理报告，也可用于文件检索和语言翻译等。但文字识别的一个最主要的用途还是作为电子计算机的输入设备，用它代替人或键孔器的工作，自动把文字和其它信息送给计算机。这是当前解决人机关系中最迫切的一个问题。但由于目前的识别机在速度上和识别率上都不能满足这一要求，因此这一问题还没有最后解决。

20世纪初期用机械来读取文字的装置，是作为视力不佳的辅助器而提出来的。40年代末出现的数字电子计算机及随后的发展，促进了自动阅读和计算机语言交换装置的研究。这种研究盛行于1960年前后。但文字识别机的最初实用，是在1958年从磁性文字识别开始的。到了60年代中期，各种类

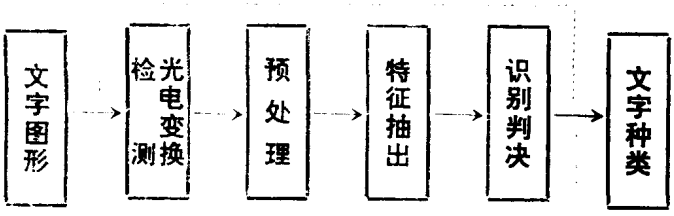
型的数字字母识别机才开始陆续问世，其中以光学文字识别机最为触目。以后的发展相当迅速，1970年日本制成*ASPET/70*型识别机，能够识别印刷质量相当差的文字。1971—1972年，美国和日本运用激光技术扫描文字，进一步发展了扫描技术，取得了新的进展。

目前文字识别技术的情况可简述如下。(1)符号识别技术已经相当成熟。符号形式以条形符居多，也有采用点形符和靶形符的。印刷符号的材料包括油墨、磁性油墨、磷光物质和萤光物质四种。各种符号识别机已在邮政、交通运输、金融和物资部门实际使用。(2)印刷体文字识别（包括字母和数字）比较成熟，已进入实际应用阶段。不仅可以识别*OCR—A*型和*OCR—B*型字，而且可以阅读多种普通印刷文字。目前全世界大约有近万台识别机在各个部门使用。但由于速度较低，特别是由于误差率和拒识率较高，只有少数机器作为计算机的输入设备使用或试用。(3)手写体文字，形状变化大，自动识别困难较大。虽然在理论研究上十分活跃，并取得了很大进展，但就整个技术而言基本上处于研究试验阶段。只有手写体数字的识别接近于成熟，在日本等国实际使用。一般认为手写体文字识别技术的解决尚需十年以上的时间。(4)普遍加强了误差校正技术的研究，使识别机具有差错校正功能，提高了可靠性和识别能力。(5)出现了手持阅读棒。这种识别装置成本比较低，一般不超过一万美元；可以识别各种字体；可以识别各种文件或物件上的信息；可与各种接口连接，具有较强的适应能力；其误差率一般为0.1%左右。阅读棒的出现为识别技术的推广使用创造了极为有利的条件。(6)文字识别机造价昂贵是推广使用的主要障碍之一。60年代一台大型识别机的成本高达一百多万美元。现在降低到十几万美元，有的可降低到几万美元。

(7)阅读速度普遍提高，目前投产的机器一般速度为400—2000字/秒，最高的达到3200字/秒。试验识别机速度可达14、400字/秒，有的在实验室中达到了$5—6\times10^4$字/秒这样的高速度。从投产的机器来看，已经接近计算机输入所必须的2000字/秒这样起码速度。(8)当前识别技术存在的最大问题是误差率和拒识率较高，机器性能不稳定。作为计算机的输入装置来说，合理的拒识率应为$10^{-4}—10^{-5}$，误差率应为$10^{-5}—10^{-6}$，以期能接近或超过人所具有的识别能力。目前的识别机的误差率一般在1%—0.01%范围，距上述要求还有相当大的差距。

文字识别装置的基本结构见下图。检测部分的主要功能是对纸面上的文字进行光电转换，然后把信号送给后面的各部分进行处理和识别。飞点扫描器、光敏器件矩阵、激光扫描器等都是常用的扫描检测设备。予处理部分的功能是将已

文字识别置装



变成电信号的信息加以区分，除去信号中的噪声（如污点、空白等），并根据一定准则除掉一些非本质信号，进行规范化（如对文字的大小、位置和笔道粗细进行规范化），以便减轻识别判决部分的复杂性。特征抽出部分是从已整形和规范化的信号中抽出有用的信息，供判识部分进行判识。例如用特殊逻辑电路抽出文字线条的端点，折点和交点等几何特征

