第5卷 第1期 自 动 化 学 报 Vol. 5, No. 1
1979 年 3 月 ACTA AUTOMATICA SINICA Mar., 1979

光学文字识别系统*

周 新 钱乐秋 执笔
(复旦大学文字识别研究组)

摘 要

复旦 789 型光学文字识别系统是为专门需要而研制的。其目的是为了直接阅读原稿上的打印的英文字母与特殊符号。它的指标是在阅读速度为每秒 100 字符的条件下，要求误识率与拒识率均达到 10^{-4}（其中还包含了红黑颜色的误分率）。由于采用通用型小写英文字母，加上待识文件制作过程为连续自动打印，以致在字符的可分性与印刷质量的扰动上面临复杂的形势。本文涉及的范围是有关识别系统的研制过程与方案的选择。在此基础上提出了一个具有硬软结合的贰级识别的方案。并给出了试验的结果与改进的方向。

序 言

789 型光学文字识别的对象是由数量众多的西门子 100 型电传打字机所制作的连续型折迭式文件。从装上新色带开始打印，至少在廿万字符的范围内，电传机不停地自动地工作。为了将原有的卅二种字符的能力扩大为六十四种，采取了红黑色带上下交错的工作方式，从而产生了红黑颜色的字符。字型为通用小写廿六个英文字母与六个特殊符号：

a b c d e f g h i j k l m n o p q r s t u v w x y z ,!/-?@

由通用小写字型与自动连续打印，从识别的角度上分析所带来的问题有：

1. 字符类之间的差异不明显，相似性大。如按照字形的相似程度，明显的可将若干字符类合并成下列几族。

"a" 族：其中成员为 a、e、s；

"o" 族：其中成员为 o、b、d、p、q、c、g；

"n" 族：其中成员为 h、n、r；

"i" 族：其中成员为 i、j、f、l、t；

"v" 族：其中成员为 v、y；

在同一族的成员间，重合部位所占的比例相当大。如 a 族经过重迭的字符图形为图 1。则 a = 重合部位 +2+3+4；e = 重合部位 +1+2+3；s = 重合部位 +1+4。

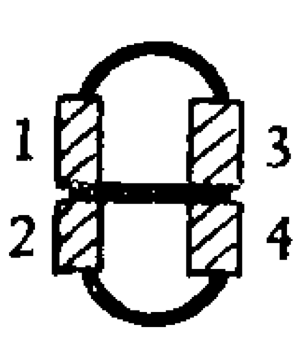


图 1

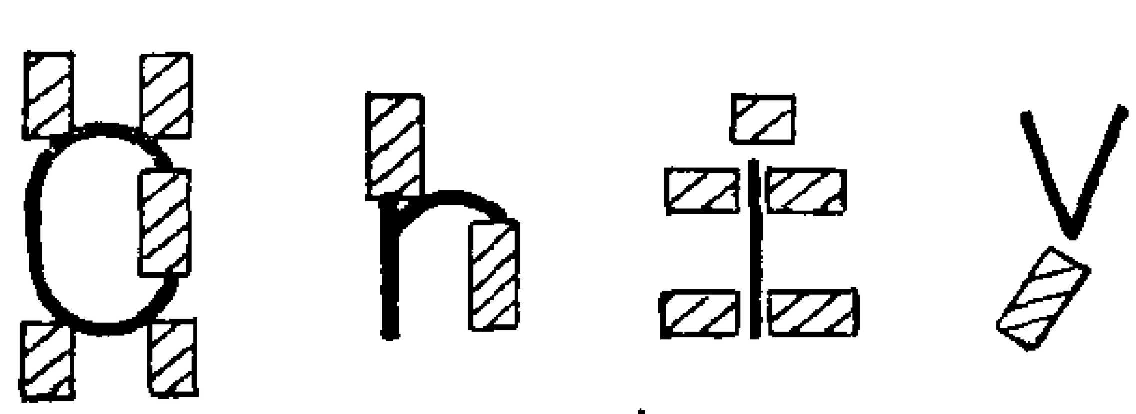


图 2

* 本文曾在中国自动化学会 1978 年年会上宣读

同样情况出现在其它各族上如图 2. 由上述可知，同一族成员间的区别仅存在于局部地区.

2. 印刷质量在一个较宽的范围内变化，如字符浓淡的程度可从炭黑到几乎看不出来，线宽的粗细可达 6:1，字符在一行中位置的参差高低可相差半个文字. 除此以外还经常伴随着一些异常的情况，诸如：缺损、沾污、重影、字相连以及出现在一个字符范围内的半边深半边淡或上部红(黑)下部黑(红).

3，连续型折迭式文件的折痕所形成的阴影，不仅会有误当作字符信息进入识别部的危险（注：当折痕位于一行字符当中时后果更严重)，而且文件纸面的不平整还会影响聚焦的效果.

由对象的复杂性与指标的严要求，将整个系统的研制过程分成予研与试验样机两个阶段是现实的也是必要的. 预研的目的是借助于通用计算机，在小样本的条件下，开展对识别系统的模拟与仿真，以探索合理的方案与手段. 试验样机的任务是硬设备的制作与试验的扩大，然后再根据试验的结果，对识别系统反复修改，使趋于完善.

第一部分　系 统 设 计

方案的选择

在预研阶段，最早考虑的识别方法是对印刷体文字识别具有代表性的相关匹配法.这种方法从信息论的观点，是经典的信号检测问题中匹配滤波器在字符识别中的拓广. 在这样一个信号检测系统中，输入信号就是待识字符，而参考信号就是标准的卅二个字符.一个相关器联接输入信号与参考信号，从而提供了一个由噪声中提取信号的输出. 在这里的噪声包括了出现在输入信号中的沾污、缺损、伸缩、旋转等自然噪声与参考信号中除去待识字符以外的字符. 相关器的输出再与一个预设的门限值相比较，大于门限值的峰值在参考字符序列中所对应的那个字符就判为待识字符. 早期的试验，将标准化字符以每毫米 8 点采样，组成 20 × 36 的量子化图形，然后在二维空间域中计算它们在规一化条件下的相关系数或称类似度.

$$R(x, y)=\frac{\sum_{i=1}^{20}\sum_{j=1}^{36} f(x_i,y_j)\cdot g(x_i,y_j)}{\sqrt{\sum_{i=1}^{20}\sum_{j=1}^{36} f^2(x_i,y_j)}\ \sqrt{\sum_{i=1}^{20}\sum_{j=1}^{36} g^2(x_i,y_j)}}$$

计算结果表明：不同族的字符类之间的类似度较小，一般在 0.5 以下，而同族的字符类之间的类似度较大，一般在 0.6 以上，其中特别突出的如 "b" 族、"h" 族中某些成员的类似度可达 0.75 左右.

