

手写体字符识别方法的探讨*

代汝为 胡启恒 执笔

(中国科学院自动化研究所)

摘 要

手写体字符识别的一项重要应用是在邮政自动化系统中自动识别邮区编码。本文扼要介绍了应用于我国邮电部信函自动分拣系统中的手写体数字识别机,并对于识别顺序逻辑的设计方法问题进行分析。试图利用有限状态文法推断的一些结果来设计识别顺序逻辑,并考虑用不分明集(模糊集)的一些概念来进行信息压缩。

一、手写体数字识别

模式识别技术的发展对于我国实现四个现代化所具有的重要意义,已经日益为人们所认识和注意了。[1]由于近代微电子学和计算技术的迅速进步,自动化技术今天面临的一个具有战略意义的重要问题已经是如何才能在更高的程度上更为完善地用机器模拟并取代人类的部分脑力劳动。人类的一项基本思维活动是通过认识和理解各种声音、图象、文字、物体等等模式来认识客观世界。用计算机实现某些高级自动化系统必须跨越的最大障碍之一也正是如何使机器能够直接利用原始的信息,即直接认识和理解客观世界中各种模式的问题。这个问题之能否解决,或解决的好坏,常常成为建立某些近代自动化系统的关键。

一个明显的例子就是邮政工业中占用人力最多,劳动强度最大的分拣工作自动化。在相当长的时期内,对这项繁重的劳动只能做到半自动化,即所谓按键入格的第二代分拣机。尽管高速传送带能代替人把信函或邮件送到应去的格子里,可是认识收件人地址并据此按键指挥邮件传送方向的这项高度紧张的工作,机器却不能代劳。这就是因为机器只能认识由按键而产生的码(即通过人的思维活动进行过再次加工转换的信息)但不能直接识别地址(在人类社会活动中产生的原始信息)。只有在模式识别领域内较早得到发展的文字识别技术逐渐趋于成熟的时候,才为解决分拣工作的全自动化问题提供了现实的可能性。例如日本从1967年才开始试用手写体数字识别机自动分拣信函,以后逐年改进,70年代以后逐步成为邮政上普遍应用的自动化手段[6]。

我们曾在识别手写体字符方面进行了一些初步的尝试,于1974年底完成了能识别手写阿拉伯数字的实验装置,对于书写者只有很少的要求,如:不过于潦草,不在字符上添装饰笔划,写在方格内等,都是一般人容易作到的。以后又在此基础上,将此项结果应用

* 参加研制手写体数字识别机工作的有:林鹏,宁汉悦,庞振太,黄美芳,吴峰风,周迪伟,刘天立,洪继光,宋志正等。(本文曾在中国自动化学会1978年年会上宣读)

于全自动信函分拣,在有关兄弟单位协作下,于1977年8月提供了信函分拣手写数字识别机样机。在测定识别机性能时,采用一个CRT扫描装置作为光电转换,用一万个并不了解对书写者有什么要求的群众所写的一万个中上等质量的字样作为测验样品。得到的结果是错识率0.49%,拒识率4.9%速度为每秒钟不低于20字符[2]。

在这个系统中的信息获取部分采用一个积分网络对视频信号进行最佳滤波,取得了很好的效果。并采用全字亮度平均法进行视频信号量化的阈值自动调整以减少扫描失真并使识别机能够适应各种纸张和钢笔、圆珠笔、铅笔等不同的书写工具。在系统的工程实现方面采取微程序控制与部分专用硬件相结合的方案,因此能以相对说来比较简单的硬件实现相当复杂的控制及识别决策功能。

关于这个系统中所用的识别方法,基本的考虑是手写体的字形和字体变化非常大这一个主要矛盾。我们采用经过改进的细化逻辑对文字信息进行信息压缩,得到文字的骨架,并达到突出和强调文字的结构信息的目的。然后从文字骨架中抽取作为粗分组依据的特征,即一个文字的骨架所具有的端点数,三叉点数和四叉点数(假定分别以 α , β , γ 代表)并据此将被识别字符归入某个 (α, β, γ) 组。经过粗分组以后,再对文字骨架进行逐点跟踪,在跟踪过程中逐点抽取由点、线、方向特征构成的特征组。所以跟踪的结果是从一个经过细化的二维文字图形中得到一条一维的特征链,链上的每一个环节则是一个点、线、方向特征组。

识别过程的最后一个步骤是把一个未知文字的特征链当作文字的代表(实际上特征链是对该文字的结构信息的描述)来判定它应归属于所在的 (α, β, γ) 组中的哪一个文字类别。代表每个文字类别的识别顺序逻辑(也就是描述该类文字区别于它类文字的结构共性的一组规则)是根据大量采样事先设计好并存储于机器的记忆装置里的。某个文字类别(例如属于“2, 0, 0”组的“2”字)的理想识别顺序逻辑应能包括凡属于“2, 0, 0”组的所有无限多个“2”字的特征链(实际上只可能是尽量多,多于某个要求的指标),而不包括“2, 0, 0”组中无限多个“非2”字的任何一条特征链(实际上也只可能是尽量少,少于某个要求的指标)。判断一个未知字符的特征链是否包含在某一文字种类的顺序逻辑中的决策过程我们称之为通链识别。

