第 12 卷 第 3 期
1986 年 7 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 12, No. 3
July, 1986

用于机器识别和学习的汉字表达式

夏 莹　　张炘中[1)]
(清 华 大 学)

摘　　要

本文提出限制性手写汉字的形式化描述——汉字表达式，扩充了文法规则 BNF 范式的元符号．利用汉字表达式，用 PASCAL 语言编写了限制性手写汉字的识别和学习程序．

一 汉字表达式

对于限制性手写汉字，同一汉字因不同的人书写往往造成模式的显著变化，用模板匹配来识别是很困难的，因此应该就汉字这个特殊的平面图形，根据汉字结构的内在规律，以人识字的先验知识建立模糊的模型，并用启发式搜索法去识别．

在用抽取笔划有序列法识别限制性手写汉字方法中[1]，把汉字笔划有序列看作一个句子，汉字的横、竖、撇、捺、左折、右折、方、叉($\{h,s,p,n,z,y,f,c\}$)作为基元，把由若干句子组成的集合称为语言 L．一种语言代表一个汉字，第 i 类语言 L_i 代表第 i 个汉字，语言 L_i 用形式化的文法规则 $G_i(V_T,V_N,P,S)$表示，常把 L_i 写成 $L(G_i)$．当用 BNF 范式 (Backus-Naur Form) 描述手写汉字规则时，因元符号只有〈 〉，::＝，|，( )，{ } 等，所以规则显得冗长和繁琐．对于笔划相对位置要求严格的部分，笔划有序列可明显地反映出来；对于笔划的相对位置要求不严格的部分．由于书写时是任意的，且由于计算机在识别时，笔划的顺序是根据严格的自上而下，自左而右的次序扫描次序，因此同一个汉字可得到的笔划有序列有多种合理的可能，即多码问题．有的字的笔划有序列码可高达 1000 种以上，这种汉字的文法规则用 BNF 范式描述时就显得很繁琐．例如手写的“心”字，对它各笔划的起笔上下位置没有严格要求，且笔划的长短、类型也允许有些变化，所以“心”的文法为：

〈心〉::＝ *pynn*|*pnyn*|*pnny*|*ypnn*|*ynpn*|*ynnp*|*npyn*|*npny*|
nypn|*nynp*|*nnyp*|*nnpy*|*synn*|*snyn*|*snny*|*ysnn*|
ynsn|*ynns*|*nsyn*|*nsny*|*nysn*|*nyns*|*nnys*|*nnsy*|
nynn|*ynnn*|*nnyn*|*nnny*，

可见很繁琐，因此笔者提出用汉字表达式形式化描述限制性手写汉字．每条汉字表达式的左边是一个非终结符，右边是用终结符和元符号组成的串．汉字表达式中元符号的定

本文于 1984 年 11 月 14 日收到．

1) 参加本项工作的还有杨德顺、李勤同志．

义如下：

〈 〉表示在〈 〉中的对象符号为非终结符.

::=表示此符号左边的字符用其后边的符号串来定义.

[ ] 是码集符. 它里面的元素(终结符或表达式)是无序的，且元素的符号可以重复出现.

/ \ 表示它里面的元素(终结符或表达式)是“或”的关系.

( ) 表示分界.

例：若终结符 $V_T=\{h,s,p,n,z,y,f,c\}$，则心、有、感这三个字的汉字表达式为：

〈心〉::=[/*pns**ynn*]，代表二十八种合法的笔划有序列.

〈有〉::=*ph*[*sz*]/(*hh*)(*nh*)\，代表四种合法的笔划有序列.

〈感〉::=[*yn*][*hp*]/(*hfp*)(*hpf*)(*phf*)\[/*psn**ynn*]，代表 336 种合法的笔划有序列.

汉字表达式的优点在于：

(1) 它扩充了 BNF 范式的元符号. 当模式结构规则中有的地方要求有序，有的地方无序时，能够精练地描述. 把汉字笔划间次序不确定的部分放入码集符内，可去掉对识别无用的非稳定信息，减少识别字典占用的存储空间，提高判别速度. 元符号[]起缩写作用，在模式规则描述中是有用的.

(2) 在汉字表达式字典中，每个汉字(或每组汉字)用一个独立的汉字表达式表示，这样便于在识别时利用分类技术和汉字的使用频度，提高常用汉字的识别速度.

(3) 便于作文法推断，可由实例学习，并归纳出汉字表达式.

二、限制性手写汉字表达式的识别和学习

当限制性手写汉字用表达式描述时，其识别和学习框图如图 1 所示. 程序用 PASCAL 语言编写. 识别的基本原理是，扫描纸上的手写汉字，数字化后得到二值汉字点阵，预处理后用启发式搜索方法抽取笔划成分，经笔划判别、合成以及三重部件分离得到笔划有序列，然后到汉字表达式字典中查找，并识别出汉字. 对于少数异字同笔划有序列，再经辅助判别后可得到识别结果，为国家标准汉字交换码. 对于 1000 个常用汉字，每个汉字平均约有 17 个笔划有序列，平均每个字的判别时间为 60ms.