以上计算是在字符取标准化的理想状态下测得的，而实际的情况要复杂得多. 试仅以字符的位移、伸缩和旋转为例，表明 R 对它们是十分敏感的. 也就是说，即使同一个字符本身，由于匹配位置的不一致、字形的胖瘦或字符倾斜，就会使 R 显著减小. 其变化趋势如图 3.

其中 α 表示旋转角的弧度，如设 m 为放大量，则有关系：$\alpha = m - 1$.

由图可见字符图形只要放大 1.06 倍，R 就从 1 下降到 0.57.

显然在各种噪声的干扰下，对同一字符类 $R\searrow$，对不同字符类 $R\nearrow$. 这样特别对某些相似性本来就较大的族间成员来说，很难找到区分它们的界面这点是不言而喻的.

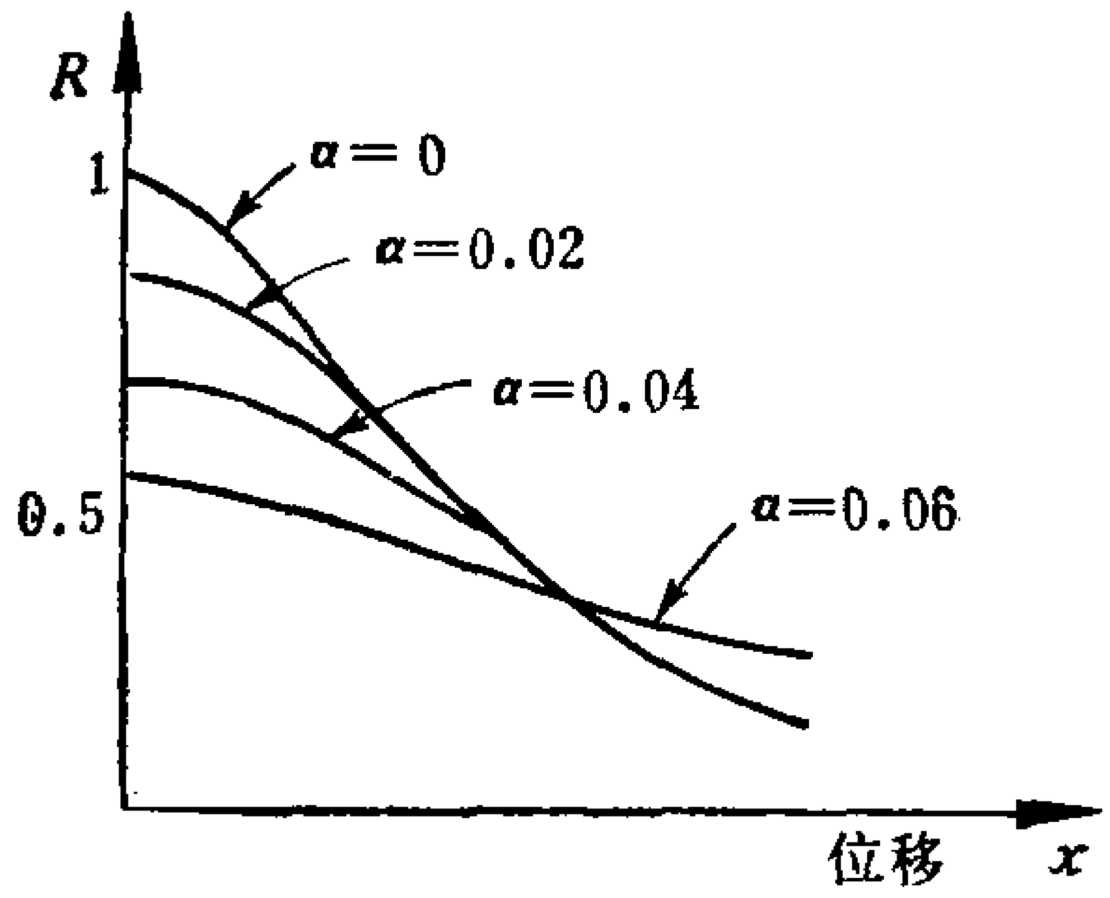


图 3

有一种补救措施，就是较严格的预处理，诸如位置与线宽的规一化，伸缩与旋转的校正所谓“弹性匹配”.但这些都要以大量的计算程序或众多的设备为代价. 日本东芝公司研制的光学文字读取装置 *ASPET/71*[1] 采取了“复合类似度”，缓和了位移对 R 的影响，但它是构筑在多于 500 只集成化的高质量运算放大器的基础之上. 基于上述情况，一种线性的相关匹配的识别方案的选择从预研中被排除.

特征抽取常用于识别手写文字，它不象相关匹配法注重于文字的全貌，而是着眼于文字的特征.所谓特征是指文字在经受各种噪声与变化的过程中，那些不变的或者说不敏感的成分. 因此只要把存在于文字类中的特征提取出来，组成一个多维的特征空间，则每个文字就对应此空间中的一点，余下的问题是如何划分特征空间，使每类文字的集群都落在一个给定的范围之内. 美国 *IBM* 公司在 1975 型光学文字识别系统[2]中，为解决低印刷质量与多字型(多达 200 种字型以上)的印刷字符的识别，设计了九十六只特征“度量”，通过最小距离判决准则，使误识率降到 10^{-4}(但有近三分之二的卡片作为拒识被剔除).