描述某个文字种类的识别顺序逻辑是分析大量的然而是有限数目的样本(包括属于该类的正样本集 S^+ 和不属于该类的负样本集 S^-)所得到的一组规划。凡以预定的特征(例如我们所用的点、线、方向特征)作为基本元素,符合这组规则的特征链,都应该代表属于该类的某一个文字;而描述不同种类文字特征链的构造规则必须是互异的,这种规则可以很方便地用状态流程图来表达。由于顺序逻辑中某些环节可以具有回归的性质,同时,由于这些规则可以任意多次地反复运用,因此用为数不多的规则可以描述一个文字种类的无穷多种变形。在实际的设计中可以做到使识别顺序逻辑在不与异类字相混淆的前提下能够允许有相当宽广的字形变化范围。因此,用一维的特征链和顺序逻辑描述文字并进行通链识别的方法,识别形状变化极大的手写体文字,可以取得很好的效果。国内外有关的工作已表明了这一点。例如日本的东芝公司为信函分拣研制的识别机,先把字符按水平方向加以分割,抽取水平特征,再用水平特征链描述字符。日立公司的手写字符识别机则是采用方向链。

在这一类识别系统中, 顺序逻辑的设计是主要问题, 在不改变预处理和抽特征方法的条件下, 识别顺序逻辑的设计质量是影响识别能力的最重要的因素. 如果从模式识别的语言学方法的角度来考虑, 可以认为设计识别逻辑的任务是一个文法推断的问题 [7, 8]. 目前在一些实际系统中主要是靠直观的方法分析样本而得到识别顺序逻辑, 这是一项技巧性比较强的工作, 而且要花费大量的时间. 如何能用计算机帮助完成这项工作, 这是一个现实而主要的问题. [3, 4] 下面将从文法推断的角度对此问题作一些探讨.

二、识别顺序逻辑与有限状态文法

在构成文字骨架特征链的点、线、方向特征中, 点特征(即端点、叉点及连点)和方向特征(8个跟踪方向)是基本的, 而线特征(即凹、凸或平)则是由方向特征派生的. 所以可以把细化后的文字骨架看作一个由节点和分支构成的有向图. 这样的有向图可以用有限状态文法加以描述.

有限状态文法是形式语言理论中最简单的一种文法. 它是一个四元式:

$$G = (V_N, V_T, S, P)$$

其中 V_N, V_T 分别代表非终止符和终止符的有限集.

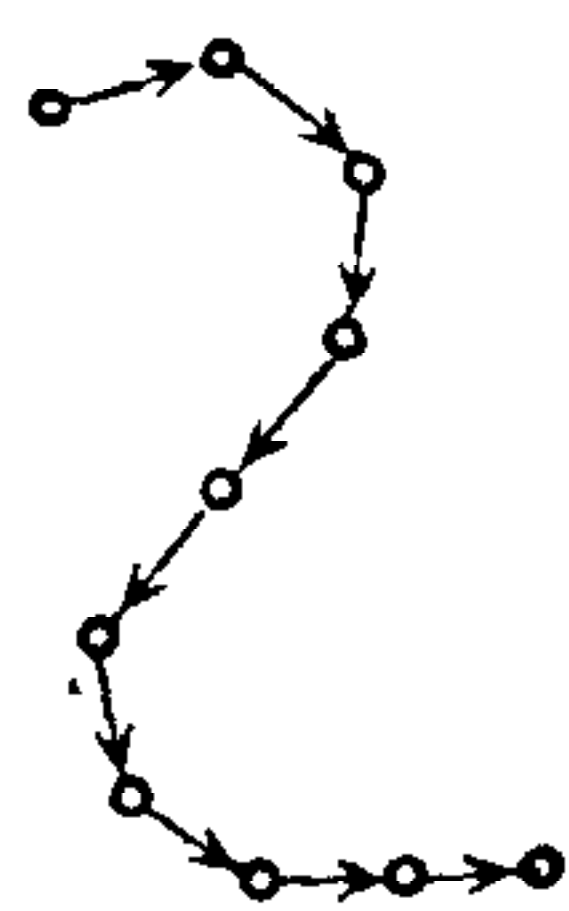
$S \in V_N$, 是一个特定的非终止符, 称为起始符, P 是一组再写规则, 或称为产生式, 有限状态文法的产生式具有下列形式:

