手写印刷体汉字识别方法2-D EAG

赵明

(中国科学院软件研究所)

HAND PRINTED CHINESE CHARACTER RECOGNITION METHOD 2–D EAG

Zhao Ming

(*Institute of Software, Academia Sinica*)

ABSTRACT

This paper introduces 2-D EAG method for hand printed Chinese character recognition. The main works are: Propose Two-Dimensional Extended Attribute Grammar (2-D EAG for short) method, in which both bottom-up reduction and top-down deduction are realized for two-way information transmitting and control; Propose a recognition scheme utilising directly two dimensional information, to avoid information losses caused by linearization of features; Provide polysemous grammars, coexisting grammars and structure inferrings which build special linkages among grammars and distinguish among similar samples by structural analogy; Besides, realize also redundant reduction scheme, stroke segment drawing algorithm with elastic tracing of contour lines, and redundant guided component drawing algorithm. The recognition experiment with 100 categories of hand printed Chinese characters is conducted and very good result is achieved in overcoming large scale structural distortions (including some cursive writings).

摘 要

本文介绍手写印刷体汉字识别方法2-D EAG。主要的工作为：提出了二维扩展属性文法模式识别方法，可实现自底向上归约和自顶向下推导双向信息传递和控制；提出了一种直接利用二维信息进行识别的方式，可避免特征线性化造成的信息丢失；提出了多义文法、共生文法和结构推断三种在文法之间建立联系，利用结构类比区分极相似字的

1989 年10 月19 日收到，1990 年2 月21 日定稿。本文得到国家自然科学基金资助。

识别算法; 提出了多冗余归约机制, 双边缘弹性跟踪笔段抽取算法, 多冗余有引导部件抽取算法。用2-D EAG 方法对100 字种实际手写汉字进行了识别实验, 在识别大畸变汉字(包括部分连笔字) 方面得到了很好的结果。

§1. 引　　言

汉字识别的工作自六十年代开展以来, 至今已取得了很大的进展。由于其巨大的实用价值及其大样本空间识别的特殊性, 这一课题吸引了众多的研究者, 形成了模式识别领域内一个重要的分枝。近年来, 随着中文信息处理、办公室自动化、机器翻译等方面工作的开展, 进一步促进了对汉字识别设备的需要, 推动了研究工作的进展[1]。

按照识别方式和识别对象的不同, 汉字识别又可分为联机手写体识别、印刷体识别和脱机手写体识别三类。由于近年内不可能实现无限制手写体的识别, 目前一般所说的手写体, 是指加以某些限制的手写体, 或叫做手写印刷体。

由于要同时处理基元级的噪声和结构级的畸变, 手写印刷体汉字识别很难直接利用联机识别和印刷体识别中已获成功的识别方法, 而不得不更多地依靠对汉字字形结构的分析来提高对噪声和畸变的容纳能力, 同时仍保持对精细结构的分辨能力。但这样一来, 就在一定程度上把模式分类问题变成了模式理解问题, 而后者是一个比前者困难得多的任务, 目前这一方面的基础研究还非常薄弱。

本文的主要工作是提出了一种用于识别手写印刷体汉字的二维扩展属性文法方法(以下简称2-D EAG), 并利用这种方法进行了初步的汉字识别实验。提出这一方法的基本动机及所要解决的主要问题是:

1. 汉字是一个二维图形, 而一般的文法只能以一维的方式接受输入符号, 这种一维化过程丢失了很多有用的二维结构信息。为了克服这一缺点, 本文提出了一种在二维平面上进行文法归约的识别方式。

2. 一般的文法方法是一种串行的决策过程, 稳定性比较差。本文利用属性文法提供的继承属性描述汉字的整体结构关系, 用于引导和约束低层的文法归约过程, 降低了对各级决策正确性的依赖。

3. 在一般的汉字识别方法中, 两个汉字的相似程度只能用一个实数来度量, 这对于区分极相似字是很不利的。为此, 本文提出了描述和区分极相似字的多义文法、共生文法和结构推断方法。