我们开始按照字符的线段、圆弧、端点、尖角等特征抽取了44只“度量”，在编排各字符类的距离表时，出现了与相关匹配法类似的问题，不同族的字符类之间特征的差距可达10以上，而某些相同族的字符类之间最小只有 3. 如维持这种极不均匀的分布，将会给等距离的决策准则带来困难. 关键是要对特征加“权”，从而在变换后的空间里，对各字符类划分的距离趋于均匀. 但特征加权决策占用通用计算机的时间，将威胁到每秒 100 字符的识别速度. 另外对于大多数不同族之间的字符类在特征空间的距离本来就比较宽而无需变换，真正需要加权决策的只是少数同族的字符类，故较为适宜的策略应是区别对待. 为此，针对各个字符类的特点，特别是侧重于同族字符类的区分，专门设计了卅二只“特殊度量” 使与卅二个字符类一一对应. 其作用在于它一方面参加到原有的四十四只 “一般度量”的决策行列，使字符类之间的距离平均增加了二. 另一方面更为重要的是当等距离决策遇到有若干个对象都合格而无法判决谁应该当选时，“特殊度量” 就起着举足轻重的影响. 这样一个复式的决策方案既节约了时间，同时从概率的角度分析，比单级的命中率肯定要高. 试验数据的统计也证实了这一点. 如单独用七十六只 “度量” 作等距离决策，其误识率为 10^{-3}. 单独用卅二只 “特殊度量” 作零距离决策，其误识率也为 10^{-3}. 而使用了复式决策，其误识率就接近降低一个数量级. 在多次每批四千多字符的识字试验中，误识统计平均为一个字符. 有关这方面的总结详见[3].

进入试验样机阶段，面临的新课题是如何在识别率、速度与设备三者之中求得协调，以全面实现指标的要求. 突出的矛盾集中于七十六只“度量”的固化将涉及到几千只固体组件. 针对当时组件的国内水平，既是不经济的，也是不可靠的.一个折衷的考虑是将“特殊度量”零距离决策做成硬件作为识别系统的前级，而“一般度量”等距离决策保留软件形

式作为后级．这种布局上的改变将对“特殊度量”的品质要求有所提高，要它独立承担 10^{-4} 的误识率，然而在拒识率上有灵活的余地．根据识别系统的品质因数 C 与误识率 S 以及拒识率 R 的内在联系，一般表示为：$C=\sqrt{S\times R}$．因而放宽拒识率的要求是换取低误识率的一种手段．但拒识率过高将会增加软件识别的负担，特别是速度上的失调．以致一个合适的拒识指标是比误识率放宽一个数量级，然后让软件在不显著增加误识率的前提下，只要识出其中的十分之九，这样一个两级识别系统就合理地分担了两个万分之一的指标．问题回到如何提高“特殊度量”的品质．总结预研阶段中复式决策的经验，其指导思想应该是：

一、反映同族字符重合部位（序言中提及的深色部位）的特征组合应具有更大的适应性，使尽量通过本族字符，而不让其它字符通过．由于矛盾主要存在于同族字符类，所以这点是易于做到的．