$$A \rightarrow aB$$

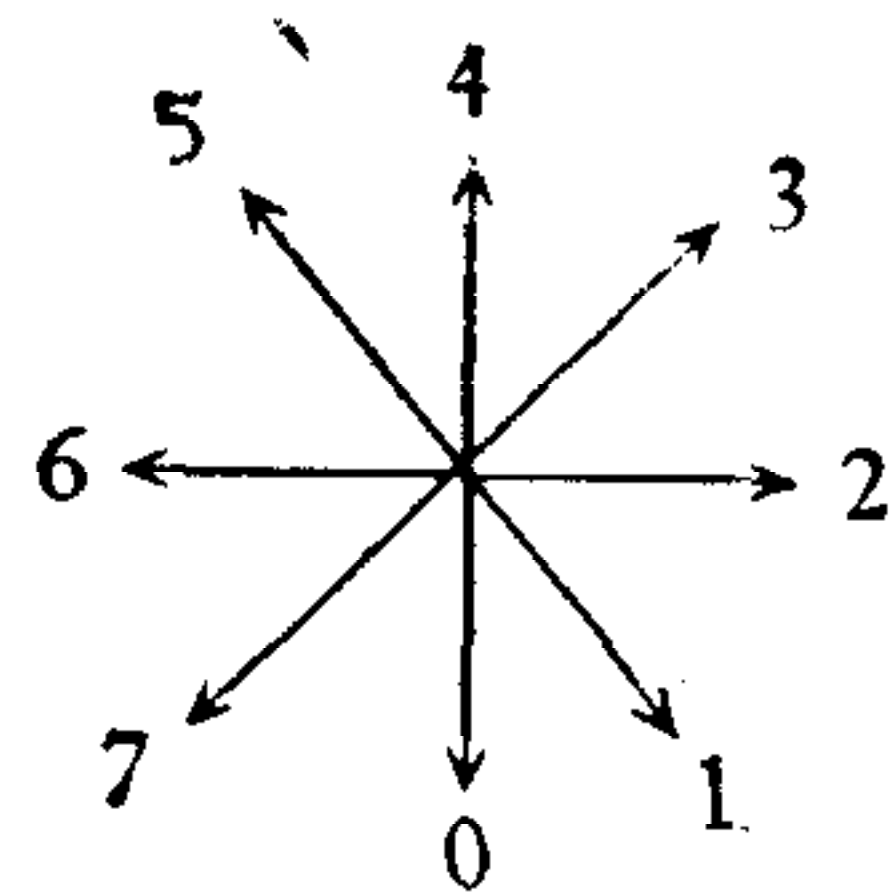
$$A \rightarrow b$$

其中 $A, B \in V_N, a, b \in V_T$.

例如一个细化后的 2 字, (图 1) 跟踪过程中产生方向链 210770122. 用有限状态文法描述.



210770122



方向特征

图 1

$$V_N = \{S, A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8\}$$

$$V_T = \{0, 1, 2, 7\}$$

$$P: S \rightarrow 2A_1, \quad A_1 \rightarrow 1A_2$$

$$A_2 \rightarrow 0A_3, \quad A_3 \rightarrow 7A_4$$

$$A_4 \rightarrow 7A_5, \quad A_5 \rightarrow 0A_6$$

$$A_6 \rightarrow 1A_7, \quad A_7 \rightarrow 2A_8, \quad A_8 \rightarrow 2,$$

这里的再写规则如果把样本集 $S = \{S^+, S^-\}$ 中更多的方向特征链考虑进去并稍加改变, 例如通过合并某些非终止符的方式扩大这个文法, 它就有可能描述更多的具有各种变形的 2 字. 如果用一个状态流程图来表达这一组文法规则, 那么可以认为这就是一种最简单的识别顺序逻辑.

需要加以说明的是上一节介绍的手写数字识别机中的识别逻辑, 远远不是这样简单的, 由于它的转移条件具有扩展的多种功能, 例如对跟踪过程中需要保留的信息加以存储, 留待到达跟踪的另一阶段时用来验证上下文关系; 对于某些具有回归性质的再写规则如 $A \rightarrow aA$ 允许重复运用的次数加以限制 (这常常体现了 S^- 对扩大文法的限制), 等等,

使得它的描述能力和正确区分容易相混的字符种类的能力远比有限状态语法要高。特别是容许考虑上下文关系这一点对于区分相似的字更为重要。在这里讨论有限状态语法是因为相对说来有限状态语法的推断问题研究得比较充分，有一些结果可以利用来作为探讨更高级语法的推断问题的基础。

对于被推断的语法 G 的要求是 G 能产生 S^+ 中所有的链，但不能产生 S^- 中的链。但目前即使是对于最简单的有限状态语法，也几乎无法考虑 S^- 的影响，其结果可能是 G 扩大得过分以致能产生 S^- 中的某些链。

根据 S^+ 推断有限状态语法的途径之一是基于形式微商的概念建立“规范微商有限状态语法”，然后以此为基础再根据某些原则合并非终止符以扩大语法。例如合并状态的原则之一是基于 K -尾等价性。

这里提出一种从尾部建立规范语法的方法进行探讨，然后再以得到的规范语法为基础，用合并状态的办法对语法加以扩大。

用一个例子说明：

例如给定样本集 $S^+ = \{3221007, 2210007, 322007, 221007, 2227, 207\}$

第一步：列出 S^+ 的尾部构造：

- 1 尾 $\{u_1 = 7\}$
- 2 尾 $\{u_2 = 07, u_3 = 27\}$
- 3 尾 $\{u_4 = 007, u_5 = 207, u_6 = 227\}$
- 4 尾 $\{u_7 = 0007, u_8 = 1007, u_9 = 2007, u_{10} = 2227\}$
- 5 尾 $\{u_{11} = 10007, u_{12} = 21007, u_{13} = 22007\}$
- 6 尾 $\{u_{14} = 210007, u_{15} = 221007, u_{16} = 322007\}$
- 7 尾 $\{u_{17} = 2210007, u_{18} = 3221007\}$