就是特征抽出方法之一。识别判决部分根据抽出的特征，运用特定的识别原理，对文字进行分类，确定其属性，即达到识别目的。

识别机按其识别的字体可分为(1)符号识别机，(2)印刷体文字识别机、(3)手写体文字识别机三种。按书写文字所用的物质可分为：(1)光学文字识别机(*OCR*)，识别白纸上的黑字；(2)磷光或萤光文字识别机，识别用磷光或萤光物质打印的文字或符号(主要是符号)；(3)磁性文字识别机(*MICR*)，用磁性墨书写文字，采用磁感应方式进行识别。按所识别的文件可分为：(1)公文阅读机(*document reader*)，能阅读6×8英寸的文件；(2)帐单阅读机(*journal tape reader*)，阅读特殊打印器打印的文件(文字一般印在纸卷上)；(3)纸页阅读机(*page reeder*)，能阅读8.5×14英寸的纸页，主要是打字机打印的文件；(4)手持阅读棒(*hand held wand*)，能阅读各种尺寸的文件，甚至是书写在各种物件上的文字。

文字识别原理大致可分为两大类。第一类是根据文字的直观特征进行识别的原理，如图形匹配法、笔划分析法、几何特征抽出法等。第二类是根据文字的统计特性而进行识别的原理，如最大似然率、矩法等。近年来模糊集合理论也应用到文字识别中，有很大的发展前途。值得指出的是，对上述这些识别原理的研究虽然在积极进行，但大多数还处在实验室阶段，只有其中的一些较为低级的方法才达到实用水平。本书的主要内容是介绍目前正在研究的和已经实用的各种识别原理。

段贵昌

目　录

（上）

目　　录

（下）

第八章　前后关系技术、语言学技术和矩阵技术

8.1　前后关系

8.1.1　使用全词的前后关系

到目前为止我们研究的识别系统都把输入字符分入识别分类，而不考虑输入字符的前后关系。考虑前后关系一般可改进识别的精确性，现在我们概述几种这样的方法。

使用前后关系的*Bledsoe*和*Browning*方法可应用于许多数值判别式系统，在这样的系统中一个未知模式x分配给第r分类，如果对于全部$s=r$，$g_r(x)>g_s(x)$，其中对于本节的目的，$g_a(x)$，…，$g_r(x)$，$g_s(x)$，…，$g_z(x)$分别为字母a，b，…，z的数值判别式函数。*Bledsoe*和*Browning*前后关系方法要求存储包含机器要识别的全部单词的字典。在未知单词的识别中，该系统实际上只查找与未知单词包含相同数目字母的存储单词。例如一个单词有4个字母x_1，x_2，x_3，x_4。对于字典中的每个4－字母单词，该系统计算与字典单词中的字母相符合的判别式函数值的和。例如对于字典单词‘*able*’，该系统计算$g_a(x_1)+g_b(x_2)+g_l(x_3)+g_e(x_4)$的和。对于字典单词‘*them*’，该系统计算$g_t(x_1)+g_h(x_2)+g_e(x_3)+g_m(x_4)$的和，对字典中全部其它4－字母单词也都是如此。未知单词被识别为具有最大和数

的那个字典单词。

统计研究认为一个单词的先验概率的对数必须加到那个单词的和数中。例如对于字典单词 *'able'* 必须计算

$g_a(x_1)+g_b(x_2)+g_l(x_3)+g_e(x_4)+log($*'able'* 的先验概率$)$的和；对于字典单词 *'them'*，必须计算

$g_t(x_1)+g_h(x_2)+g_e(x_3)+g_m(x_4)+log($*'them'* 的先验概率$)$的和；对于字典中全部其它 4 －字母单词也是如此。未知单词必须被识别为和数最大的 4 －字母单词。

为数学地了解这一点，令 ω 是某一输入单词中的字母数目。根据最大似然判决律，构成输入单词的字母必须分别分配给分类 R_{r1}，…，R_{ri}，…，$R_{r\omega}$，使 $\rho(r_1, \cdots, r_\omega | x_1, \cdots, x_\omega)$ 大于字母的任何其它识别分类的分配。在这里 $\rho(r_1, \cdots, r_\omega | x_1, \cdots, x_\omega)$ 是 ω 个字母分别属于分类 R_{r1}，…，$R_{r\omega}$ 的条件概率，假定该输入单词的字母分别是模式 x_1，…，x_ω。根据贝叶斯（*Bayes*）公式

$$P(r_1, \cdots, r_\omega / X_1, \cdots, X_\omega) = \frac{P(r_1, \cdots, r_\omega) \cdot P(X_1, \cdots X_\omega / r_1, \cdots, r_\omega)}{P(X_1, \cdots, X_\omega)}, \tag{8.1}$$