二、反映同族字符差异部位（序言中提及的斜线部位）的特征组合应具有更精细的刻画，使将同族中的字符类区分开来．

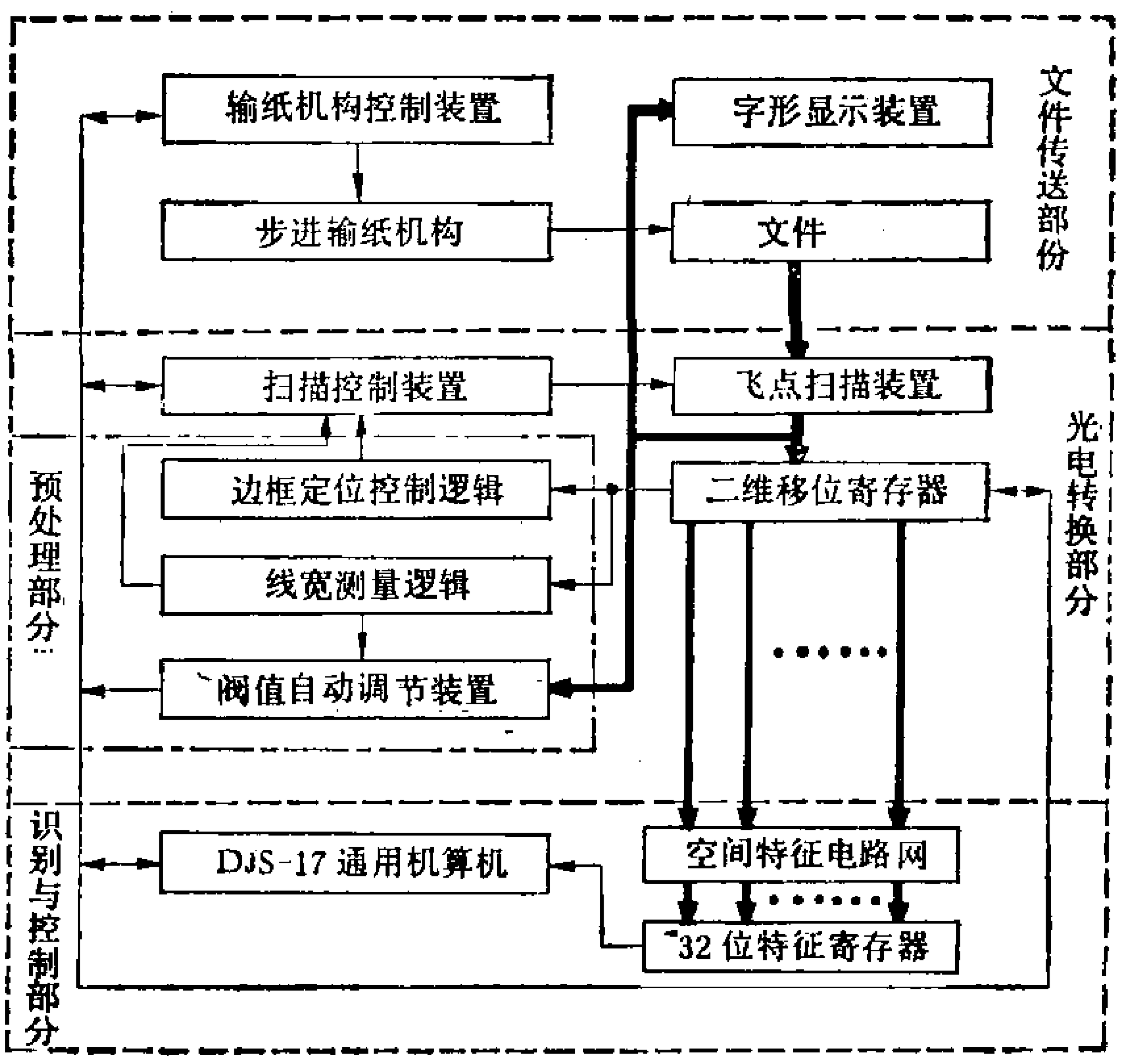
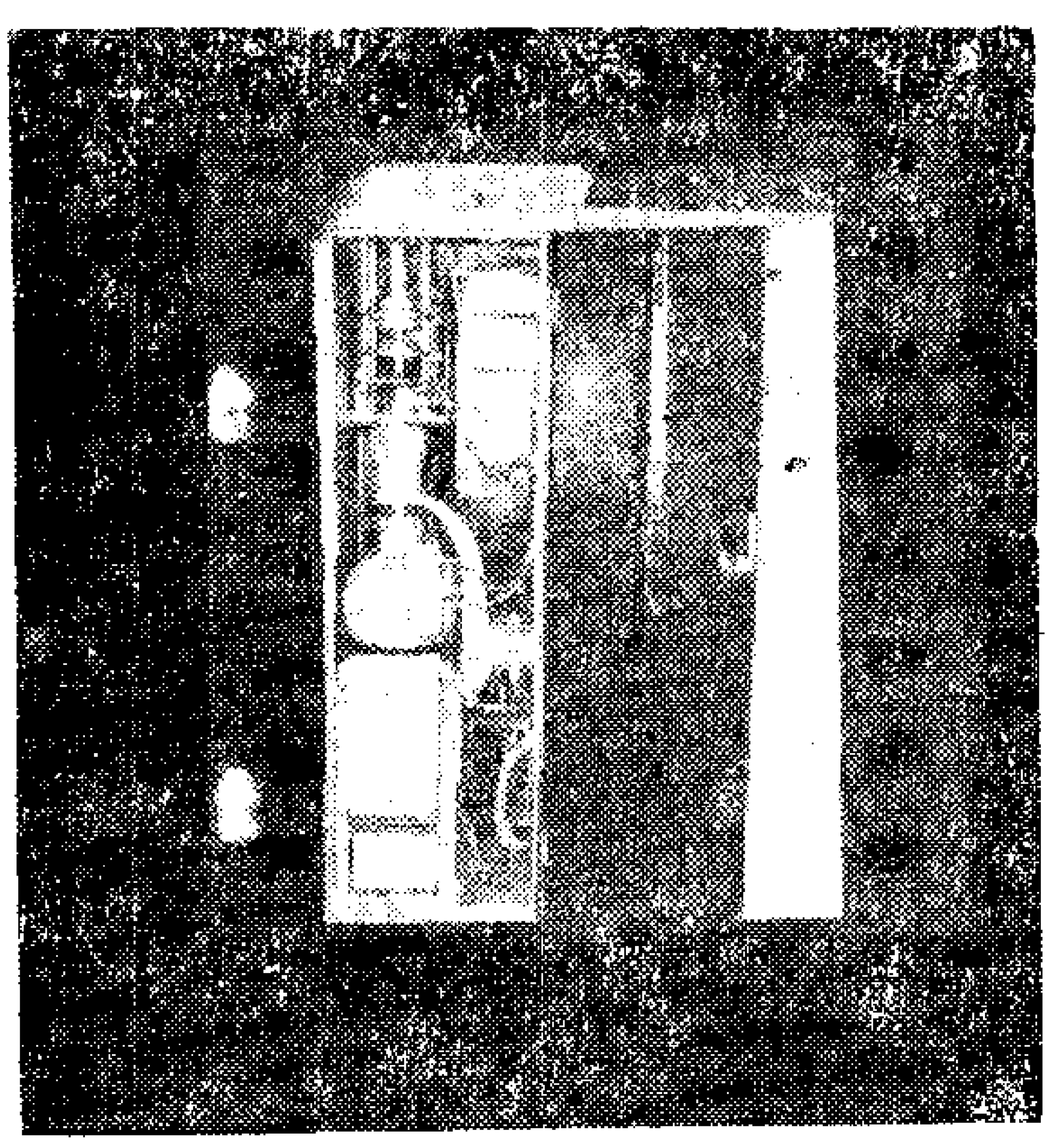
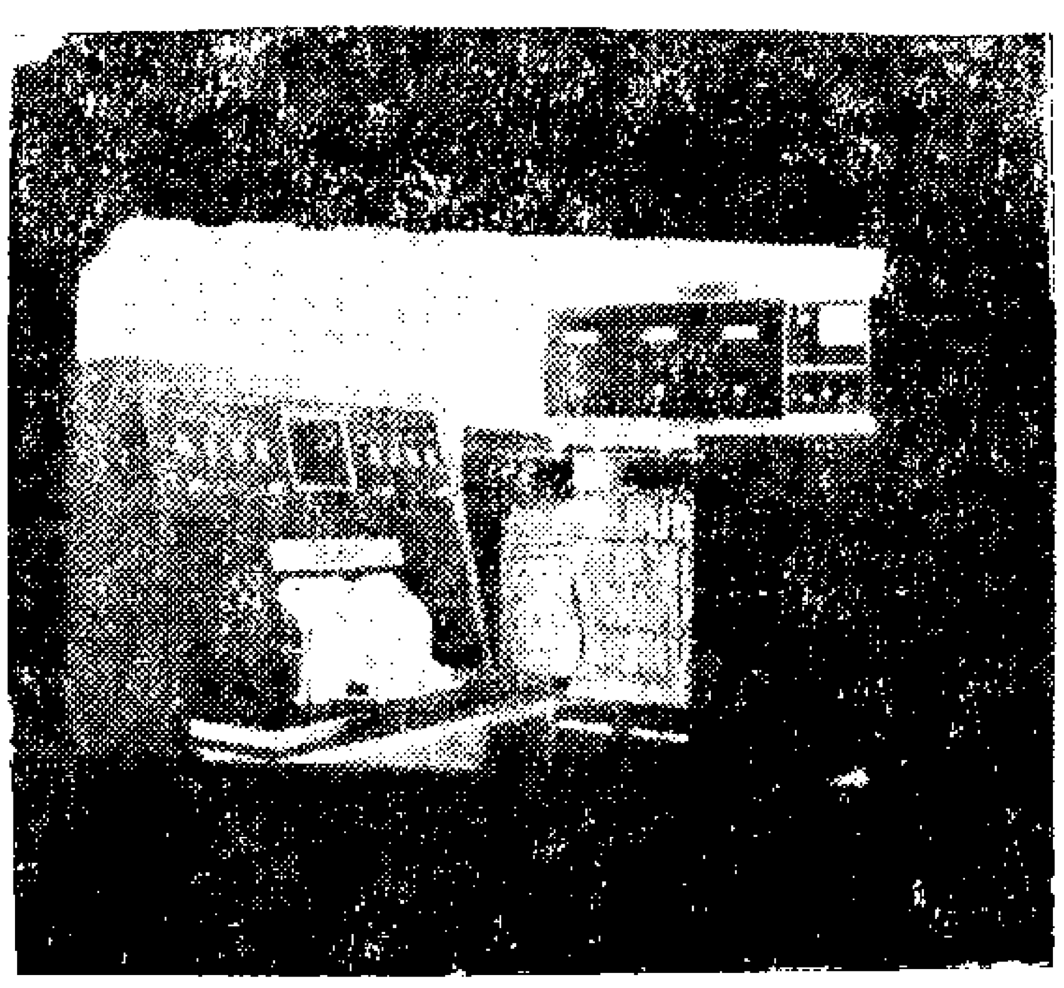


