文章编号: 1001-2486 (2000) 05-0093-05

基于主动知识库的地形图汉字注记智能提取方法*

徐战武, 刘肖 琳

(国防科技大学机电工程与自动化学院, 湖南 长沙 410073)

摘 要: 地形图中包含了大量的字体丰富的汉字注记, 正确提取并识别这些汉字是图纸处理中的关键组成部分。简要分析了交互提取及自动提取两种方法的优缺点, 提出并实现了一种基于主动知识库的结合两者优点的汉字注记智能提取方法, 取得了很好的应用效果。

关键词: 地形图; 汉字注记; 主动知识库; 人机工程学

中图分类号: TP391.4 文献标识码: B

Intelligent Chinese Character Extracting Method for Topographic Map Based on Active Knowledge Base

XU Zhan-wu, LIU Xiao-lin

(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: There are a great number of Chinese character annotation with different size in topographic map. Extracting the Chinese character annotation in scanned topographic map image is the key condition and foundation. The advantage and shortcoming of interactive extracting method and automatic extracting method is analyzed briefly. An intelligent extracting method based on the active knowledge base is presented and realized, combining the advantages of both. The algorithm is efficient.

Key words: topographic map; Chinese character annotation; active knowledge base; human engineering

纸质地图的扫描矢量化是地理信息系统的主要数据来源。地形图中包含了大量的字体丰富的汉字注记, 正确提取这些汉字是图纸处理中的关键组成部分。目前, 汉字交互提取方法以其简单直观性占主导地位, 但对于复杂的地图, 工作量很大。另一方面, 注记自动提取的研究也取得了较大的进展, 文献[1-3]中的算法, 对于简单的字符具有较好的效果, 但对于诸如粘连等的复杂情况就难以适用; 文献[4-5]中介绍了地形图中汉字、粘连汉字的提取算法, 取得了很好的效果, 对于一般的汉字具有 85% 以上的提取率, 但对于很复杂的情况, 仍然无能为力。低效率的矢量化手段是制约地理信息系统发展的瓶颈之一。

1 基本原理

一般而言, 计算机适合于大量的简单的重复运算, 而人擅长于执行启发性的、相互联系的任务, 但在大量的数据和重复任务方面则表现很弱。汉字交互提取以人为主, 把所有的工作都交由用户处理, 而自动提取则充分依赖计算机。交互汉字提取对于简单的对象每个都进行处理, 降低了整体效率, 而这正是自动汉字提取的优势所在。

地图是按照制图规则综合而成的空间拓扑图, 制图规范作为一个综合的知识库决定了各个对象之间的相互关系。应用人机工程学的原理, 提出了基于主动知识库的智能汉字提取方法, 以充分发挥人的主观能动性和计算机的快速计算能力。将复杂和简单的工作分别交给人和计算机来完成: 由自动提取算法提取和识别出简单的对象, 它具有很高的正确率; 由人来交互地提取和识别出复杂的对象。人的交互提取部分与汉字自动提取算法并不是分开的, 而是相互影响, 相互促进的并行实体, 人(交互提取部分)和计算机(自动提取程序)共享一个统一的主动知识库, 每一部分都能不断地更新知识库, 知识库统

* 发稿日期: 2000-03-03
基金项目: 国家预研基金项目(16.9.1.4)
作者简介: 徐战武(1973-), 男, 硕士。

一指导每一部分的下一步汉字提取工作。两部分以知识库为纽带和核心, 不停地发挥作用, 直到将所有的对象都提取识别出来, 并努力做到两个部分的无缝连接。图 1 示意了这一过程。

2 主动知识库系统的设计

2.1 主动知识库系统的模型[6,7]

一个主动知识库系统(aKBS) 定义为一个传统知识库系统(KBS) 之外加一个事件驱动的规则库, 简称事件库(EB) 及其相应的事件监视器(EM) , 即: aKBS= KBS+ EB+ EM。在这里 KBS 中主要存放着当前事实、状态、规则等陈述性知识; 事件库由系统和用户定义的各种事件驱动的规则组成, 事件库的规则具有以下的一般形式:

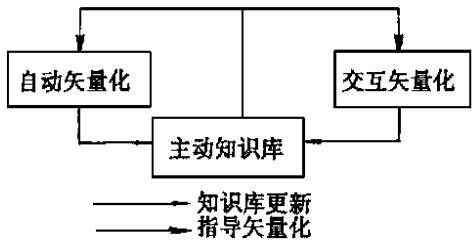

图 1 汉字智能提取示意图

Fig. 1 The sketch of intelligent chinese charcter extracting method

RULE < 规则名> [< 参数> , …, …]

