基于边界码的非 Manhattan 格式下的文本图象的自动区域分割方法

高 翔 张 利 吴国威

(清华大学电子工程系,北京 100084)

摘 要 随着网络和通信技术的发展,电子图书馆逐渐发展为信息检索和资料提供的重要途径。非 Manhattan 格式的排版方式因具有形式自由、富于变化的特点而得到越来越广的应用,但是对于非 Manhattan 格式的文本图象的自动区域分割的研究还只是停留在实验的水平,迫切需求一种快速实用的自动区域分割方法提高电子图书馆的文献录入速度。此文提出的基于边界码的文本图象的区域分割方法在实际的应用中达到了令人满意的效果。

关键词 文本图象,自动区域分割,Manhattan 格式

1 引 言

1.1 提取文本区域部分信息的重要性

当今已进入信息社会,信息的需求量与日俱增。报纸、杂志和图书等以纸张为信息媒介的刊物已不能完全满足人们对传递信息和获取信息的需要,电子刊物就应运而生。怎样快速准确地生产和获取电子文件已成为当前研究的一个重要课题。

文本图象是电子文件的重要形式之一,它是将印刷文本文件(Printed documents)通过扫描仪等输入设备转换而成的二维数字图象。文本图象的处理过程是一个内容广泛的研究课题,包含从图象的获取到结果的输出等一系列的技术。在整个的分析处理过程中有二种分析过程是必不可少的:一种是光学符号识别(Optical Character Recognition OCR);另一种则是版面分析。所谓的版面分析就是对组成文本图象的各个元素(标题、图象等)之间的位置结构关系进行分析。

文本图象的自动区域分割是文本图象版面分析的重要课题之一,也是 OCR 识别的重要的前端处理步骤。目前,OCR 技术已经有了很大的发展,机器对印刷体字符的识别率已经达到了高于 96%的水平。而文本图象的自动区域分割及版面分析的水平却离广泛的实用化阶段还有一定的距离。这主要是因为:在实际的印刷过程中,为了吸引读者的注意或为了更好的表达自己的意图,往往有大标题,有关作者、日期的标注,在大块的正文中穿插小标题和一定量的图片、表格和公式等。这使得文本图象中的版面结构千变万化。由于目前自动区域分割算法的适用范围小且分析结果的正确率和算法分析的速度随版面结构复杂程度的提高而降低等问题,在正式的电子图书馆的建设过程中,文本图象的版面分析及文献检索内容的提取等工作都是由人工完成的。这就形成了抑制文献录入速度的一个瓶颈。为了有效地解决该问题,各国学者提出了多种算法。

1.2 区域分割和版面分析的技术的发展现状

从 80 年代末开始,文本图象分析(Document Image Analysis DIA)的技术迅速发展。针对在文本图象的分析处理中所遇到的问题,提出了很多种可实际应用的专用算法:如文本图象的自动倾斜校正算法,图象的噪声消除算法,游程平滑算法(Run-Length Smoothing Algorithm RLSA),在周期性背景下的字符提取算法[1],灰度图象中的字符提取算法等等。该技术已从单一的区域分割发展到与计算机视觉、模式识别等理论相结合的版面结构的区域分割技术。总的来说,文本图象的自动区域分割技术主要有两类:自顶而下(Top-down)算法和自底而

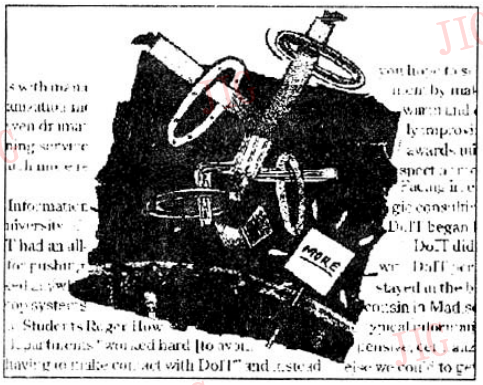
上(Bottom-up)算法。

自顶而下的区域分割算法采用了从整体到局部的设计思想。其中最为有效的算法有:X-Y 树(X-Y tree)的文本图象的层次分割算法,基于黑象素游程长度的等块基本结构特征的文本区域和图象区域的划分方法,参考黑白象素对游程(Black-White Pair Run Length)和黑白黑联合游程(Black-White-Black Combination Run Length)的判别文本区域和图象区域的方法。这些算法的主要应用对象是符合 Manhattan 格式[2]的文本图象(可以用一组水平和竖直的直线段完成区域划分的文本图象)。它们的共同特点是对版面结构简单或版面结构已可以用先验知识表示的文本图象进行文本图象的自动区域分割时,有较快的运算速度,通过有回溯功能的树搜索算法,还可提高区域分割的成功率。但是随着文本图象版面结构的复杂程度的提高,算法的运算时间和数据结构的存储空间会成指数增长。

