第 20 卷 第 1 期
1994 年 1 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 20, No. 1
Jan., 1994

简 报

交替运用两种信源模型的汉字识别法

张彩录 郭宝兰 张宇桐 韩 勇 张宇铮

(河北大学汉字信息处理研究室 保定 071002)

关键词: 汉字识别,相关处理,信源模型.

1 对现有汉字 OCR 的分析

如果用 C. E. Shannon 信息论的观点,对当前各种汉字 OCR 的识别方法进行理论概括,可把它们归结为“无记忆信道对无记忆离散信源的信息传输模型”. 在汉字 OCR 中,被识别的汉字集合 A 与每个符号相应的概率为:

$$A = \{a_1, a_2, a_3, \cdots, a_i, \cdots, a_n\}$$

$$P(a_1), P(a_2), P(a_3), \cdots, P(a_i), \cdots, P(a_n) \tag{1}$$

按这种观点,解除一个待识汉字的不定度平均所需的信息量为:

$$I_1 = -\sum_{i=1}^{n} P(a_i)\log P(a_i) \tag{2}$$

但在多数汉字 OCR 中,特征提取、模式分类,只注重汉字的光学信息,并未重视语言的统计规律. 这等效于把集合 A 中的各个符号看作是等概率的. 这时(2)式将简化为(3)式,并满足(4)式所表达的关系.

$$I_2 = -\log P(a_i) \tag{3}$$

$$I_2 \geqslant I_1 \tag{4}$$

这说明 OCR 如果只注重汉字的光学信息,就会使解除待识汉字的不定度所需要的信息量为最大. 试比较 27 个英文字符(包括空白)和国标 6763 个汉字的等概率熵和一维熵:

$$H_{e1} = \log 27 = 4.75\text{bit} \tag{5}$$

$$H_{e2} = -\sum_{i=1}^{27} P_i \log P_i = 4.03\text{bit} \tag{6}$$

$$H_{c1} = \log 6763 = 12.72\text{bit} \tag{7}$$

$$H_{c2} = -\sum_{i=1}^{6763} P_i \log P_i = 8.8\text{bit} \tag{8}$$

其差值分别为:

本文于 1992 年 3 月 15 日收到.

$$D_e = H_{e1} - H_{e2} = 0.72\text{bit} \tag{9}$$

$$D_c = H_{c1} - H_{c2} = 3.92\text{bit} \tag{10}$$

如果英、汉均采用等概率模型，则识别一个汉字比识别一个英文字符需多开销 7.97bit 的信息．如果英、汉都采用一维概率模型，则英文相对等概率模型将节约 17.86% 的信息开销，汉字将节约 44.54% 的开销．这说明文字识别的研究，即使都采用无记忆信源模型，对于汉字识别来说，更应注意一维概率模型的运用，以减少巨大的信息开销．

2　Markov 信源模型的运用

把汉字 OCR 待识字符集 A 看作 Markov 信源，从信息开销的角度将得到更多的好处．如 A 为 m 阶 Markov 信源，其符号表及其转移概率为：

$$A = \{a_1, a_2, a_3, \cdots, a_i, \cdots, a_n\} \tag{11}$$

$$P = (a_i / a_{j1}, a_{j2}, \cdots, a_{jm}) \quad i, j = 1, 2, \cdots, n \tag{12}$$

这时某汉字的发生概率依赖于其前的 m 个汉字．该种信源的状态数为：

$$C = n^m \tag{13}$$

对于汉字识别，当 m 值大时，信源的状态数会大的惊人．为使问题简单，可选择 $m = 1$ 这时信源的状态数为 n，如果考虑的是常用汉字即 $n = 6763$ 则对于汉字的一阶 Markov 信源模型，消除一个待识汉字的不定度平均所需的信息量为：

$$I_3 = -\sum_{i=1}^{n}\sum_{j=1}^{n} P(a_i, b_j)\log P(a_i / b_j) \tag{14}$$

容易证明：

$$I_3 \leqslant I_1 \tag{15}$$

参考文献[2]所提供的英文一维熵 4.03bit，二维熵 3.32bit，借助它们的比例关系，类比求出汉字的二维熵，即消除一个待识汉字的不定度所需要的信息量为 5.74bit，可见采用一阶 Markov 模型在信息开销方面比一维概率模型可以得到更多的好处．这反映在特征提取上，可以采用较少维数的特征便能识别出相同集合中的汉字．交替运用两种信源模型的汉字识别法，就是采用一维概率模型识别单字，采用一阶 Markov 信源模型完成相关识别的一种方法．