图　4

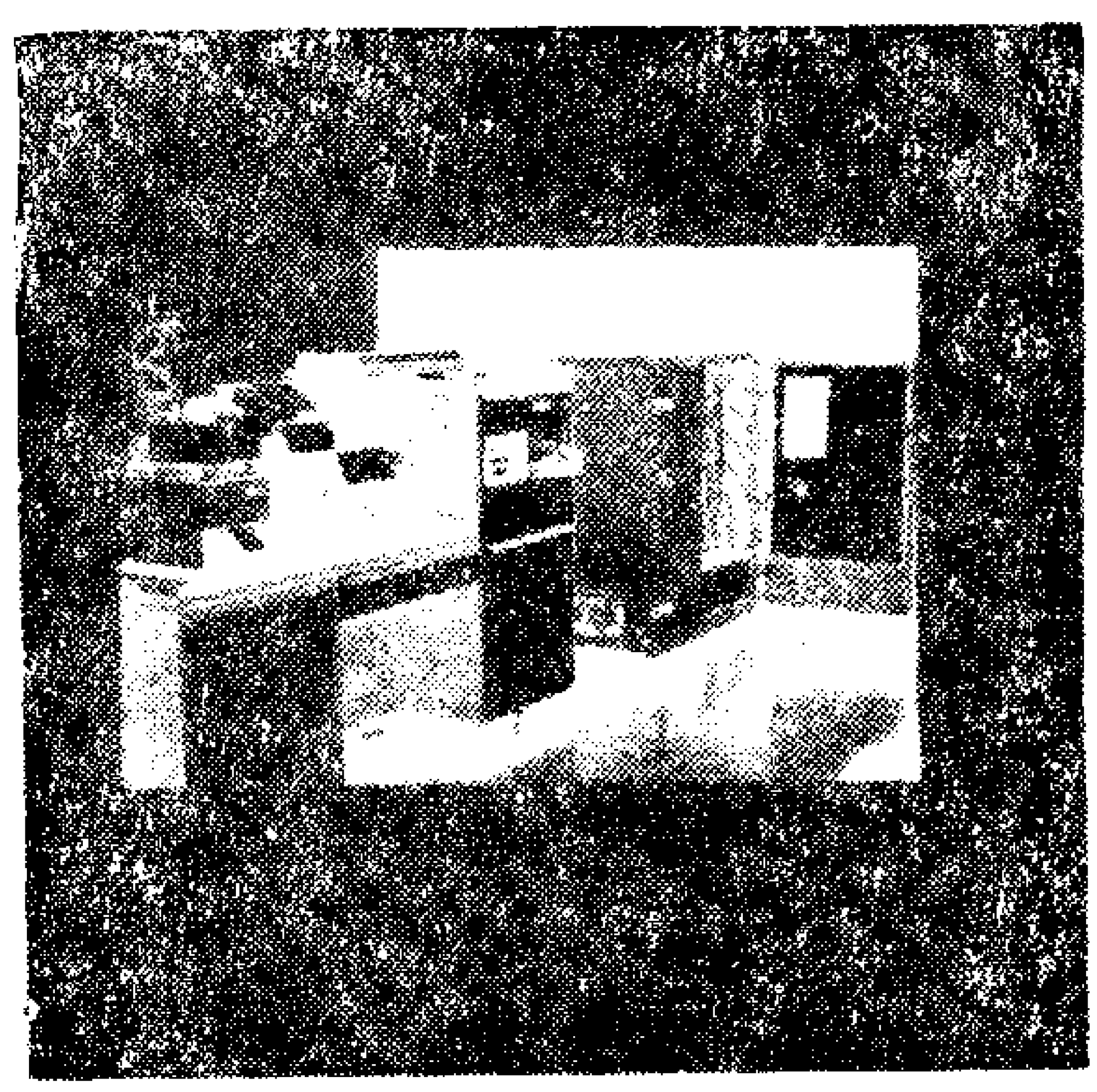
这样的设计思想，在识别同族字符的意义上，相当于前者代替了“一般度量”粗分类的功能，而后者完成了“特殊度量”加权的细分类作用．至于距离决策思想可以通过构成“特殊度量”逻辑表达式的“或”关系来体现．经过反复的训练与修改，新的度量逐渐适应了环境，达到了预定的要求，余下的问题是软件的识别率．基于对硬件拒识对象的分析与识别速度的考虑，一个以笔划分析加上字符位相关系的综合识字方法取代了“一般度量”．它的速度提高到每秒五个字符以上，以致可以利用对字符扫描的同时进行软件识字，而无需单独占用计算机的时间．这样一个硬软结合的贰级识别系统，在利用一台小型通用计算机作为处理机的