u_i 代表语法中的非终止符，凡下面有横线的符号是指起始状态，两个状态间的关系同样由形式微商定义：即，语法规则中包含

$$u_i \xrightarrow{a} u_j$$

的充分必要条件为 $D_a u_i = u_j$

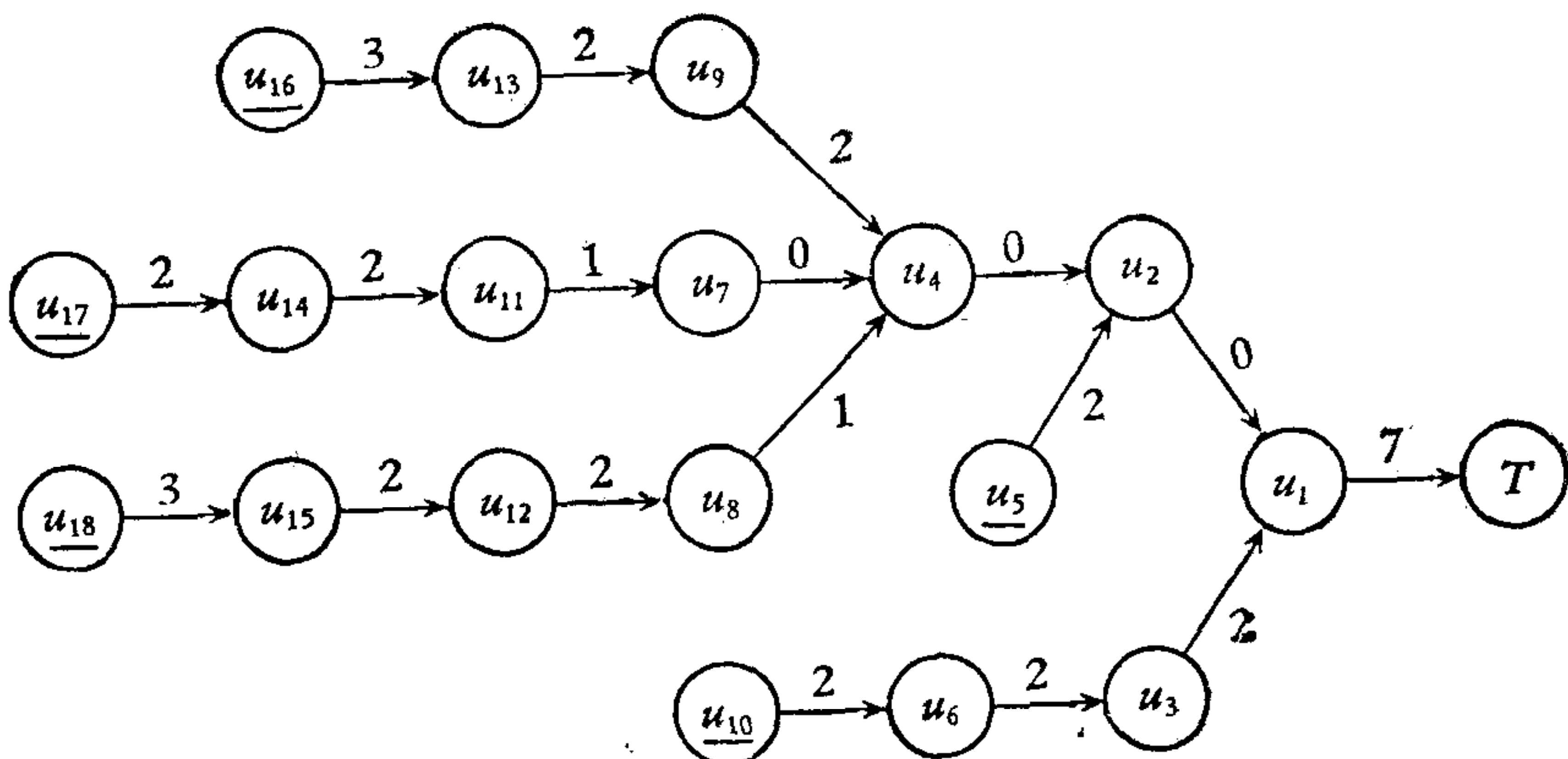


图 2 尾部构造的状态流程图

第二步：于是对给定的 S^+ 得到尾部构造规范文法的状态流程图：

第三步：对状态进行合并以简化得到的文法。合并的原则是，接连出现三次同一个终止符，则可以认为等价于一个回归的状态。所有起始状态应合并在一起。

图 2 中的状态 u_i 可以合并的如：

$D_{000}u_7 = D_{00}u_4 = D_0u_2$ 因此 $u_2u_4u_7$ 合并； $D_{222}u_{10} = D_{22}u_6 = D_2u_3$ 因此 $u_{10}u_6u_3$ 合并等。合并后得到