机器学习功能是指计算机在设计者或用户的教授下，系统能不断增添和改善所拥有

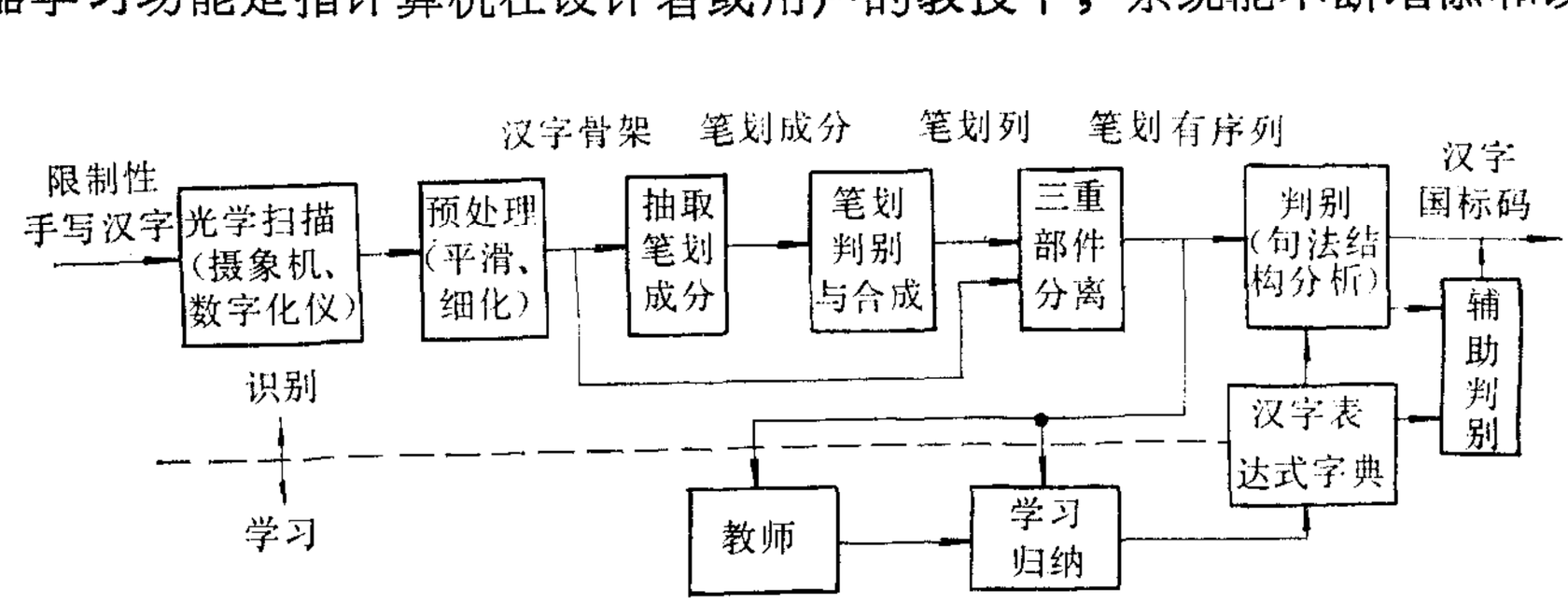

图　1

的汉字知识. 初级的学习是由设计者或用户直接向系统提供汉字规则——汉字表达式. 高级的学习是用户提供正、反训练示例,由机器归纳出汉字表达式. 一般的文法推断要求供系统学习的正示例必须是完整的. 对于手写汉字能提供的正示例很难是完整的, 因此要求计算机有更高的学习能力,有一定的归纳能力,并可以把正反示例上升为概念[2]. 因汉字表达式是独立的,计算机可以由实例归纳出汉字表达式.

参 考 文 献

[1] Zhang Xinzhong, Xia Ying, A Method of Recognizing Handprinted Chinese Characters by the Extraction of An Ordered Sequence of the strokes, Procedings of 1982 International conferece of the Chinese-Language Computer Society, p. 388—398.

[2] Edited by Paul R. Cohen and Fdward A. Feigenbaum, The Handbook of Artifical Intelligence, Volume III, Heuristech Press Stanford, California, 1982.

THE EXPRESSIONS OF CHINESE CHARACTERS USED IN MACHINE RECOGNITION AND LEARNING

XIA YING ZHANG XINZHONG

(Tsinghua University)

ABSTRACT

A formal description of handpritten chinese characters—— the expressions of chinese characters is presented in this paper. The metalanguage symbols of BNF of grammar rules are generalized. By means of the expressions of Chinese characters, a program for the machine to recognize handpritter Chinese characters and learn experessions are given in PASCAL.