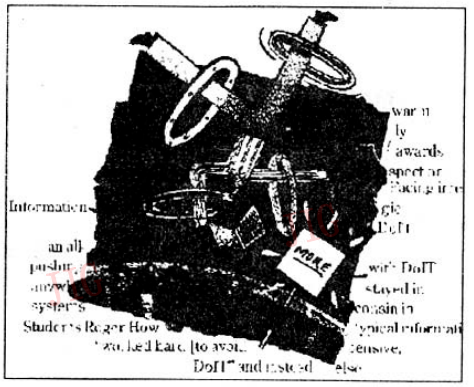
自底而上的区域分割算法则是采用了从局部到整体的设计思想,它从文本图象的象素点分析入手,将象素点逐渐聚集为一个个的连通区域,根据这些连通区域的几何和结构特征进行文本区域和图象区域的属性判断,达到完成文本图象自动区域分割的目的。这方面的典型算法有游程平滑算法,连通域的边缘检测算法等子算法。其核心算法都离不开使用变换域的谱分析手段[3,4]进行连通子区域的合并。虽然变换域的谱分析技术对所处理的文本图象是否符合 Manhattan 格式没有特别的要求,但是考虑到从空域到变换域的算法的运算量随被变换的点数的增长而急剧增长的特点,使游程平滑法成为必不可少的算法之一。不过在非 Manhattan 格式的文本图象中,用游程平滑算法很容易改变文本图象的排版结构(即会在游程平滑时将文本区域和图象区域连通在一起),造成文本区域和图象区域的误划分。

图 1 中的(a)、(b)就是依据文献中所述的算法进行处理后的结果,从图 1 中可以看出,存在少量的文本区域被误识为图象区域,同真正的图象区域混杂在一起。如果不能在进行文本区域的 OCR 识别之前改正文本区域的误划分,则必然会导致最后的印刷文本文件的录入结果出现错误,限制了文本图象的自动区域分割的适用范围。

本文的目的就是提出一种可在非 Manhattan 格式下进行快速的文本图象自动区域分割的方法。



(a)原始文本图象及经过自顶而下的算法处理后的图象区域



(b)经过自底而上的算法处理后的图象区域

图 1 非 Manhattan 格式的文本图象的自动区域分割结果

Fig. 1 document image autozoning result of non-manhattan format

(a) the Original Image and the Graphic zone processed by top-down method

(b) the Graphic zone processed by bottom-up method

2 解决问题的方法

在本文提出的算法中,文本区域的提取和分割只是基于在文本图象中的连通元素的几何特征、结构特征和相互之间的位置关系。其次,本算法的适用条件是:(1)文本区域的元素要与图象区域的元素相分离。即当以连通域的角度观察待分割的文本图象

的时候,文本区域和图象区域是两个相互独立的文本图象区域。(2)在图象区域内部可以找出有决定作用的中心图象区域。本算法的指导思想就是先提取误识区域中的中心图象区域,然后再根据具体情况对中心图象区域周围的文本区域进行提取。

2.1　名词定义

为了便于对问题进行讨论,我们给出两个定义:

"峰"和"谷"——在文本图象不规则的外边缘上,凸出的部分称为"峰",凹进的部分称为"谷","峰"和"谷"交替出现。

文本区域的纹理特征——所谓文本区域的纹理特征,就是指文本区域都是由字母组成的单词构成的,这些单词组成了在水平方向上的文本行。文本行又在竖直方向上排列起来而形成文本段落。文本行与文本行之间也存在着一定长度和一定高度的行间隔。从整体上看,就如同一行行的黑色文本与白色的行间间隔构成了有规则的黑白相间的水平纹理。

2.2　具体算法

本文中讨论的非 Manhattan 格式部分的文本图象自动图文区域分割算法(图 2)主要包含动态游程平滑,基于连通域的文本图象的基本连通区域的检测,文本区域和图象区域属性的预分类,图象区域的再次分割,基于区域属性的区域合并等几个步骤。