4. 由部件构造汉字这一级存在着大量的信息冗余, 对孤立字识别来说, 利用这一信息冗余是很有帮助的。为此本文提出了多冗余归约的文法归约方式。

5. 利用2、3 两种能力, 2-D EAG 可以在归约过程中动态决定是否要对某个局部进行识别及要对哪几种情况进行区分, 对存在冗余结构的汉字, 不需要基元一一匹配成功。这对于识别结构复杂的汉字是很有用的。

本文第二节讨论手写汉字识别中存在的问题, 第三节介绍2-D EAG 方法的设计思想, 第四节介绍2-D EAG 的总体结构, 第五节报告识别实验的结果及分析。

§2. 手写汉字识别中存在的问题

对于识别数量多达六七千的中国手写汉字来说，统计方法的分辨能力明显地显得不足，若要完成细分，必须更多地依靠结构方法。但使用结构方法必须考虑下列的几个问题，并能给出初步的解决办法:

1. 粘连和断线问题。尽管这个问题往往只涉及几个象点，但对笔道抽取的影响却是巨大的。由于这个问题的存在，以笔道个数为特征是不可靠的。

2. 笔段编码问题。很多笔道抽取方法在抽出笔道线段(笔段) 之后将其编码为横竖撇捺之一，部件抽取也类似地唯一给以一个编码，这是结构方法不稳定的重要原因之一。

3. 自顶向下和自底向上(目标驱动和数据驱动) 问题。由于汉字数量太大，单纯目标驱动的耗费是不可忍受的；同样由于畸变太大，单纯的数据驱动也是靠不住的。因此必须寻求两种驱动方式的最佳结合。

4. 学习问题。对于结构方法来说，学习一直是一个薄弱环节。从描述方面来说，希望能力越强越好，但从学习方面来说，文法越复杂则学习越困难。

5. 相似性度量问题。对识别畸变模式来说，这是一个极为重要的问题。由于结构方法不连续，相似性度量一直未能很好解决。在这一方面，统计方法也是很困难的。统计方法中虽然可以使用各种距离函数，但对于在不同字的不同部位允许不同的畸变程度这一问题，也没有什么好的解决办法。

§3. 2-D EAG 的基本设想及所要解决的问题

针对手写印刷体汉字的特点[2] 及上节提出的问题，2-D EAG 方法采取了相应的对策。

3.1 笔段抽取

2-D EAG 方法以笔段作为基元，抽取笔段之后记录其二顶点坐标，并按圆周24 等分为其编码。为了方便后阶段处理，抽取笔段后构造了一个笔段关系图。同样为了减少关系编码造成的信息丢失，笔段间的关系是多重的，可以表示多种可能。

3.2 汉字的层次描述及文法终结符

结构方法中汉字结构的描述通常分为两类，一类把汉字直接看作由笔道(笔段) 序列组成，没有中间层次；另一类按照汉字的概念结构分解部件，直到分解为笔道(笔段)。前一种方法的好处是可避免部件分割及抽取的麻烦，但匹配计算量较大，更为严重的问题是，由于直接用笔道描述汉字，必须在笔道之间排序，从而又引入了一种新的不稳定因素。后一种方法尽管抽取部件较为困难，但汉字的描述比较简单，部件之间的关系比较稳定，而且可以利用部件组字这一级的信息冗余。基于这些考虑，2-D EAG 采用了层次的汉字描述方式。

在文法描述中，需考虑文法终结符集合的设定，这是基元抽取与文法归约的界面。2-D EAG 不象通常的结构方法那样以笔段作为文法终结符，而是以部件作为文法终结符。其原因是: 由于使用方法的不同，笔道的描述变化很大，用笔道做终结符的文法方法难以兼容各种不同的基元抽取方法。若上升到部件这一级，就比较容易用一种统一的方式描述了。用部件做文法终结符可以把文法的处理与基元抽取尽可能地隔离开来，从而有助于独立进行两方面的研究。

3.3 部件的识别

为了提高部件识别的稳定性，部件识别这一级采用有引导(目标驱动) 的识别方式。由于部件数量不多(约100 多个，更低一层的基本结构仅20 多个)，这种花费是可以容忍的。部件抽取阶段并非只取一个最佳候选，而是抽出所有可能的部件留待后阶段处理，这就更进一步提高了部件抽取的可靠性。