A．飞点管暗箱



B．飞点扫描器



C．移位寄存器　　D．分析计算机

条件下,基本上解决了识别率、速度与设备上的矛盾,从而达到了预期的效果.

最终完成的试验样机的框图见图 4.

有关试验样机的概貌见上页的照片.

试验结果

经过多次十万以致一百五十万字符的识字试验，证明最终完成的这套二级识别方案的误识率为 10^{-4}，拒识率亦为 10^{-4}. 其中硬件拒识率为 10^{-3}，误识率小于 10^{-4}；软件拒识率为 10^{-1}，误识率小于 10^{-2}.

在一九七八年五月的鉴定会上进行了 96596 字的识字试验，被这套方案拒识的字符有 8 个,误识的字符也是 8 个.

结 束 语

本识别系统由于采用了硬软结合的二级识别方案,从而提高了整个系统的效率,增加了系统的灵活性,保证了两个万分之一指标的实现. 如字型选用 *OCR-A* 或 *B*，识别率还可望提高. 在识别速度上，如果采用超短余辉的飞点管，扫描频率提高到 1*MC*，同时将通用计算机改为专用控制机，每秒一千字左右的识别速度是可以达到的. 空间特征逻辑网络在识别过程中几乎不占用时间，但它不便于修改. 而度量设计又需要经过一个冗长的反复的人工参预过程. 发展的方向是希望由具有自学习功能的自动设计度量的方法来代替，但目前可以采取的现实途径是充分利用人——机对话系统，将计算机处理的快速、判断的精确与人的聪明才智巧妙的结合起来，研制出适应性更广、识别率更高的识别系统.

第二部分：识 别 方 案

一、识别方案的工作过程

经过飞点扫描系统的光电转换,把文件上的文字转换成如图 2-1 所示的二进制信息. 它们保存在 $20 \times 36 = 720$ 位的移位寄存器中. 有信息的点为 1，没有信息的点为 0. 本方案识别的对象就是这样的二进制信息.

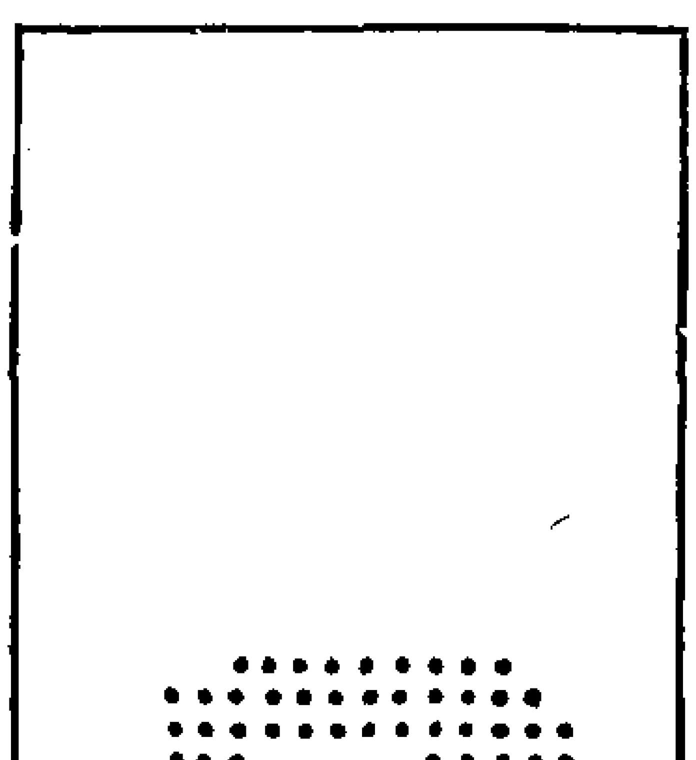


图 2-1

文件规定,按 1mm 8 点来计算，每个文字的宽度 $\leqslant 20$ 点，高度 $\leqslant 36$ 点. 一行中平均每个字约占 20 点的位置.

考虑到识别速度及电子器件的节省等因素，整个识别方案是由二级识别所组成. 第一级识别方案全部由硬件实现，因此速度快，但器件费. 第二级识别方案是由软件实现，因此速度慢，但不花费器件. 只有在第一级方案中拒识的文字才到第二级方案中去进行识别. 实践证明,目前采用的方案中,第一级识别方案可识出 99.9%的文字,即在一千个字中只有一个字拒识. 而第二级识别方案中识一个字大约需要 200ms 的时间，一千个字的扫描时间约 3.5 秒，所以利用计算机多道程序的特点,第二级识别方案几乎不影响整个识别速度.