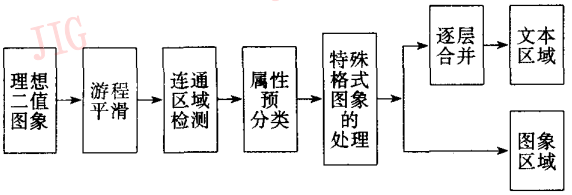


图 2　文本图象自动区域分割的算法流程图

Fig. 2　The flowchart of document image autozoning

在动态游程平滑阶段,用动态统计的文本区域的行高和行间距,字符的宽度和字符间的间距分别作为竖直方向和水平方向的平滑门限选取的参考,使得在一定范围内的达到尽可能连通文本区域中的黑色象素点,减少将文本区域和图象区域的元素连通在一起的概率。

在基于连通域的文本图象的基本连通区域的检测过程中,采用 8—邻域的连通区域检测算法进行操作,减少最终结果中的连通域的数目。为了简化连通域的数据之间的相互关系,只对连通区域的外边缘进行检测,忽略各个连通区域内部的边缘。

当进行文本区域和图象区域的预分类的时候,选用连通区域的各种几何特征(如区域的高度、宽度、体态比、周长和区域内部的黑色象素点的密度等)作为特征向量,组成特征空间。通过用大量已知属性的样本区域对连通域分类器进行训练和监督,可以得到使用效果良好的划分门限,达到以连通域为单位的分类目的。

当文本区域和图象区域在游程平滑的运算中连通在一起的时候(这通常发生在非 Manhattan 格式部分的文本图象区域),由于区域中图象部分的内容是区域的主体,对所提取的特征的贡献远大于文本部分的贡献,故这种混有文本区域和图象区域内容的连通域会被连通域分类器判别为图象区域。

本文提出的算法优于其他算法的最核心的方法之一就是图象区域的再分类方法。它主要包括 3 个步骤:

2.2.1　判断图象区域中是否混有文本区域的内容

在非 Manhattan 格式的文本图象中的图象区域的内容是千变万化的,要判断是否在图象区域中混有文本区域的内容,首先就要了解文本区域的内容与图象区域的内容在几何特征上的最大的差别是什么。通过对近百幅非 Manhattan 格式的文本图象进行观察与比较,作者认为两者在几何特征上最大的差别就在于文本区域的内容在水平方向上含有强烈的纹理信息。不论在误识区域中的中心图象区域的组成结构如何变化,都无法掩盖文本区域所独有的文本纹理特征。

这样,只需找到一种可以很好反映出文本区域纹理特征的数学模型,就可以比较容易的判断出在一个被连通域分类器判别为图象区域的内部是否存在着文本区域围绕在真正的图象区域的周围。结合模糊数学理论,通过对采用各种数学模型时的判断时间和判断效果进行比较,决定采用经过动态游程平滑技术处理过的文本图象的连通区域的外边缘轮廓作为图象区域的纹理特征的反映。这是因为当连通区域内存在图象区域的内容周围有文本区域的内容的时候,必然会导致文本区域内容的外边缘取代了在相应的位置上的图象区域内容的外边缘。根据连通区域检测中得到的图象区域的外边缘的数据,就可以寻找图象区域中的水平方向上的"峰"的信息。当某一图象区域的外边缘上存在着这样的"峰"的时候,就表示在这个图象区域中可能有文本区域的内容。

2.2.2　寻找中心图象的边界

在找到可能含有文本区域内容的图象区域以后，接下来的任务就是将图象区域中可能包含的文本区域的内容提取出来。在具体文本区域内容的诸多提取方案中，简洁而有效的途径是首先找出位于连通区域内的中心图象区域的位置。当中心图象区域被提取出后，再对其四周的文本图象部分进行区域分割。提取中心图象区域的最直接的方法是图象的边缘跟踪算法。为了正确的使用边缘跟踪算法，最重要的问题就是找到一个真正的隶属于中心图象区域的外边缘上的象素点。如果以一个其他附属图象(非中心主图象)边缘上的点作为边缘跟踪算法的起点的时候，边缘跟踪的结果将只是附属图象，仍无法顺利的避开中心主图象而进行文本区域的提取工作。