2-D EAG 只从四角抽取部件，对包围结构的部件，只抽取外框部分。这样做的理由是: 1. 汉字信息量较为丰富的地方是在边框部位; 2. 由于识别对象是单字，边框笔道粘连较少，抽出较为可靠; 3. 汉字中存在着大量的冗余信息，一般用不着取出所有部件后再做识别(尤其是对结构复杂的字)。

3.4 模糊度量

在手写印刷体识别中，待识字与模板的相似程度的度量是很困难的。按统计方式计算距离可能是失准的，按照模糊方式度量虽然可以更好地反映人类识别的主观性，但如何度量仍然是一个问题。在计算机上，尽管对位置关系的相似性度量已经有了很多算法，但对更重要的拓朴结构(不是图论意义上的拓朴结构)，却没有什么成功的算法。

考虑到这种情况，2-D EGA 中的模糊度量主要是放在部件这一级。计算模糊隶属度值考虑到的因素有: 横竖的方向和平直度、笔道交接处的完好程度、非关键笔段的有无对该部件的影响程度，等等。隶属度值低于阈值的部件不作为下一步文法归约的候选。

3.5 二维的文法归约

抽取部件之后，并不对部件排序(由于每个角上可能有多个候选，大小不一，实际上也无法排序)，而是按照各部件的位置直接在二维平面上归约，这就是把这种文法称作“二维”的原因。这种归约方式与黑板结构[3] 中的“拼板游戏”很相似。由于位置属性带在结点本身，因此位置算子就不需要了; 由于取消了树结点间顺序关系的限制，文法树的结构也不是十分重要的，归约实际上可以在任意两个部件之间进行。归约成功的部件再用于归约，直到归约成整字。这种做法摒弃了文法方法按符号串匹配的识别方式，增强了对畸变的容纳能力。

对结构复杂的字，只取四角部件可能无法自底向上地完成归约。对这种情况，2-D EAG 根据已取出部件及其在字框中的位置，在文法定义的引导下，建立相应的结构推断树，然后根据需要再做进一步的内部部件识别。

3.6 相似字的区分

相似字的存在给识别造成很大的困难，很多方法不得不对相似字做特殊区分处理，显得非常麻烦，2-D EAG 由于部件抽取不完全，归约到整字级可能会有多个候选字。作为细分手段，提出了多义、共生文法这两个概念及与之相应的文法处理模式。在存在多个候选的情况下，根据开始符号上的语义信息，确定要对哪几个字、对字框中的哪个部位做进一步的区分，然后在继承属性的引导下，沿树下降到要做区分的结点，启动相应的语义函数做出区分，得出最后识别结果。由于利用了扩展属性文法的语义处理能力，2-D EAG 可以把需要做的工作定义为文法结点上的语义函数，从而在获得较强的灵活性的同时，仍保持了控制结构的规整性。考虑到模糊度量的可靠性不高，2-D EAG 首先依靠可靠性较高的部件组合关系进行区分，然后再利用部件的模糊隶属度。

§4. 2-D EAG 的总体结构

如流程图(图1) 所示，首先对字形点阵轮廓进行边缘跟踪、拟合，获得轮廓折线段。然后，在双侧轮廓折线的引导下，抽取汉字的笔段并建立笔段关系图。在此基础上，从左上、右上、左下、右下四个方位抽取基础部件，构造文法终结符，填写相应的属性变量。文法终结符构造好之后，即可进行文法归约。由于只是从四角取出基础部件，归约最多只做两次。对于结构简单的字(如口、十), 取基础部件就可得到候选字，此时即可转入终级识别(流程图中虚线所示)。归约完成之后，根据已做归约的成功情况和基础部件覆盖整个字框的情况，决定是否要做结构推断。如果需要，则进行结构推断。最后，终级识别对归约和结构推断得到候选字进行判断，给出识别结果。