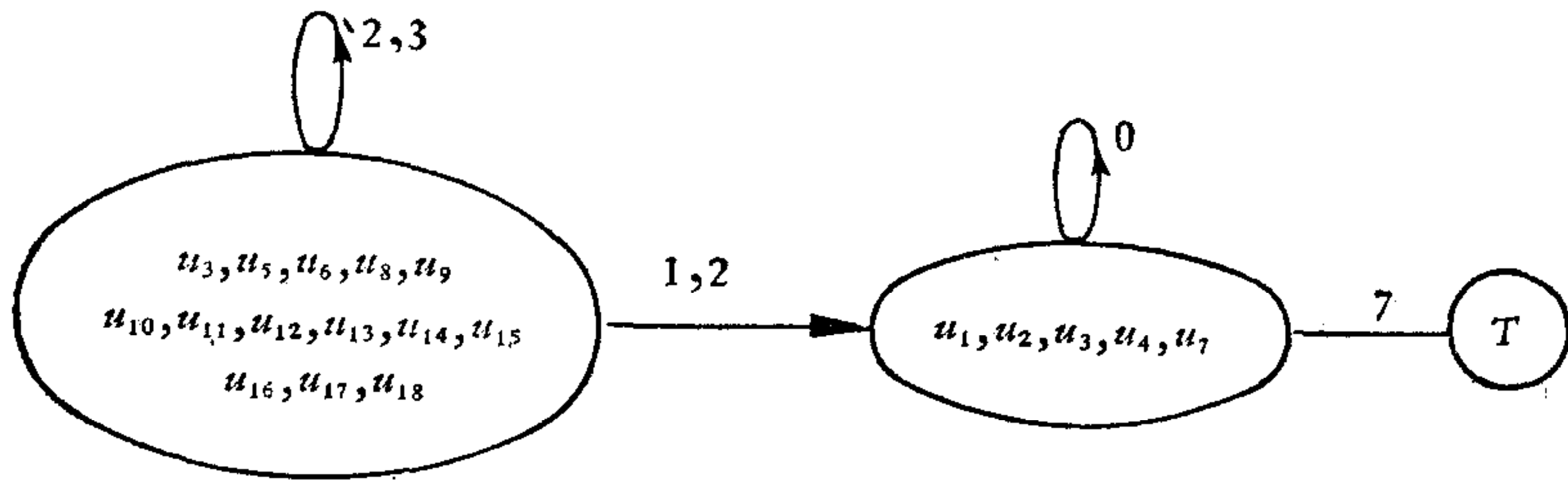


图 3

简化后的状态流程图为

图 4 是一个具有最简单功能的识别顺序逻辑，它不仅描述了给定采样 S^+ ，而且还可以产生具有类似于 S^+ 中的链的结构无穷多的链。

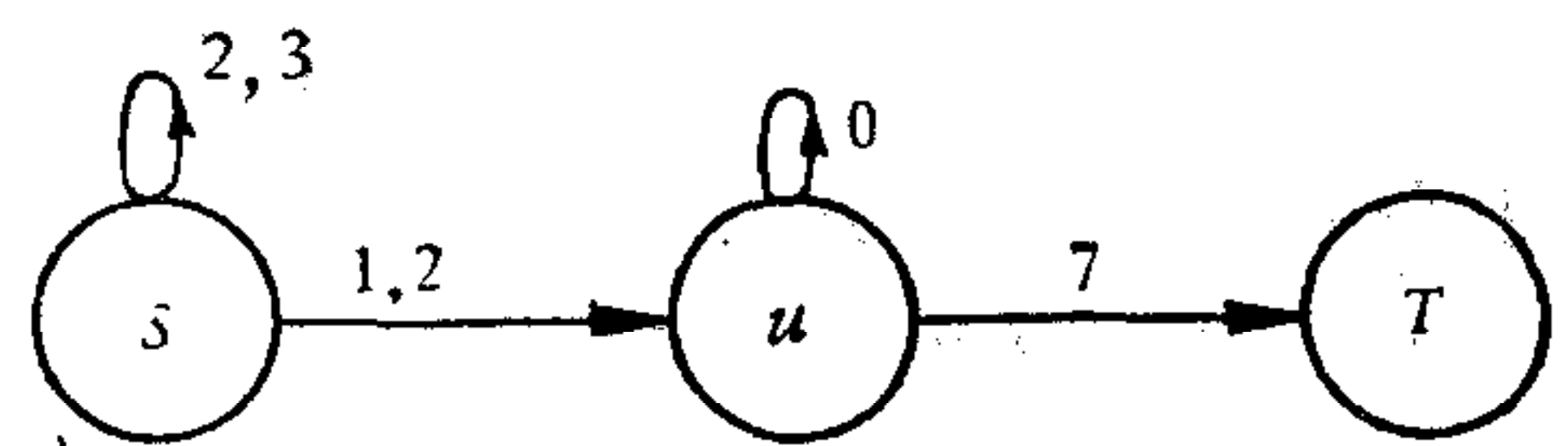


图 4

从以上例子可以看到，从样本集中得到的规范文法，即使是对一个仅包含 6 个极为简单的链的 S^+ ，就已经具有相当多的非终止符（状态），不难设想，对稍具现实意义的采样集来说，其规范文法中包含的状态数目将会有惊人的增加，因而使得无论用 K -尾等价原则或上面介绍的办法进行合并，实际上都会由于计算量的庞大而难于实现，更难考虑对 S^- 信息的合理利用问题。为了试图解决这个问题，下面将利用不分明集的一些简单概念对信息压缩的问题进行讨论。