3　算法及实现

在识别中交替使用两种信源模型，需解决如下两方面的问题：(1)相关识别，(2)单字识别．一般的识别系统，只使用汉字的光学信息，象“大、太、犬、丈”等相似字在多维特征空间中将靠的很近，这给识别带来很多麻烦．如果依据文本汉字之间的约束关系，合理地收集某汉字之后的后继字，并依此为根据采用非特征相关分类，相似字的识别就容易得到解决．图 1 是两种分类方法的特点比较示意图．

由于相关识别只把待识文本看做一阶 Markov 信源，这显然是忽略了文本汉字之间的弱相关联系，保留了强相关部分．这种只处理强相关的办法，使得一句话或一段的起

始字处于较难运用相关分类的状态. 这部分汉字仍需采用一维概率模型进行识别. 为提高系统的识别效率，在两种信源模型的交替使用中，必需使系统更多的处于相关识别状态. 实验显示，这种系统在最不利的情况下所能达到的识别指标，与一维概率模型的识别结果是一致的. 一般情况下均能给出优于一维概率模型的识别结果. 该算法已在 AST386/20M 系统上实现，在提取同样特征的情况下，对普通印刷品上的汉字，一维概率模型识别率达到 92—93% 时，交替使用两种模型时可达 95—97% 的识别率，平均识别速度可达 1380 字/分.

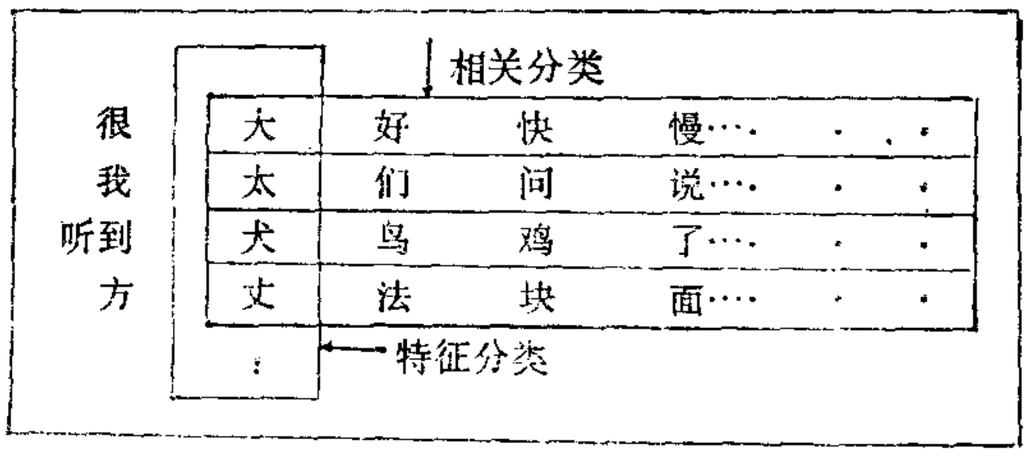

图 1 相关分类与特征分类的比较

参 考 文 献

[1] 郭宝兰、张彩录: “汉字识别中正确识别率与识别速度的探讨”通信学报，7(1986),(5),53.
[2] Normam Abramson, Information theory and coding, McGraw-Hill Book Company Inc, (1963), 33—35.
[3] 赵伯璋,徐力;计算机中文信息处理,宇航出版社,(1987),29.

A CHINESE CHARACTER RECOGNITION METHOD USING ALTERNATELY ZERO MEMORY AND MARKOV SOURCE

ZHANG CAILU, GUO BAOLAN, ZHANG YUTONG, HAN YONG, ZHANG YUZHENG
(*Chinese Information Processing Lab. Hebei University, Baoding* 071002)

Key words: Chinese character recognition; relative handling; source model.