有关2-D EAG 的详细实现，见[4-6]。

§5. 手写印刷体汉字识别实验结果及分析

识别实验工作在主频10 兆赫、内存1 兆字节的DUAL 68000 上进行，程序用C 语言书写。实验使用字样除自行收集、经HP 扫描仪输入的字样外，还使用了湖北自动化所胡家忠制作、录制在磁带上的字样，大小均为64×64 点阵。

为了检验2-D EAG 方法的识别能力，选择了100 个汉字进行了实际的识别实验(表1)。这些字中包括具有各种部件组合方式的字(日、胡、古、枷、屋、旬、周、固等), 若干极相似字(汨、汩、干、千、于等), 需要或不需要做多义、共生区分和结构推断的字(共、其、钢、锏、钥、南、葫、撂等)。由于基本结构和基础部件抽取相当于已做了粗分类，因此所选择的实验集合可具有充分的代表性。对五个人写的共500 个字(包括部分连笔字), 得到了80% 的正确识别率。

从识别结果来看，文法归约和控制算法具有很强的排除不可能候选，区分极相似字的能力(图2、3)。

部件抽取算法可以在畸变较大、甚至存在连笔的情况下正确地抽出笔道，完全不受部件大小、位置偏移的影响，但当笔段误差过大时取不出正选部件(图4)。

笔道抽取算法对粗细不同(笔宽变化可从1 到8 个象点) 及弯曲的笔道工作良好，因而可用于字母数字或行书的笔道抽取，但当有污团时不能正确地抽取笔道(图5)。

综观整个识别过程，笔道抽取是一个最薄弱的环节，在这方面还需进行改进。例如，笔段抽取也可以采用多冗余方式，考虑直接从象点抽取基础部件，等等。

§6. 结　论

本文提出并初步实现了2-D EAG 手写印刷体汉字识别方法，进行了实际的识别实验。工作的结果表明:

1. 利用汉字的构成知识和整字的结构关系对识别大畸变汉字是一种强有力的手段，它可以在很大程度上补偿基元抽取阶段造成的失误。

2. 多义、共生文法和结构推断提供了一种在汉字之间进行比较的有效手段，它在相似字之间建立了特殊的联系，有助于区分极相似字。

3. 多冗余归约是提高识别方法稳定性的有效措施，其实现简单，效率也比自顶向下

的回溯方式高。

4. 以基础部件作为基元抽取和文法归约的界面是一个合理的选择。基于这一界面，高层的结构分析可以独立研究，摆脱了完全依附于基元抽取的从属地位。同时，基元抽取研究仍可以独立进行，并可以期望得到高层信息的引导。

本文工作中存在的问题表明：要开发实用的手写印刷体汉字识别方法，仅从狭义的模式识别观点出发寻求好的分类算法的做法是难以胜任的。由于存在着大幅度的畸变，对它的解释只能从视觉的结构分析的角度着手，同时这一分析还必须克服大量存在的图象级噪音的影响。因此，按照Ballard [7] 划分的从广义图象到结构关系这四个层次，每个层次都要有胜任的处理方法并且互相补充。然而到目前为止，数值近似类和结构分析类方法各只能适应于一端，它们之间的结合仍存在着很多问题。本文在这一方面进行了探索，但要得到较为满意的结果，还有很多工作要做，尤其是基本理论方面的工作。

本文是在童锺美教授的精心指导下完成的，在工作中还得到了戴汝为教授的帮助，樊建平同志在汉字属性库方面进行了合作，实验用部分字样是湖北自动化所胡家忠同志提供的，本文的编辑和打印分别使用了张旭波开发的PLED 编辑软件和曾云峰开发的SP排版软件，在此一并致以谢意。