其中 $\rho(r_1, \cdots, r_\omega)$ 是其字母分别属于分类 R_{r1}，…，$R_{r\omega}$ 的单词的先验概率。在印刷或正楷手写单词的识别中（可能不是在草体单词的识别中），我们假定输入单词中的任何字符的精确形状不依赖于输入单词中任何其它字符所属的分类。

在这种情形中我们有

$$P(X_1, \cdots, X_\omega / r_1, \cdots, r_\omega) = \prod_{i=1}^{\omega} P(X_i / r_i)$$

为极大化目的我们可忽视（8.1）中的 $\rho(x_1, \cdots, x_\omega)$，因为这对于 x_1，…，x_ω 的全部可能的分类分配是相同的。象

第3.3节那样取对数，x_1，…，x_ω必须分配给R_{r1}，…，$R_{r\omega}$，$log\,\rho(r_1, \cdots, r_\omega)+\sum_{i=1}^{\omega} log\,\rho(x_i/r_i)$被极大化。在第3.3、3.4、4.6.1、5.1，和5.2节的识别系统中，数值判别式是（我们希望）一个函数，如果$log\,\rho(x_i/r_i)$被极大化，这个函数也至少大致被极大化。在这样的系统中，x_1，…，x_ω必须分配给R_{r1}，…，$R_{r\omega}$，使$log\,\rho(r_1, \cdots, r_\omega)+\sum_{i=1}^{\omega} g_{ri}(x_i)$被极大化，其中$g_{ri}(x)$是对于分类$R_{ri}$的判别式。

使用前后关系的另一种方法是为*IBM* 1975型机发展的。*IBM* 1975要识别100万个不同的名字，包括名字*'Janes'*。有可能把*'Jones'*中的'0'误识为'*a*'，因此*'Jones'*被误识为*'Janes'*。*Fosdick*和*Hennis*叙述过减少这类差错数目的方法。一共存储两个名字表，第一个名字表包括不大可能与其它名字混淆的名字。如果属于这个名字表的名字是字符识别机的明显输出，这个名字被接受为该系统的最后输出。第二个名字表包括很容易产生误识结果的名字。如果属于第二个名字表的名字是识别机的输出，这个名字被改变成它可能是正确的型式，即*'Janes'*或*'Jonss'*被改变成*'Jones'*。校正的型式作为该识别系统的最后输出。这个校正过程只适用于某些名字，对这些名字要有可靠的统计根据，即平均地把错误的识为正确的概率必须大于把正确的识为错误的概率。对于上述概率比处在规定范围内的任何名字都不列入两个表中，识别系统总是拒识这样的名字。

8.1.2 使用n－字符的前后关系

把*'thick-er-est=en=ens'*作为一个单独的复合款而不作

为 'thick'，'thicker'，'thickest'，'thicken' 和 'thickens' 等五个单独的款存储起来，可能减少在字典中查找单词所花费的时间。这种想法的一个著名的发展是用树形结构而不是用表的形式存储字典。但甚至当这种想法应用时，使用大的字典也是浪费的。

不使用包括全部单词的字典而使用2－字符、3－字符等一般地是更经济一些，尽管与使用全部单词相比误差率的减少要低一些。为我们本节的目的，一个字符是一个字母或一个空白。一个2－字符是一对连接的符号。一个3－字符，4－字符，…，n－字符分别是3，4，…，n个连接的字符的集合。例如 $*CA$，CAT，$AT*$ 是包括在单词 $*CAT*$ 中的3－字符，其中 * 代表空白。

一个简单的方法是存储出现过的每个2－字符的先验概率。让我们研究单词 $*x_1x_2x_3x_4*$ 的识别。假定起始的'*'（代表空白）已经被识别。为了识别x_1，识别系统估计数值判别式 $g_a(x_1)$，…，$g_r(x_1)$，…，$g_z(x_1)$，并把 x_1 分配给 $g_r(x_1)\cdot P(\#r)$是最大的第 r 分类，其中 $P(\#r)$是由'#'和随后的第 r 分类的一个元组成的2－字符出现的先验概率。假定x_1被识别为'C'。然后x_2被分配给 $g_r(x_2)\cdot P(C_r)$ 是最大的第 r 分类。假定x_2被识别为'A'。然后把 x_3 分配给$g_r(x_3)\cdot P(Ar)$ 是最大的第 r 分类，等等，对于单词内的其它字符也是一样。这种类型的技术已由 *Denes* 应用于语言识别，但这有一个缺点，误差趋于随着构成单词的字符列而扩展。