注意到，已找到的连通区域的外边缘上的“峰”只是区域内含有文本区域内容的一个必要而不充分的条件。并不能将是“峰”的位置就当作文本区域处理。处在“谷”上的象素点也不一定是就是中心图象区域外边缘上的象素点。真正的图象区域的外边缘上同样会出现“峰”而纯文本区域的外边缘上也会有“谷”存在。这在当中心图象区域是由多个小块的图象所组成的时候这种现象尤为明显。由概率统计及模糊数学的观点可以导出：在连通区域外边缘的“峰”上象素点隶属于真正图象区域边缘的置信度与该象素点隶属于文本区域边缘的置信度近似相同。而连通区域外边缘的“谷”上的象素点则不同，其对真正图象区域边缘的隶属程度远高于对文本区域边缘的隶属程度。通过对大量的非 Manhattan 格式的文本图象部分进行统计，发现由文本区域内容形成的“峰”的宽度和与整个图象区域高度的比例总是很小。可见，当“谷”的长度越大时，它代表的象素点是中心图象区域的边缘的可能性就越大。可以用宽度最大的“谷”上的一象素点作为边缘跟踪算法的起点。由于在图象区域中由多个间隔很近，彼此并不连通的子图象组成中心图象的情况是相当普遍。所以再简单的运用 4－邻域或 8－邻域的算法进行边缘检测并不总是能得到正确的中心的图象区域组的外边缘。结合文本图象中的文法，只要得到中心图象区域的确切的左右边界，就可以方便的进行文本区域的提取工作。这样做不仅减少了算法的复杂度，还增加了算法的稳定性和适用范围，可以一次成功的提取中心图象组整体的边界信息。

2.2.3 文本区域的提取和与图象区域的区分

找到了连通区域中的中心图象区域，就可以对中心图象区域周围的区域进行进一步的区域分割。对于任一幅文本图象而言，文本区域和图象区域的内容都可以出现在中心图象区域的周围，这就给进一步提取中心图象区域周围的文本区域的内容带来了困难。

为了降低剩下的文本图象区域分割的难度，可采取分而治之的办法。先假设在中心图象区域周围的全部是文本区域的内容，将它们分别的提取出来，形成一个个的子区域。然后依据文本区域与图象区域在几何特征和结构特征上的差异逐个的对提取出来的子区域进行属性辨别，过滤出图象区域，并将它们与中心图象区域进行合并。从而得出正确的文本区域和图象区域。下面就上述的两个子步骤进行讨论。

在提取中心图象区域周围的子区域的时候，由于事先做了全部是文本区域内容的假设，使得在这部分的工作得到了简化。依据文本区域具有的独特的纹理特征，对连通区域的剩余部分采用一次反复投影的方法就可以达到提取子区域的目的。

然而在文本区域和图象区域预分类中用的几何特征的特征空间算法却已不适合这里的子区域属性辨别工作。主要原因是从这种误识区域中分割出的文本区域已经不具有经过 RLSA 处理所得到文本区域的常有的几何特征，如区域的长度应有几个单词的长度等等。必须寻找新的方法来辨别文本区域和图象区域。经多次实验，使用下面三种方法进行多数表决，就可以正确的对区域的属性进行辨别。

(1)文本区域纹理方面的检测。如果子区域是文本区域，那么它应该具有文本区域相应的纹理。通过竖直方向的投影运算，应可以检测到在单词与单词之间，字母与字母之间的有规律的间隔。

(2) 文本图象文法的检测。一篇文章中，孤立的有限几个字母并不能准确地表达作者的意图。在一个小的文本区域的相应位置上(如水平位置)总会有其他的文本区域和它对应。因此可以用那些已经经过判断的文本区域的位置信息作为一个依据来检测未知的区域属性。

(3)文本区域结构方面的检测。图象中的文本区域和图象区域在结构特征上的最主要的差别是内部形状结构的差别。构成文本区域的主要内容是具有一定形状的字符，图象区域中的象素分布则具有一定的连续性。我们以区域的线结构特征作为区域的结构特征对未知属性的区域进行判断，得到了不错的效果。区域中的线结构特征是利用形态学中的击

中或击不中变换（Hit-or-miss Transform）来获取未知区域中的线条信息。

区域有效部分的集合定义为：$g(x,y)=\bigcup_{n}(f(x,y)\otimes B)$。其中 $f(x,y)$ 是区域中象素点的集合，B 是结构特征算子。

