第3卷 第12期
1998年12月

中国图象图形学报
Journal of Image and Graphics

Vol. 3, No. 12
Dec. 1998

中文地图文字图象理解的研究

夏波涌 刘政凯
(中国科技大学信息处理中心,合肥 230027)

摘 要 地图的识别和理解是一项复杂而艰巨的工作。本文主要针对地图的字符要素,提出了分割合并算法,实现对字符要素的特征抽取,进而实现对字符要素的分割和提取。文章最后给出了分裂合并算法的字符分割提取的实验结果。本算法也可用于其它类型图的字符分割和提取。

关键词 图象理解,合并,分割

1 引 言

在当前的信息社会中,仍有大量的地图信息或类地图信息以纸质的形式保存、管理和使用着。而将这些纸质地图的信息进行计算机管理、存储和使用,并建立相应的数据库,是必然的发展趋势,具有重要的社会现实意义[1]。早期实现这一过程主要采用数字化仪将地图信息采集到计算机,同时实现矢量化处理。显然这种方法缺陷较大,难以满足实际应用的需求。目前主要采用扫描仪将纸质图信息转化为数字图象,然而,这种数字图象的数据量非常大,不利于管理、交流和使用。为了满足实际应用的需求,需要对地图的各种要素进行提取分割以及识别和理解。

地图的主要要素总体可分为三大类:(1)文字要素;(2)线条状要素;(3)块状要素等[2]。本文主要考虑到中文地图的文字提取有别于其它文字地图的文字提取,进而对中文地图的文字要素的提取分割作有关讨论。

目前,有些文献也提出了文字提取分割的有关算法,但是这些算法主要针对文档、专用工程图纸等,而且涉猎的文字主要是英文字符的提取和分割,而对中文地图的文字提取和分割很难适用。这些算法在一定程度上依赖于文档和图形本身质量以及扫描质量,从而中文地图的文字的提取分割的研究就显得独特和重要。国内有些文献虽然涉及地图要素的提取分割和识别理解,但具体到文字的提取分割也较少见到[3]。本文试图在此方面作些有益的探讨和研究。

2 基本原理和方法

彩色地图图象具有比黑白图象更丰富的信息。彩色地图直接转化为黑白二值图象势必会丢失许多有益的信息,造成对下一步的处理工作的被动[4]。但是,彩色地图虽然含有较黑白图象丰富的信息,但就目前的彩色地图的颜色特征空间的聚类分析的研究亦非简单而准确的。如采用彩色地图进行颜色特征空间分类,并对每类颜色特征空间进行文字的提取和分割,这样势必由于颜色特征空间的聚类分析结果的准确性造成最终文字提取分割的困难。再者,很多地图并不具备颜色特性,如采用彩色地图的颜色特征空间来提取和分割文字,显然是不恰当的。另一方面,颜色特征空间的聚类分析会花费一定的时间开销。基于上述考虑,我们采用灰度图象作为原始处理图象。

图1给出了地图文字提取分割的结构流程图。

与西文字符相比较,方块形的中文文字显得较为复杂,对西文文字的提取切割方法并不一定适应

收稿日期:1998-01-24

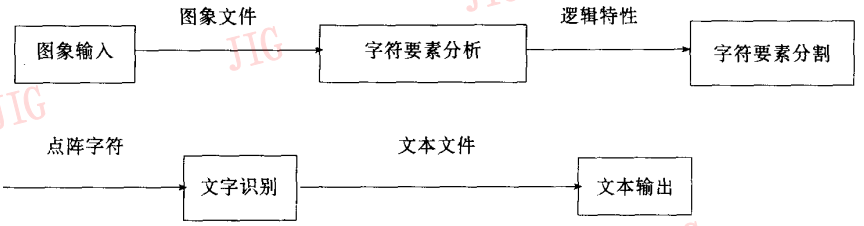

图1 一般的地图文字识别处理流程图

中文文字的提取分割[5]。而传统的中文文档的OCR识别系统中的文字的行字切割方式主要采用投影方法，而比此版面复杂得多的地图文字的提取分割，如采用投影方法是显然行不通的。然而，中文文字虽然较其它文字如英文的模式和类别复杂得多，但是，它同样有自身独特的特征：

(1) 外轮廓呈矩形；

(2) 长宽比近似等于1；

(3) 大多数文字的笔画是连通的，如中、文；

(4) 具有一定的空间结构特性：上下结构、左右结构、嵌套结构等，如益、传、国；本文就是基于上述特征，对地图图象进行版面分析和特征抽取，进而对文字进行提取和分割。

3 文字提取分割算法

3.1 灰度图象的初步二值化及噪声消除

首先，针对原始灰度图象进行二值化处理。这里二值化的阈值的选取较为关键。若阈值取得过大，则保留的信息过多，其中许多杂点无用信息，造成对以后处理的干扰。若阈值取得过小，则丢失的信息过多，其中许多文字信息产生续断或丢失，造成最终的文字提取分割的文字丢失。地图灰度图象中的线条要素和文字要素的灰度值较周围的象素的灰度值小，这里阈值的选取我们按以下原则来取：