Carlson 研究过这样一个问题：名字中的一个字母是被识别机拒识的一个字符，而名字中的所有其它字母都已被识别机正确地识别。为了对被拒识的字符的识别分类作出较好的猜测，*Carlson* 使用3－字符。为了得到3－字符的先验概率的估计，他使用大的名字表。对于在名字表任何名字中

出现的每个3－字符，*Carlson*的系统计算名字中包括该3－字符的名字的数目。这个数目除以名字表中名字的总数目，作为该3－字符的先验概率的估计。

为了说明3－字符概率的使用，让我们研究名字*MIC•AEL*，其中•代表拒绝。*Carlson*的系统用‘*A*’代替(•)，把*ICA*，*CAA*，*AAE*的先验概率的和作为‘*A*’的得数。该系统然后用‘*B*’代替(•)，把*ICB*，*CBA*，*BAE*的先验概率的和作为‘*B*’的得数。然后用‘*C*’代替(•)，等等，最后(•)用得数最高的字母来代替。为每个3－字符使用两种先验概率，识别性能得到改进。

*Thomas*和*kassler*研究过这样一个问题：在一个英文单词中，每个字符或者被识别或者被拒识，从识别机不能得到其它信息。假定文字识别机的输出是#*PIG•ON*#。*Thomas*和*Kassler*依次用全部字母代替被拒识的字符。在包括英文单词中出现的全部4－字符的表中查找这个4－字符#*PIG*，*PIGA*，*IAGO*，*GAON*，*AON*#。如果这5个4－字符都存在于表中，#*PIGAON*#暂时地作为一个可能的单词被接受，否则被拒绝。对于单词#*PIGBON*#，…，#*PIGZON*#这一过程重复地进行，某些被暂时地接受，另一些被拒绝。

暂时被接受的单词再经受第二次试验，该试验设计用来拒绝在包括全部单词的字典中找不到的单词。一般地说，第二次试验使用n－字符的两个主表，其中n是个变数。在第二表中n－字符的一个集合与第一个表中的每个n－字符相对应。在第二个表中的每个n－字符包括第一表中与它相对应的n－字符。当第一表中的任何n－字符在被接受的单词中发现时，这个单词被拒绝，除非它还包括第二表中的n－字符之一，这个n－字符与第一表中的n－字符相对应。

使用两个表为下述问题提供了解决办法，即在待选单词

中哪一个 n －字符必须被检验是否在表中存在。而如果使用固定值的 n 并规定全部 n －字符都必须检验，上述问题就可避免。但 n 的任何固定值对于使用性有时不够大，有时又太大。例如4 －字符或5 －字符试验将接受假单词 *'SPATENT'*，因为 *'SPATE'* 和 *'PATENT'* 是合法的英文单词，而2 －字符试验将拒绝假单词 *'PQN'*。

文字识别机仅仅是产生识别错误的根源之一。另一个识别错误的根源是写字的人，如打字员。识别机可能误识或拒识某一字符，但它不会不作任何反映就放过任何字符。因此造成错拼的单词，使某一单词中包含的字母数目发生错误。

我们仅仅考虑了识别单词中和名字中的字符所使用的前后关系。实际上文字识别机还经常用于识别任意排列的字符，如会计帐目。另外在要识别的字符中还可能加入校验字符，因此只有当字符列中的全部字符都正确识别时，某种规定的算术试验才能得到满足。

8.2 场景分析（Scene Analysis）

在现代的结构语言学中，一列符号不仅被识别，而且被解析地描述。有人建议语言学方法对于模式或场景的形式描述或分析也许是可以使用的。一个场景是包含许多物体的视野。场景分析是关于描述2维或3维场景的技术术语。全面的介绍场景分析方法不是本书的目的，下面仅就场景分析与模式识别不同的某些方面加以说明。

8.2.1 某一物体可能分类的多重性

把一个物体识别为一个立方体就是把“立方体”这个名字分配给这个物体。在场景分析问题中可能没有必要把这样