附 录

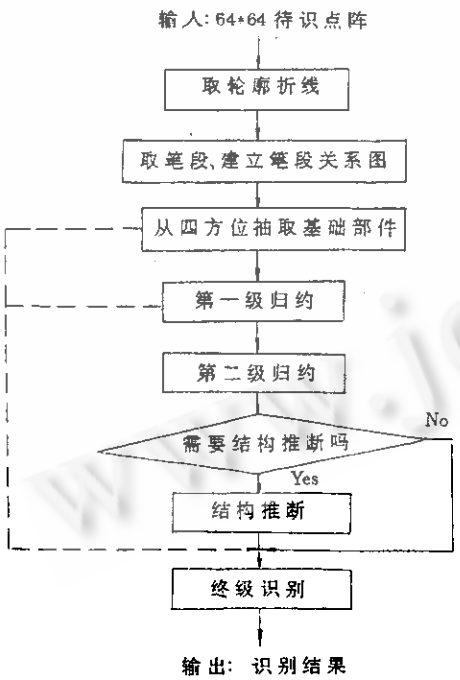


图1 2-D EAG 识别流程

表1 100 个识别实验字

口田吕品只识江扛杠日曰汨汩古右吉土士干千于石计汁讨叶汗讦芊芋共其基拱棋栱明晴目睛腈相木林呆杏杳枷南钢铜钥撂搞居厕厘句甸过迁迂门问间闲周回固困咽扪搁啃腩捐涓葫胡葫谰洁诘刚肌虫钊铡铝期朝潮滴摘壁旬几凡讥讽

(a) (b) (c) (d) (e)

图2-1 部件位置变化对识别的影响 a-d 正确识别，e 被拒识

图a 方位1取出辶、讠，方位2取出工、(干、于)，归约中曾产生上半部件讦但最后只有迂成功归约。

图b 方位1取出扌、女，由于不存在拈字，只有"姑"能够成功归约。若待识字为如，则此时要靠部件扌、女各自的模糊隶属度值来决定。

(a) (b)

图2-2 利用部件组合关系的识别

方位1取出部件𠂉、钅、𠂉、𠂆，由于钅的存在，𠂆被筛掉。

图2-3 本方位内筛选

图2 文法归约识别的例子

不管部件日写成长或扁，由于只有一个"昔"字，故不必再区分日、曰。

即使门内部件无法识别，但由于具有门结构的只有"阑"字，因而已能唯一确认。

尽管门内部件无法识别，但由于只需在"阑、扪"之间做出选择，因而问题得到了简化(判断门内有无部件)。

图3-1 多义区分 图3-2 共生区分

(a)

图a 根据已取出部件"月"进行结构推断可唯一确认"朝"(此时与利用四角编码识别相似)。

(b)

图b 结构推断得到如图所示推断树，表明需要进一步识别口、丷(或区分同、冏)才能识别。

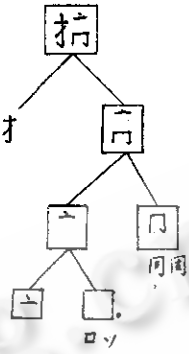


图3-3　结构推断

图3　利用结构类比的识别

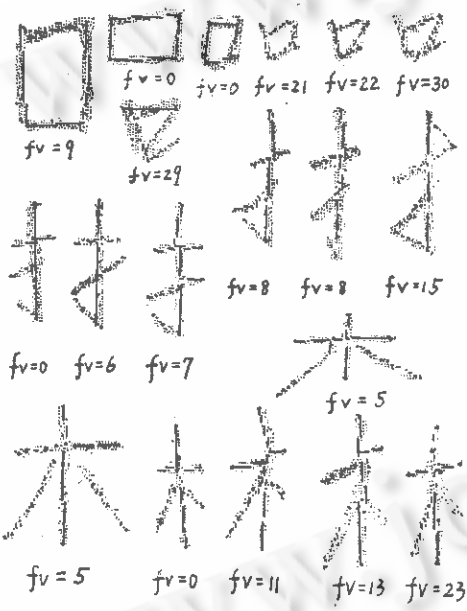
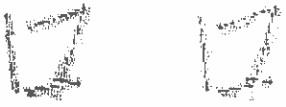


图4-1　成功取出的部件，fv是部件的模糊隶属度值



(a)　(b)