在第一级识别方案中，对应于26个英文字母和6个特殊符号，我们设计了32个度量．每个度量实际上是用硬件组成的移位寄存器上的某些点的一个逻辑表达式．每个表达式反映了一个文字在移位寄存器中的特征．一个文字扫描到移位寄存器后，即在移位寄存器的各个不同的位置上进行测试，如果它满足某个表达式的逻辑关系，则它就是该度量所对应的字母．

然而，由于种种原因，扫描进来的文字除了图2-1这样的正常文字外，还有满格（见图2-2）、二字连（见图2-3）、多字连——这往往是由于纸上的折痕引起的（见图2-4）等情况．另外在某些文字缺损时，使一个文字变成二个字，称为字拆散（见图2-5），在这种情况下，必须由程序来进行字分割或字合并，然后重新扫描和识别．

为了减少识错和拒识，文字在初次拒识后，先由程序调整扫描器的阈值后重扫．如果若干次重扫后都拒识，则这个字才真正作为第一级拒识文字进入第二级识别方案．

在遇到字拆散的时候，前半个字扫入后，往往是被第一级方案拒识的，显然此时就把它送到第二级方案去识别是没有必要的．因为当后半字扫入后发现字拆散时，前半字的识别结果是无效的．考虑到这个因素，因此在扫完下一个字后，确认前字不是字拆散，并且前字是拒识的情况时，才启动软件识字程序（即第二级方案）．

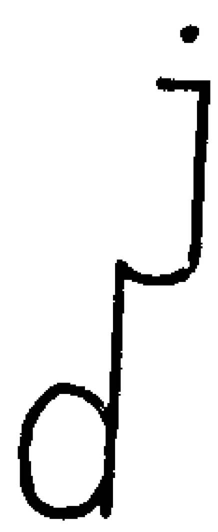


图　2-2

图　2-3

图　2-4

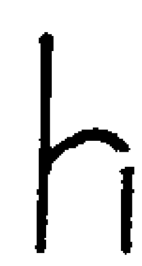


图　2-5

框图说明：

文字识别方案的工作过程是由通用计算机通过主控程序来控制执行的．主控程序借助于管理程序使得扫描器、步进与穿孔输出、主控程序与软件识字程序之间并行工作．这样提高了系统的工作效率，并使各部件达到较好的匹配．工作流程图见图2-6．

主控程序在使各部件处于初始状态后启动扫描器工作，在扫描器扫描过程中，由于扫描还在进行，因此主控程序暂时退出运行，当扫描器扫完一个字并发出结束中断后，主控程序重新投入运行．此时主控程序根据扫描结束状态进行分类处理．主控程序对正常结束则就直接进行决策处理，对其它情况分别采取相应的处理，依据处理结果决定下一扫描位置，使扫描器继续工作直到一幅（四行文字）结束，然后启动步进换一幅，在扫下一幅的同时输出上一幅识别结果．主控程序在决策处理中出现拒识时，就把有关信息提交给软件识字程序，软件识字程序利用扫描器扫描过程中，主控程序等待扫描中断的空隙时间对拒识文字进行识别处理．

二、第一级识别方案（硬件度量）的设计思想

1. 为了工程化的简便起见，我们在720位移位寄存器上先构筑一些基本小匣子．

基本小匣子按其性质来分有四种：① 表示几点中至少有一点为1，记为 0_1．② 表示几点全部是0，记为 A_0．③ 表示几点中至少有一点为0，记为 0_0．④ 表示几点全部