三、信息压缩

在识别系统中，模式信息的预处理，抽特征与识别决策互相之间存在密切的联系。如果能在抽取最简单最原始的特征（例如方向特征）以后，先采取适当措施进行信息压缩，就有可能使识别顺序逻辑的设计变得简单一些。

从上节的例子可以看出，以 8 个方向特征作为基本元素建立起来的规范文法中，对于 S^+ 中每条链的描述是完全精确的，反映出每条链中每个细节（方向单元），因此在这样的文法中不可避免会存在大量的非终止符。其实这样精确的描述对于解决识别的任务来说是完全不必要的，前面已经提到，为了识别的目的，恰恰是需要不分明的描述。

如果在对字符进行细化处理及粗分组以后，先着手分析文字骨架从端点到端点（或到叉点）之间的一段文字笔划究竟是直线还是曲线，是沿哪一个方向倾斜的直线，是哪种形状的曲线，用不分明的描述定义具有一定意义的线条特征，然后把文法推断建立在这样一些线条特征的基础上，则问题的解决将会现实而有效得多。

例如考虑手写体阿拉伯数字和英文字母可能有以下的各种线条特征:

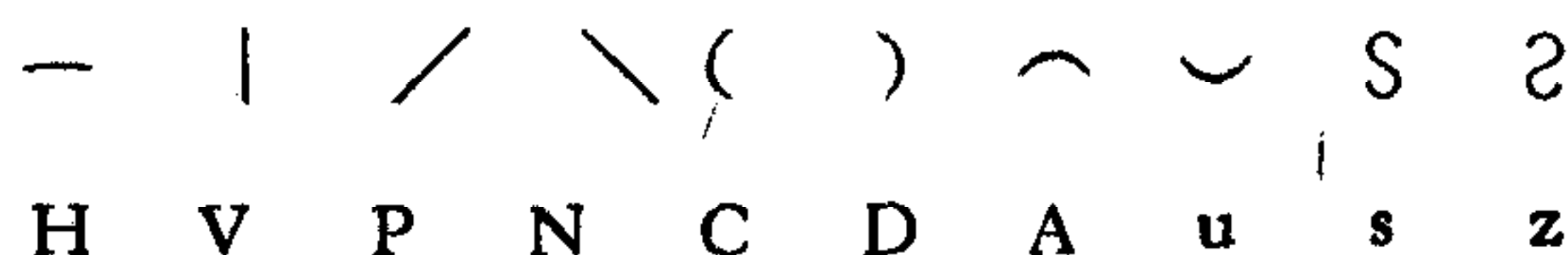


图 5

试图利用不分明集理论的一些简单概念来定义这样的线条特征。

不分明集理论是 1965 年提出来的,它考虑变量在某种范围内变化是可以容许的。对一些控制和识别的问题,用不分明描述比起精确的描述将更为有效。一个不分明集可以叙述如下: 对于一个通集 u , 它的一个不分明子集 A 是由从属度函数 $\mu_A: u \rightarrow [0, 1]$ 来表征的, 对于 u 中的每一个元素 y , 有 $[0, 1]$ 中的一个数 $\mu_A(y)$ 来表示元素 y 在 A 中的从属度, 例如考虑平面上所有通过原点的直线集合, 要定义不分明水平线, 用 $m(x)$ 表示直线 x 的斜率, 那么过原点的直线 x 属于水平线的从属度:

$$\mu = \begin{cases} 1 - m(x) & \text{当 } |m(x)| \leq 0.5 \\ 0 & \text{当 } |m(x)| > 0.5 \end{cases}$$

当 $m = 0$, $\mu = 1$, μ 的值为 1, 即表示水平线; 当 $|m| = 0.5$ 时, $\mu = 0.5$ 。这时 x 已经不能较好的反映水平线了, 这里容许有一定程度的不分明性。当从属度函数 μ 在 0.5 与 1 之间时, 则认为 x 在某种不分明意义上是水平线。这样就用 $[0, 1]$ 区间内的数给出了接近水平线的从属度。

关于由点 E 至 N 的一个线段 b 究竟是直线还是曲线的问题, 可以根据 EN 与 b 所包围的面积 S 来确定, 即定义下列函数:

$$\mu = \begin{cases} 1 - \frac{S}{2\sigma} & \text{如果 } S \leq \sigma \\ 0 & \text{如果 } S > \sigma \end{cases}$$

其中 σ 是一个给定的阈值。如果 S 满足 $0.5 \leq \mu < 1$, 则认为 b 是某种意义上的直线, 否则是曲线, σ 是一个可供选择的参数, 例如可以选择具有一定深度的凸或凹所范围的面积作为 σ 。