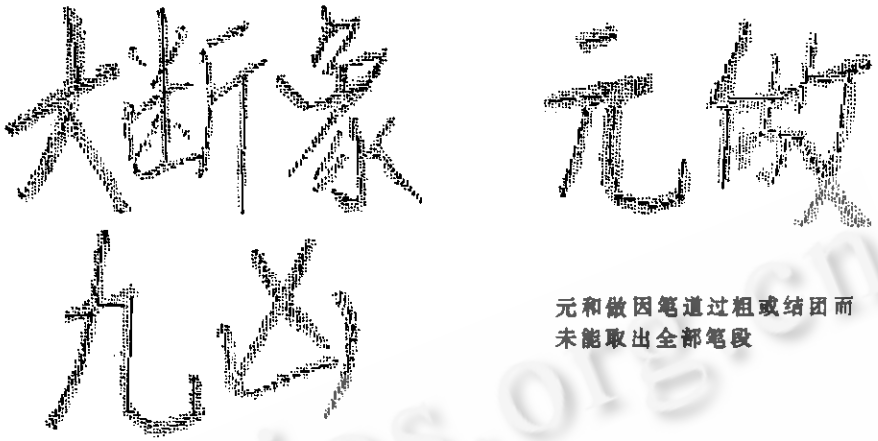
图a 中因左上角"丨"断开距离超过阈值而未能取出"口"，图b则能正确取出(fv=26)

(c)　(d)　(e)

图c-e　因"十"结构破坏未能取出正选部件

图4-2　部件抽取失败

图4　部件抽取的例子



元和做因笔道过粗或结团而
未能取出全部笔段

图5　笔段抽取的例子

参考文献

[1] 张中，我国汉字识别研究的进展，中文信息学报，Vol.1, No.3, 1987。

[2] 赵明，手写印刷体汉字识别方法综述，第三届全国汉字及汉语语音识别学术会议，1989。

[3] H.P.Nii, Blackboard System, The AI Magazine, Summer, 1986.

[4] 赵明，识别手写印刷体汉字的二维扩展属性文法方法，中国科学院软件研究所博士论文，1988。

[5] 赵明，手写印刷体汉字部件的抽取，中文信息学报，Vol.2, No.4, pp.59-64, 1988。

[6] 赵明，识别手写汉字的二维扩展属性文法中的文法归约，计算机学报，Vol.13, No. 7, 1990。

[7] D.H.Ballard and C.M.Brown, Computer Vision, Prentice Hall, New Jersey, 1982.

(上接第4 页)

我们有$S(L)=S_{f'}(L)$, 从而$S(L)$ 是ω-NTB 语言，由定理4 知$r(S(L))\leq 2^{r(L)-1}+1$。

3) 设τ 为正则替换。对于任意的$a\in\Sigma$, 设确定的$fsaA_a=(Q_a,\Sigma_a,\delta_a,S_a,F_a)$, $T(A_a)=\tau(a)$。令$S_{f'}=(Q,\Sigma',\Sigma',\delta,S_0,F')$ 其中$Q=\{S_0\}\cup\left(\bigcup_{a\in\Sigma}Q_a\right)$

$$\Sigma'=\Sigma\cup\left(\bigcup_{a\in\Sigma}\Sigma_a\right)$$

$$F'=\{F\in 2^Q\mid S_0\in F\}$$

以及δ: i) $(S_0,a,\lambda,S_a)\in\delta,\quad a\in\Sigma$,

ii) $(q,\lambda,b,q')\in\delta,\quad q,q'\in Q_a,\quad iff\quad \delta_a(q,b)=q',\quad b\in\Sigma_a$,

iii) $(q,\lambda,\lambda,S_0)\in\delta,\quad q\in F_a$

我们有$\tau(L)=S_{f'}(L)$, 从而$\tau(L)$ 是ω-NTB 语言，由定理4 知$r(\tau(L))\leq 2^{r(L)-1}+1$。

参考文献

[1] S, Ginsburg and E. H. Spanier, "Finite-Turn Pushdown Automata", SIAM formal on control, Vol. (1966), No. 3, pp. 429–453.

[2] 郭清泉，陈力行，"有穷转向的ω 前后文无关语言"，山东大学学报(自然科学版)，1986 年，第2 期。

[3] 郭清泉，"ω 超线性语言"，山东大学学报(自然科学版)，1987 年，第1 期。

[4] R. S. Cohen and A. Y. Gold, "Theory of ω-Languages", JCSS, Vol. 15 (1977), No. 2, pp. 169–208.