WHEN < 事件表达式>

IF < 条件 1> THEN < 动作 1>

…

IF < 条件 n> THEN < 动作 n> , n ≥1

上述事件驱动规则的语义是, 一旦< 事件> 发生, 计算机就主动触发执行其后的 IF- THEN 规则, 也可定义为, 一旦< 事件> 发生时, 就主动触发执行其后的规则集合(规则库) , 按产生式系统执行的模式(即匹配- 解决冲突- 执行循环) 来解释执行相应规则集合。在一个主动知识库中, 一方面包含了称为“被动知识”的传统知识库, 另方面包含了称为“主动知识”的能根据事件的发生主动激活执行的事件库。事件监视器一直主动地监视着事件库, 一旦发现某个事件发生, 就立即触发执行其后的规则。

如图 2 所示的基于主动知识库系统地形图汉字智能提取系统, 交互提取和自动提取是知识库的获取知识的两种不同手段, 对知识库而言, 两种汉字提取手段是统一的、透明的。当用其中的交互或是自动汉字提取得到新的事实的同时, 将生成一个事件, 事件监视器捕获到这个事件后, 在事件库中匹配相应的规则, 然后触发执行规则所确定的动作。推理和执行机构将根据当前知识库的状态, 进行相应的处理, 处理的结果很可能得到新的知识或状态用来指导下一步的汉字提取(包括自动汉字提取与交互汉字提取) 。从一个汉字提取动作开始, 产生一个事件, 事件监视器捕获该事件, 匹配规则, 触发一系列的动作, 得到新的知识, 并可能产生新的事件, 如此循环反复构成一个闭合的逻辑环路, 直到汉字提取工作完成为止。

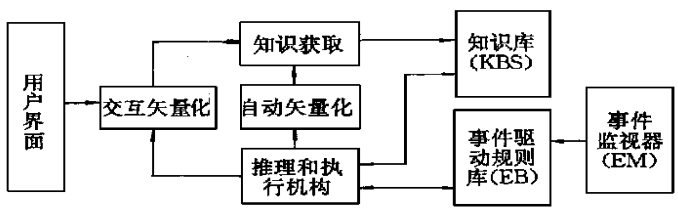

图 2 基于主动知识库的汉字智能提取系统

Fig. 2 Intelligent Chinese character extracting system

2.2 知识库的设计

在地形图中, 汉字名称注记一般由 3~ 4 个汉字组成, 一个完整意义的一组汉字注记具有相近的尺寸, 水平、垂直和雁行字列标注的汉字注记, 在空间上具有均匀性。根据这一制图规则和常识, 建立一个简单的知识库。知识库中保存着所提取出的汉字, 以及汉字词组。每一个汉字注记作为一个描述性事实知识, 采用框架表示, 具体有位置、大小、相邻汉字链表、所属的词组性等槽, 如下所示:

< 汉字注记>

< 位置> < X> < 值 11> < Y> < 值 12>
< 大小> < 宽度> < 值 21> < 高度> < 值 22>
< 相邻汉字表> < 值 31, …>
< 汉字词组> < 值 41>
……

汉字词组则保存一指针, 标明该汉字所属的词组。由 KBS 负责维护该知识库的完整性。

2.3 事件库与事件监视器的设计实现

在汉字提取算法中, 考虑“汉字插入”, “汉字删除”这两个主要事件。当提取出一汉字时, 并将汉字插入到知识库的同时, 还将生成“汉字插入”事件; 事件监视器捕获该事件, 根据当前汉字的位置、大小信息, 触发汉字近邻检查、可组合性检查等动作。动作的结果将很可能找到新的“候选汉字”, 指导进一步的提取工作。同样地, 产生“汉字删除”事件时, 将校验知识库中其它汉字的相邻特性, 以及组合特性等。

事件驱动规则可用下列方式实现:

RULE < 汉字插入> [< 汉字位置> , < 汉字大小>]
WHEN < 汉字插入>
IF < 左(右, 上, 下) 相邻项为空> THEN < 在左(右, 上, 下) 边检测已存在的相邻汉字>
IF < 左(右, 上, 下) 相邻项为空> THEN < 在左(右, 上, 下) 边检测潜在的相邻汉字>
IF < 相邻项空间上均匀> THEN < 插入/ 更新汉字词组>
……
RULE < 汉字删除> [< 汉字位置> , < 汉字大小>]
WHEN < 汉字删除>
IF < 左(右, 上, 下) 相邻项不空> THEN < 左(右, 上, 下) 相邻项的右(左, 下, 上) 相邻项置空>
IF < 相邻项空间上不均匀> THEN < 更新/ 删除汉字词组>
……