(1) 计算原灰度图象的均方差和灰度均值；

(2) 取均值与均方差之差作为阈值；

(3) 若该象素的灰度值小于阈值，则保留该象素；

(4) 用四邻域法对孤立点象素进行消除；

从而，原灰度图象中的文字信息得到保留，但同时也保留了其它的非文字信息。

3.2 标号和滤波

下面，对3.1处理过的图象进行标号和滤波运算，初步提取文字信息。

(1) 对每个连通区域进行标号 i；

(2) 计算每个标号区域的外接矩形框；

(3) 计算每个矩形框的长宽比 $R(i)$ 和面积 $S(i)$；

(4) 面积和长宽比滤波：如果 $|R(i)-1|\leqslant\Delta R$ 和 $|S(i)-S|\leqslant\Delta S$，保留输出。否则，转3.3处理。其中 ΔR、ΔS 和 S 为预先选取的阈值。

3.3 分裂

(1) 如该标号部件的外接矩形的 $|S(i)-S|\geqslant\Delta S$，则对其分裂运算，否则，进行合并运算；

(2) 计算该部件的外接矩形内原图象的灰度均方差和均值；

(3) 以此灰度均值和均方差之差为阈值；

(4) 如该部件的象素的灰度值小于阈值，则保留；

(5) 重新标号和滤波；

(6) 检查是否有满足条件(1)，如有，继续分裂；如无，分裂结束；

这里，分裂的主要过程也就是进行二值化的阈值的动态选取。在原图象中，同一文字的灰度值近似相等，且与周围的其它要素的灰度值相差较大，这就是我们能进行动态二值阈值选取的实际原因，进而我们能达到分裂的目的。

3.4 合并

经过上述步骤处理的图象，必然有许多未被提取的文字部件，造成这种状况的主要原因是汉字结构的复杂性，如上下结构的“会”字，左右结构的“北”字，如何将它们提取出来，这里我们采用合并算法：

(1) 对任一剩余标号部件 i，与其余的某任一标号部件 j 合并(i 不等于 j)；

(2) 计算其合并的外接矩形框的长宽比 $R(i,j)$ 和面积 $S(i,j)$；

(3) 对任何部件 j，如果 $|R(i,j)-1|\leqslant\Delta R$ 和 $|S(i,j)-S|\leqslant\Delta S$，计算 $\Delta RS(i,j)=|R(i,j)-1|+|S(i,j)-S|/S$；

(4) 对任何部件 j，选取 $\Delta RS(i,j)$ 为最小的部

件 j，则 i 部件与部件 j 最终合并；

(5) 将合并的部件保留输出；

(6) 如无可合并的部件，算法结束。

4 实验结果和分析

图 2 是实验用的原始地图图象，这里采用 600dpi 分辨率，256 级灰度图象，图象大小为 450×450 个象素。图 3 为本算法处理后的结果图象。

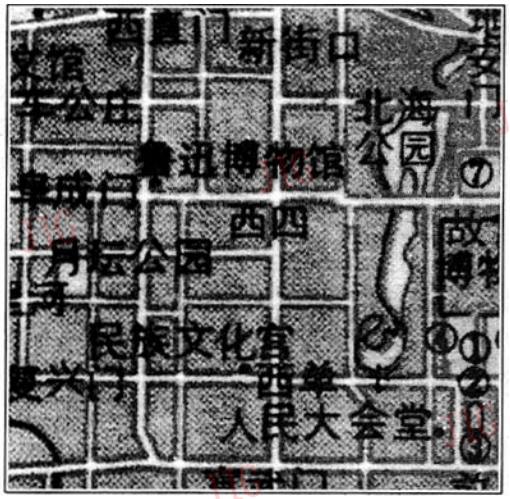

图 2　原始图象

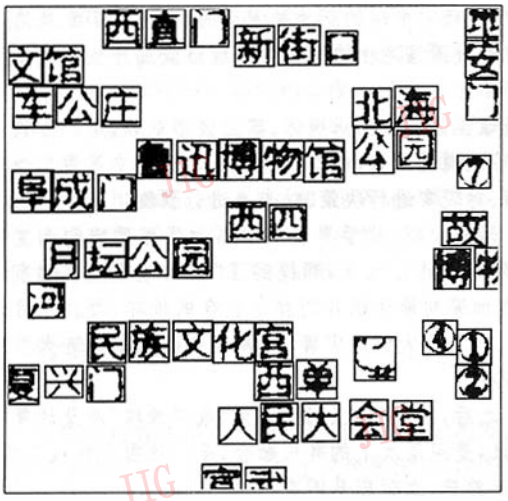

图 3　处理结果图象

由处理结果图象可知：对全部连通及整体形态较好的文字较好地分割和提取出来，如“新”“大”等字；对具有一定结构的文字如“门”“会”“北”等字，经过合并运算，也能较好地将之提取和分割出来；对一些粘连部件的文字如图 2 中的“海”“公”等字，经过分裂运算后，基本上能提取和分割出来，但是，提取后的文字图象存在一定的干扰如图 3 中的“公”“海”等字。这主要是由于原图象中的文字和其它标识之间存在较强的重叠和干扰。另外，由于采用外接矩形长宽比作为合并参数，处理结果也存在了一些非文字图象，如图 3 中的“  ”。