对于 b 是直线的情况, 还要解决 E 和 N 的连线究竟属于 H 、 V 、 P 、 N 中的哪一种情况。如果用 Q 表示线段 EN 与 x 轴所成的角, 对于过原点在 x 轴上下 22.5° 范围内的直线认为是水平线, 则可以定义从属度函数如下^[9]:

$$\mu_H(\theta) = 1 - \min \left\{ \min [|\theta|, |\pi - \theta|] / \left(\frac{\pi}{4}\right), 1 \right\}$$

$$\mu_V(\theta) = 1 - \min \left\{ \min \left[\left| \frac{\pi}{2} - \theta \right|, \left| \frac{3\pi}{2} - \theta \right| \right] / \left(\frac{\pi}{4}\right), 1 \right\}$$

$$\mu_P(\theta) = 1 - \min \left\{ \min \left[\left| \frac{\pi}{4} - \theta \right|, \left| \frac{3\pi}{4} - \theta \right| \right] / \left(\frac{\pi}{4}\right), 1 \right\}$$

$$\mu_N(\theta) = 1 - \min \left\{ \min \left[\left| \frac{3\pi}{4} - \theta \right|, \left| \frac{7\pi}{4} - \theta \right| \right] / \left(\frac{\pi}{4}\right), 1 \right\}$$

线段 b 的两端 EN 的连线与 x 轴的夹角 θ 使上面 4 个从属度函数中的哪一个数值最大, b 就属于哪一种线条特征, 这样一来, 就有相当的不分明性。

对于曲线类的线条特征, 包括 C, D, A, U, S 和 Z 可以进而分为近似垂直的 C, D (即 EN 连线近似于垂直线), 近似水平的 A, U 以及 S 和 Z . 主要取决于 EN 两点连线的性质. 至于判断 b 是直线或哪一类曲线, 可以通过从 E 到 N 逐步跟踪时计算面积. 例如

跟踪到线段 b 上的一点 M (见图6), 那末线段 \widehat{EM} 与 x 轴所包围的面积为

$$A(\widehat{EM}) = \sum_{i=1}^M \frac{1}{2} (y_i + y_{i-1})(x_i - x_{i-1})$$

这里 $(x_i - x_{i-1})$ 是常量, 但可以为正或负值. 直线 EM 与 x 轴所包围的面积为 $\frac{1}{2} (y_m + y_0)(x_m - x_0)$ 于是线

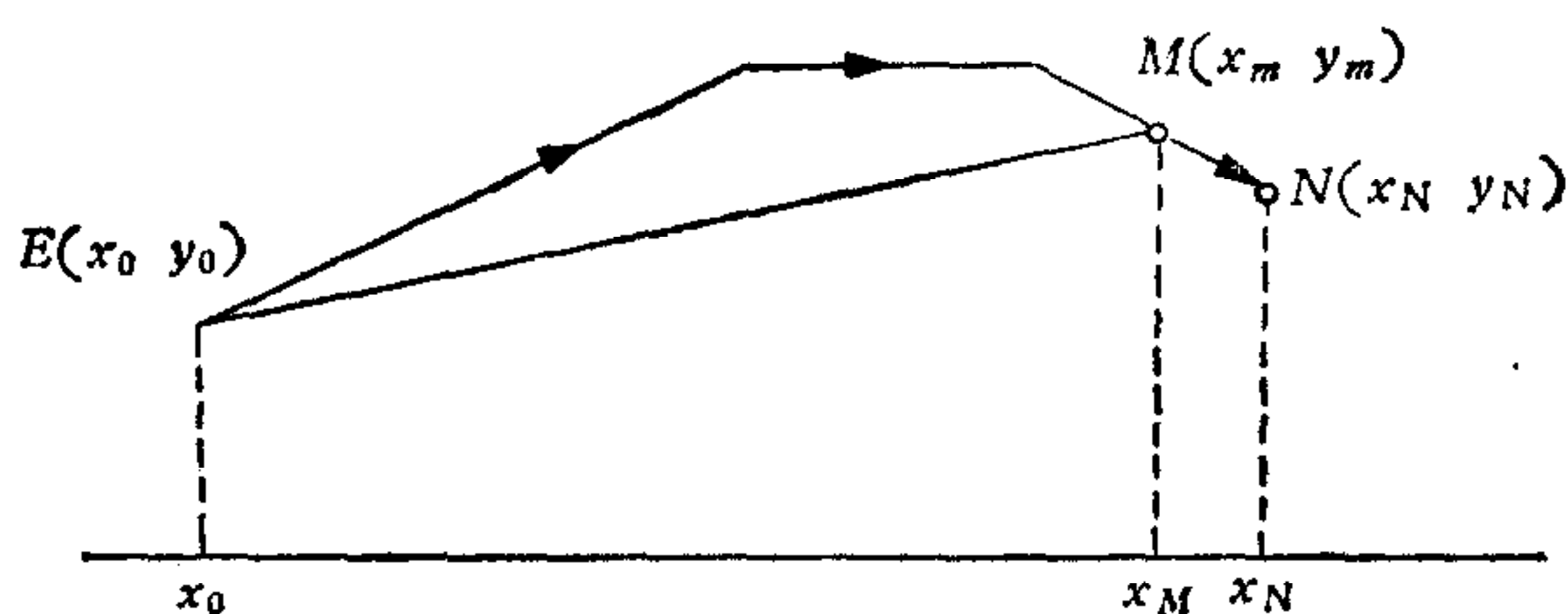


图 6

段 \widehat{EM} 与直线 \overline{EM} 所包围的面积为.

$$S(\widehat{EM}) = \frac{1}{2} \sum_{i=1}^M (y_i + y_{i-1})(x_i - x_{i-1}) - \frac{1}{2} (y_m + y_0)(x_m - x_0)$$

每跟踪一步就可求出相应的面积 S , 在求 S 的过程中, 可能得到正值或负值. 这里不再详细列举了.