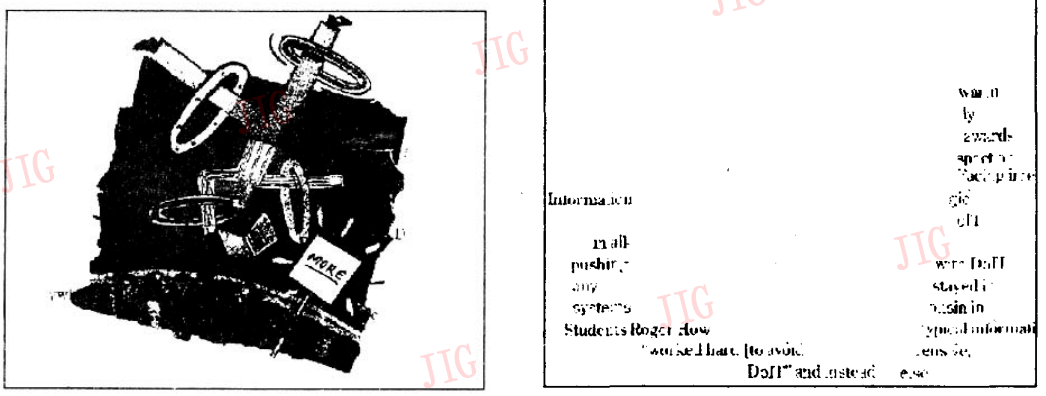
区域的线结构特征定义为：$LSV=\frac{S_{g(x,y)}}{S_{f(x,y)}}$。其中 $S_{g(x,y)}$ 和 $S_{f(x,y)}$ 分别表示相应集合的面积。

虽然此种分类方法的成功率和正确率都明显高于基于区域的几何特征的分类方法，但是在时间的花费上也是明显的增多的。

将文本图象以小块区域的形式进行了分割，并对每个区域的属性都做了区域属性的辨别后，再按照文本图象的文法进行相同属性的区域合并就可以得到正确的文本图象的区域数据，实现文本图象的自动区域分割。

3　实验结果

经过使用上述的文本图象自动区域分割方法对近百幅来源于西文杂志的各种类别的文本图象进行区域分割的实验，可以得出结论：分割算法可有效的将相距较近的文本区域和图象区域分割开。图 3 显示了本算法对图 1 的自动区域分割效果。



(a)图象区域的提取结果　(b)文本区域的提取结果

图 3　经过基于边界码的文本图象的自动区域分割结果

Fig. 3　The result of document image autozoning processed by boundary code method

(a) The result of Graphic zone

(b) The result of Text zone

4　结　论

本文提出的基于边界码的非 Manhattan 格式的区域分割算法主要是根据文本图象的纹理特征、文本图象的文法及文本图象区域的几何和结构特征来实现文本区域和图象区域的属性辨别。从对大量的文本图象进行区域分割的结果可以看出，本算法的效果良好，并注意到了区域分割的正确率和分割时间之间的平衡。目前已取得了很好的实际应用效果。

参 考 文 献

1　Liang S, Ahmadi M. A morphological approach to text string extraction from regular periodic overlaping text/background images CVGIP: Graphical Models and Image Processing, 1994,56(5): 402～413.

2　Baird H S, Jones S E, Fortune S J. Image segmentation by shape-detected covers. Proc. 10th Int. Conf. Pattern Recognition, 1990, 820～825.

3　Frank Hönes, Jürgen Lichter. Layout extraction of mixed mode documents. Machine Vision and Applications, 1994.

4　Lawrence O'gorman. The Document Spectrumfor Page Layout Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(11).

高翔，1996年获清华大学电子工程系学士学位，现在清华大学电子工程系攻读硕士学位。主要进行文本图象的自动区域分割及版面理解方面的工作。

Boundary-Based Non-Manhattan Document Image Autozoning

Gao Xiang, Zhang Li, Wu Guowei

(*Electronic Engineering, Tsinghua Univ, Beijing* 100084)

Abstract Today, more and more printed documents use non-Manhattan format. The traditional autozoning methods don't adapt to the format properly. A new autozoning method based on boundary code is presented in this paper and implemented for document image of non-Manhattan format.

Keywords Document image, Autozoning, Manhattan format

(上接873页)

A Kind of Fractals Based on Part of Real Number

Huang Tiejun, Liu Jian

(*Image Recognition and Artifical Intelligence Institute,*
HuangZhong University of Science and Technology, Wuhan 430074)

Abstract By extending the concept congruence from integer to real number field, this paper defines a concept named 'part' of real number. A real number can be represented as an infinite sequence, a finite sequence of it is called 'part'. If two real numbers have the same part, then they are called isopart. By Studying the part of a real value function, we have the conclusion as following: the set which includes all the isopart points is normally a binary fractals, The image of the new function generated by taking the part of the given function is multi-value fractals. The infinity of real number leads to the complexity of this kind of fractals, and the given function leads to the regularity. The method does not need iterative operations which are essential for traditional fractal generation methods. In addition, it makes the connection between numbers and fractals.

Keywords Fractals, Number, Part, Isopart