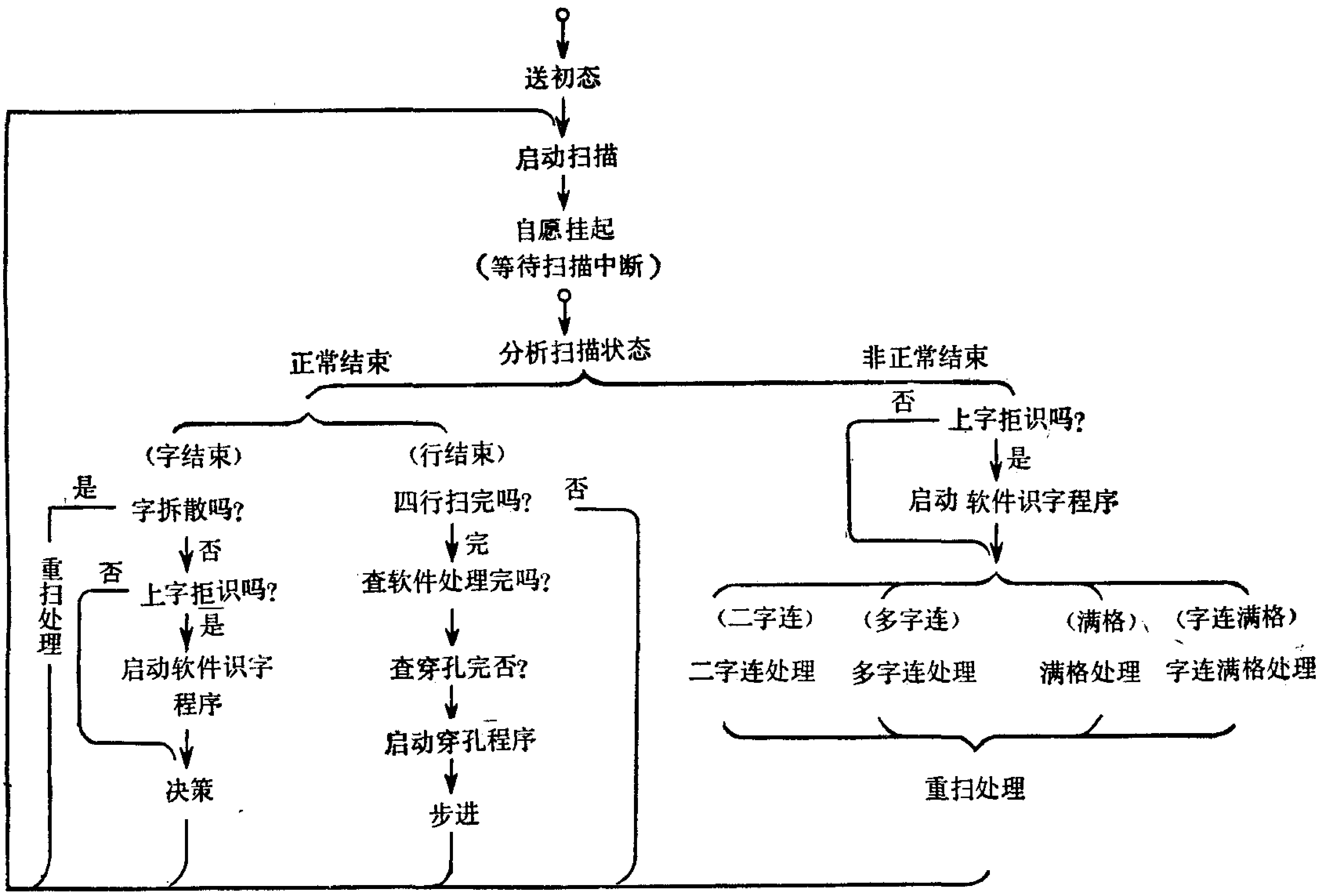


图 2-6

是 1，记为 A_1.

基本小匣子按其位置来分有四种：① 横向相邻二位组成的匣子记为 2W. ② 横向相邻三位组成的匣子记为 3W. ③ 竖向相邻二位组成的匣子记为 2S. ④ 竖向相邻三位组成的匣子记为 3S.

考虑到度量大致都设计在移位寄存器的中间，所以 720 位移位寄存器的中间部分都设计成 2S、2W. 边缘部分都设计成 3S、3W.

2. 根据文字的基本形状抽取特征，确定那些位置上必须要有 O_0、A_1、O_1、A_0 等信息. 再根据这些位置用 2S、2W、3S、3W 代入组成一个逻辑表达式. 这样的逻辑表达式即表示一个度量.

3. 度量设计中的几个问题：

① 为了区分形状相近的文字，例 c 和 O，则在关键部位——c 的缺口处，要比较细致地刻划它的逻辑关系.

② 对于不易与其它文字相混淆的地方，逻辑关系可以极其粗糙，以节省器件.

③ 逻辑关系中要考虑文字的缺损. 例 b d p q 等字经常由于打字机锄头的位置偏移造成O部分有缺损，例 bd pq 或 bd pq 等，因此在这些文字的度量中的O部分，应至少允许有一个方向的缺损.（当然这个缺口不能太大，太大了要与其它字混淆. 例 b 字下面缺损太大与 h 就分不清了.）

对噪声也同样处理.

4. 逻辑上作化简及合并，以节省器件.

5. 度量初步设计后，经过大量的反复试验和反复修改，最后定型.

三、第二级识别方案（软件度量）的设计思想

软件度量的识别对象是第一级识别方案中拒识的文字.

这套度量的设计是采用分析文字的笔划及其相对位置的方法. 为简便起见，我们把文字看作由竖线、横线、斜线所组成. 文字的圆弧部分用竖线和横线来近似. 例 b 字就用 ㅂ 来近似.

对每个软件度量规定一条主线. 例 ㅂ 的主线是左面的竖线. 描述主线的信息是线型和线的长度范围.

度量中的其它线条——不妨称它为付线，都是描述它与主线的相对位置.

我们知道，如果确定了二点的位置，一条直线的位置也就确定了. 在付线的度量信息中，我们给出付线的起点和终点与主线的起点或终点的相对距离（± n 行、± m 列）.

对于每个被第二级识别方案识别的文字：

1. 先求出它所有的竖线、横线、斜线的绝对位置.

2. 然后通度量. 在通某个度量时，先在文字的线条中寻找是否具有该度量的主线，如果找到的话，则根据被识文字中主线的绝对位置及度量中付线对主线的相对位置，则可计算出付线在该字中应有的绝对位置. 然后在文字的线条中找是否存在这些付线，如果有的话，则该度量被通过，否则该度量不被通过.

由于文字的笔划是有一定的宽度的，即满足度量所指出的主付线相对位置的线条组，在一个文字中可能存在若干组. 所以一个文字往往可以多次通过一个度量. 例 ㅂ|、ㅂ就可能通过二次、四次 b 的度量.

3. 对所有度量进行测试.

4. 根据这个文字对各度量的通过次数进行决策，我们选通过次数最多的度量作为该文字的识别结果.

几点说明：

1. 为了提高度量的适应性，在寻找付线时，允许位置有一定的偏移.

2. 为了区分相似的二个文字，在度量中可指明某些在文字中必须不具备的付线.

3. 由于噪声会改变文字线条的起终点位置，在度量中可指明某些付线允许被文字线条所包含.

4. 考虑到文字的缺损，可对一种文字设计几个度量，只要通过其中的一个度量即可.

参　考　文　献

[1] 饭岛泰藏、森健一«日經エレクトロニクス»1972 年 30 号66—80.页人間の識別能力に迫为"OCRASPET/71".

[2] R. B. Hennis The IBM 1975 Optical page Reader pare I: *System Design. IBM. J. R. and Dtv.* pp. 345—420, Semp. (1968).

[3] 齐拔求，文字识别的一种方法，复旦学报——自然科学版，1973，№1 39—50.

OPTICAL CHARACTER RECOGNITION SYSTEM (OCR)

CHOW SHING, CHEIN LOW-CHUNE

(University of Futan, Shanghai)

ABSTRACT

The Optical Character Recognition System of Futan 789 model has been developed for a special demand. The aim of the system is to recognize the alphabet and some special marks which have been machine-printed on certain document papers. The researching target of the system is that under the recognition speed of 100 characters per second both the sabstitution rate and reject rate approaches to 10^{-4} (including the black/red colour recognition substitution rate). Because of the lower case alphabet the system is to recognize and automatically continuously printing in the document-making process, the printing quality varies a great deal. So we are confronted with a difficult situation in the aspeet of character recognizability.

This paper will decribe the developing process of the system and how to select the scheme. Basing on it, a two-stage recognition scheme with the combination of hard ware and soft ware will be set forth. Finally this paper will present the experimental results and the orientation of its improvement.