总之，由处理结果图象图 3 可以看出：本算法基本上将原图中的文字提取出来。

5 讨　论

本文提出的算法，从实验结果来看，获得较为满意的结果，但是这种分割提取算法仍有待改进。

(1) 最初图象的二值化的阈值的选取很大程度决定了最终的处理结果。故此，作者认为：如何正确且稳定地开发出分割和提取算法，应尽可能减少对阈值选取的依赖。

(2) 当然，文字分割的最终目标是为了识别。本文的重点在于分割和提取，如何将分割提取和识别结合起来，有待进一步研究和讨论。

(3)针对不同字型、字体的分割，我们将在以后的工作中继续探讨。

参 考 文 献

1　蒋庆权译. 地图识别输入动向. 现代电子工程，1989，1，69～75.

2　周源华，权淑媛，刘惠娟. 地图的计算机识别. 上海交通大学学报，1993，6，26～32.

3　Kasturi R，Alemang J. Information extraction from images paper-based maps. IEEE Trans. on Software Engineering，1988，14(5)：671～675.

4　朱文忠. 基于颜色特征的地图要素的分割和识别. 模式识别与人工智能，1996，9(6)：194～200.

5　Taylor S L，et al. An Intelligent Document Understanding System. Proc 2 ICDAR，1993，107～110.

夏波涌 中国科技大学博士研究生。主要从事图象理解、地理信息系统、模式识别及多媒体通信等方面的研究工作。

刘政凯 1964年毕业于中国科学技术大学，现为中国科学技术大学教授，博士生导师。主要从事遥感图象处理，人工神经网络及模式识别方面的研究，已发表论文60余篇，出版专著6本。

A Study of Image Understanding of Character of Chinese Map

Xia Boyong, Liu Zhengkai

(*Information Processing Center, USTC, HeFei* 230027)

Abstract It is a complicated and formidable task to recognize and understand the map. The paper gives a algorithm of split and combination. It realizes the feature's extraction of character element of map and character's segment and extraction. Finally, we give the experiment result of the algorithm of split and combination. The algorithm is applicable of the other type map character's segment and extraction.

Keywords Image understanding, Combination, Segment

“数字地球”已经在中国转动

人们说“数字地球”是知识经济的基石，10月24日教育部科技司在北京大学主持召开了“数字地球”高级研讨会。来自国家有关部委的领导及领域的专家学者就“数字地球”技术、“数字地球”与信息高速公路的关系、“数字地球”对我国社会经济发展的影响以及与知识经济的关系等一系列重大课题，进行了深入研讨。专家们认为，“数字地球”是全球经济一体化、高技术化、信息化过程中出现的新的技术动向，也是我国发展信息产业新的切入点和机遇。“数字地球”技术的运用和推广，将对我国信息化建设和国民经济发展产生重大推动作用。

“数字地球”是指信息化的地球，或地球的信息化，是地球实体的虚拟对照体。建立数字地球，是因为我们迫切需要利用有关地球的各种信息。利用数字地球，我们可以通过模拟，在人和计算机之间建立更为自然的界面，真正实现人脑和电脑的对话，并以次对过去进行反演、对现实进行决策、对未来进行预测。

“数字地球”首先由美国于今年初提出，立刻引起许多国家政府、科学界及产业界的强烈反响和高度重视。江泽民主席于今年6月在接见两院部分院士及军队外事工作会议上，都提到了“数字地球”。“数字地球”之所以得到如此广泛重视，不仅因为它可以帮助我们更加深刻地认识我们赖以生存的地球，而且还可以帮助我们解决许多最迫切的社会问题，如保护资源和环境、预测重大自然灾害、发展农业生产和促进大型产业的增长，这和“信息高速公路”同等重要，是知识经济的基石。

与会者在听取了有关专家关于“数字地球”的技术报告之后，经过热烈讨论，认为“数字地球”涉及计算机科学、信息科学、地球科学、系统科学和社会科学等众多领域，是深层次上的有机融合，是一项重大科技工程，是前瞻性与实用性相结合的项目，有重大的科学意义和实用前景，立项应是国家行为。

现在，我们正身处人类历史发展的关键时刻，我们将用科学技术与全球信息，克服自古以来的局限性，重新认识地球。用人类独特的能力和卓越的智力，利用地球丰富的信息资源达到人与自然、人与人之间的和谐相处。