至于字符骨架中包含圆环的情况, 这时需要考虑的特征为:

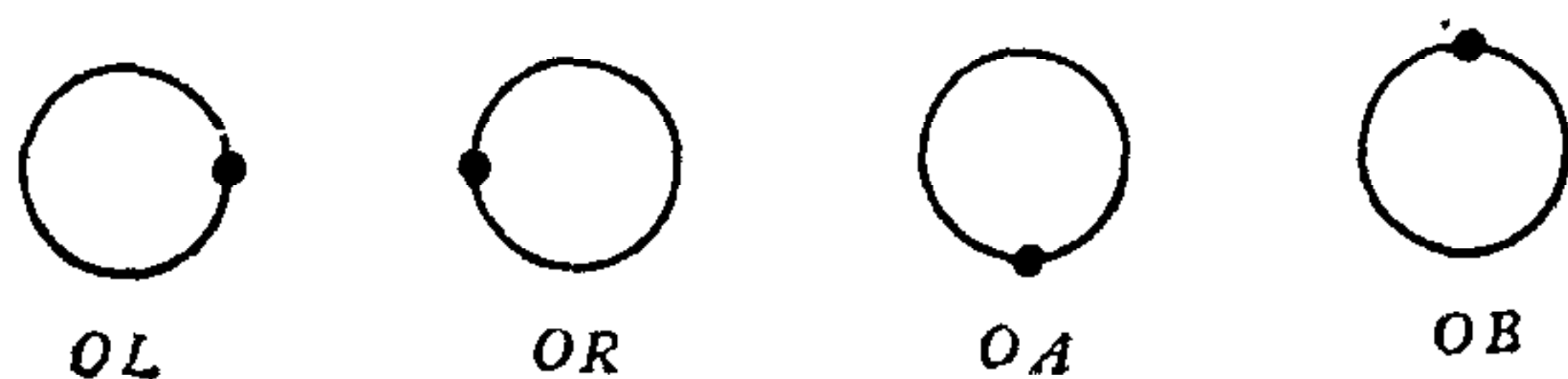


图 7

即位于叉点之左圆 (OL), 位于叉点之右圆 (OR), 位于叉点之上圆 (OA), 位于叉点之下圆 (OB), 可以借助叉点 (x_0, y_0) 和封闭曲线的重心 (x_c, y_c) 的连线的斜度来确定.

如果按上述方法把端点到叉点 (或端点) 的一段折线表示为线条特征, 那么识别顺序逻辑就可以大大简化. 如果对于由线条特征构成的特征链集合解决文法推断问题, 将获得相对来说比较简单的规范文法, 从而使得通过合并某些非终止符以简化文法的问题以及合理估算负样本集 S^- 的限制作用的问题都比较容易考虑了. 至于这些问题的解决, 则尚有待于进一步的工作.

参 考 资 料

- [1] 中国科学院自动化研究所, “图象识别”, 《国外科学》第一集, 科学技术文献出版社 1977 年 1 月.
- [2] 中国科学院自动化研究所二室四组, “手写字符的识别”, 《科学通报》1978 年第 4 期.
- [3] 山本真司等, “手書き数字認識理論の設計”《电子通信学会论文志 C》oct 1970.
- [4] 山本真司等, “辞書の自動作成”《情報処理》1976 年 7 月号.
- [5] R. O. Duda, P. E. Hart Pattern Classification and Scene Analysis New York Wiley 1973.
- [6] Mori K. Genchi, H. et al. Microprogram Controlled Pattern processing in a Handwritten Mail Reader-Sorter “Pattern Recognition” Vol. 2, No. 3, Sept 1970
- [7] K. S. Fu Syntactic Methods in Pattern Recognition Academic Press, 1974.
- [8] K. S. Fu and T. L. Booth Grammatical Inference. Introduction and Survey I. IEEE Trans Vol. SMC-5 Jan. 1975.
- [9] P. Siy and C. S. Chen, IEEE Trans SMC-4, No. 6 (1974).

THE INVESTIGATION OF HANDWRITTEN CHARACTERS RECOGNITION METHODS

TAI JU-WEI HU CHI-HENG et al
(*Institute of Automation, Academia Sinica*)

ABSTRACT

One of the important application of handwritten characters recognition machine is to read postal codes in the automatic letter sorting. In this article, we shall briefly introduce a handwritten numeral recognition machine which is used for automatic letter sorting, and discuss the design of sequential recognition logics by means of results of grammatical inference of finite state grammar, as well as the idea of Fuzzy sets for information compression.