规则中, 各个条件项是逐个顺序执行的。汉字新插入时, 它的各相邻项为空。事件监视器采用一个独立的线程来实现, 它始终监视着是否有“汉字插入”、“汉字删除”这两个事件产生, 一旦捕获了其中的某一事件, 则立即进行规则匹配, 并触发相应的动作序列。

2.4 汉字自动提取与交互提取

汉字自动提取与交互提取都是将相应的对象从图中提取(识别) 出来。自动提取由程序自动地定位候选汉字的位置, 然后用一定的模式识别算法进行判断提取, 而交互提取则应用了人机工程学的原理, 采用交互点击的手段, 直观快速地确定候选汉字的大致位置, 然后也采用相同的模式识别算法判断提取。图 3 示意了自动汉字提取与交互汉字提取的区别与联系。

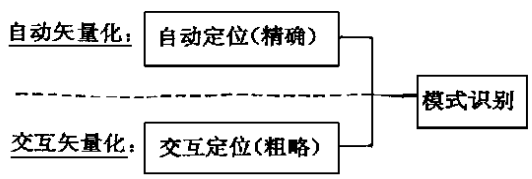

图 3 汉字自动提取和交互提取的联系与区别

Fig. 3 Chinese character automatic extraction and interactive extraction

汉字除少数独体字外, 均由基本笔画构成部首, 再由若干部首组合为单字。汉字数目繁多, 笔划由简到繁, 绝大多数的汉字由几个部件构成, 没有一个明显的特征, 并具有丰富多样的字体, 但其结构原则基本相同, 大体上符合平衡稳定、布白均匀、参差有变的原则。紧紧扣住汉字的方块性及结构均匀性这一关键特点, 通过分析判断各连通成份组合成的图形对象是否满足汉字的特点来提取出候选汉字。另一方面, 针对地形图中水平与垂直字列的汉字注记间隔比较小的特点, 在提取出的汉字注记周围搜索是否有粘连汉字, 若有则去除粘连, 并提取之。在 1: 50000 的地形图中, 大多数的汉字注记是村庄名称, 所以主要对这级大小的注记实现自动提取。具体算法可参见文献[4]、[5]。

3 实验与结论

如图 4 所示, (a) 是原始扫描图黑色分版图, (b) 是利用一次自动提取算法后提取出的汉字, 可以看到在(a) 中共有 60 个大小不一的汉字, 相应的在(b) 中提取出了 55 个汉字, 有 5 个汉字被拒取, 其中有 4 个字(‘ 赵’ 、‘ 滋’ 、‘ 齿’ 、‘ 门’) 是由于与其它图形对象相粘连的, 可见汉字正确提取率达 92%。(c) 是应用了基于主动知识库的智能提取算法后的提取结果, 实际上, 在 5 个拒取的汉字中, 只是交互地提取出‘ 国’ 、‘ 齿’ 、‘ 滋’ 三个汉字, 而另外的两个汉字则利用主动知识库通过触发“ 汉字插入” 事件自动得到。可见基于主动知识库的智能提取算法大大减少了交互提取的工作量, 效果显著。

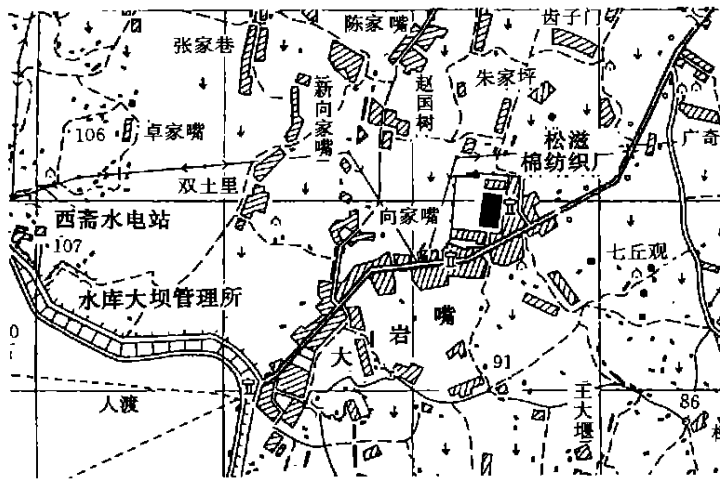

图 4 (a) 原始扫描地形图

Fig. 4 (a) Original topographic map

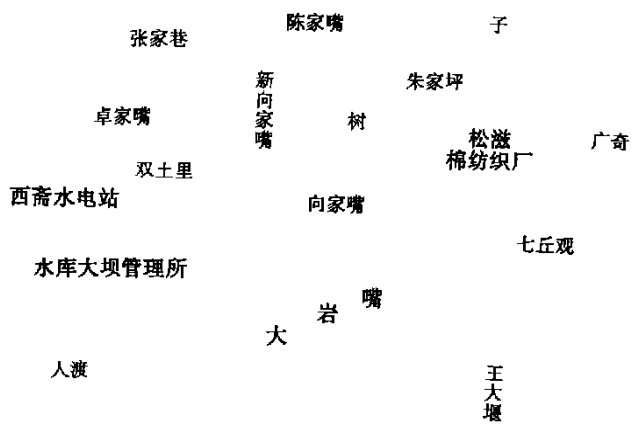

图 4 (b) 一次自动提取结果

Fig. 4(b) The result of (a) after automatic extraction

本文针对目前地图汉字提取的现状, 分析了汉字交互提取和自动提取两种手段的优点和缺点, 从工程应用的角度出发, 应用人机工程学的基本原理, 以主动知识库为核心, 提出了地形图智能汉字提取方法。该方法结合了交互提取及自动提取的优点, 克服了自动提取对复杂对象不适应性以及交互提取的反复性, 实现两者的无缝连接, 大大降低了自动提取的研究难度, 提高了整体效率, 取得了很好的应用效果。

仅仅考虑汉字注记时, 知识库相对来讲比较简单, 容易实现。但是, 地形图是复杂的, 不仅仅是汉

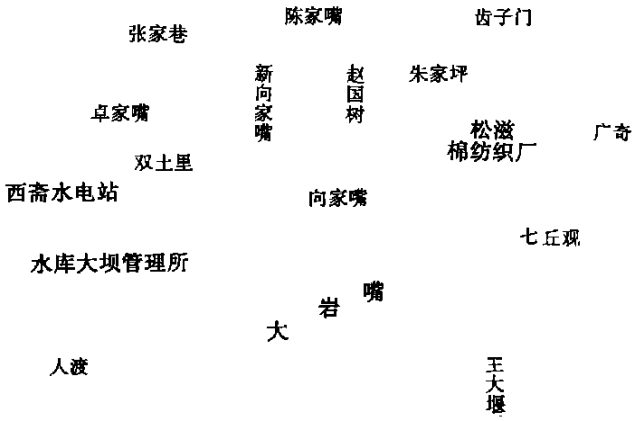

图 4 (c) 提取结果

Fig. 4 (c) The final result of (a)

字, 所有图符之间都存在一定的依赖关系, 如何建立和维护一个完善的主动知识库系统, 结合交互和自动矢量化的优点, 以实现智能、高效的地形图矢量化, 是值得进一步研究的课题。

本文提出的汉字智能提取方法, 是面向用户、面向工程应用的, 力求为用户提供一个快速可靠的汉字提取系统。它避免了交互汉字提取和自动汉字提取谁更有效的争论, 降低了自动汉字提取的研究难度, 大大提高了整体的汉字提取效率, 对地形图/ 工程图的字符提取以及矢量化研究有重要的意义。

参考文献:

[1] 邵子缨, 朱淼良. 一种图纸文字过滤器[J]. 计算机辅助设计与图形学学报, 1998, 10(2): 124- 130.

[2] 李伟青, 彭群生. 一种新的字符提取和组合算法[J]. 工程图学学报, 1997, (2- 3) : 38- 45.

[3] 杜建强等. 工程图纸上的字符提取和识别系统[J]. 计算机工程, 1995, 21(1): 62- 65.

[4] 徐战武, 刘肖琳. 基于结构分析的地形图汉字自动提取[J]. 计算技术与自动化, 1999(11).

[5] 徐战武, 刘肖琳. 一种地形图粘连汉字提取算法[J]. 中文信息学报, 2000, 14(2): 43- 48.

[6] He Xingui. Event Algebra and Active Knowledge Base System[J]. Chinese Joural of Advanced Software Research, 1994, 1(1): 56- 65.

[7] 何新贵. 模糊知识处理的理论与技术(第 2 版)[M]. 北京: 国防工业出版社, 1998.