基于模式链分析的文本页面图像的分割与分类

李艳玲 王加俊
(苏州大学电子信息学院,苏州 215021)

摘 要 为了能对复杂版式的文本图像(如包含镶嵌在文字中的形状不规则的图片区)的页面进行图文分割与分类,提出了一种新的基于模式链分析的文本页面分割与分类算法。该算法首先使用外接矩形框出图像中的所有黑像素,并且存入矩形框链表中,再组合所有相邻的矩形进而形成模式,最后依据各模式的统计特征分类,输出文字区和图片区两类图像。另外,对大图片模式周围个别不确定的模式,本文采用了上下文分类的算法进行再次分类。实验结果表明,该算法不仅运算速度快,而且能够对复杂版式的页面图像进行正确的图文分割和分类。

关键词 矩形框链表 模式链表 模式上下文 页面分割和分类
中图法分类号:TP391.4 O4 **文献标识码**: A **文章编号**: 1006-8961(2005)06-0741-05

Document Page Segmentation and Classification Based on Pattern-list Analysis

LI Yan-ling, WANG Jia-jun
(*School of Electronic and Information Engineering, Soochow University, Suzhou* 215021)

Abstract In this paper, a new algorithm based on pattern-list analysis is introduced for page segmentation and classification of document images with irregular-shaped halftone regions embedded in the text regions. This algorithm is composed of three steps. The first step, all the black pixels are extracted by the bounding-boxes and are stored in a linked rectangle-list. The second step, all connected rectangles are grouped to form patterns and pattern-list. At last, the page images are classified into text regions and halftone regions according to their the statistical features. After above three steps, still uncertain patterns are further classified by the type of contextual patterns. Experimental results show the fastness of the proposed algorithm in segmenting text and halftone regions and its excellent performance for complex document images.

Keywords rectangle-list, pattern-list, pattern context, page segmentation and classification

1 引 言

随着计算机,因特网的普及,网络数据库和电子图书馆的产生,用电子格式存储的文件越来越多。相对于传统的纸张文档而言,电子文档具有存储空间小,方便检索,便于传输,易于更新等优点,因此将纸张文档转换成电子文档具有十分重要的意义。由于文本页面图像版式多种多样,通常不仅含有文字,还可能含有图片、图表等非文字区域。光字符识别系统(OCR)只能识别文字区域,而对于文字、图片同时存在的区域就无法工作,因此分离文字与非文字区域将成为一种迫切的需要。

对于文本页面图像的分割与分类的研究从 20 世纪 80 年代就开始了[1],这些算法大体上可以分成几何分析法[2,3]和纹理分析法[4]两类。纹理分析法主要根据文本图像中的不同区域(文字、图片、图表等)在纹理上存在显著差别的特点,对文本图像进行分割和分类。几何分析法较有代表性的算法是投影轮廓切分算法[5],通过页面的各个分离部分生成外接矩形框,然后对这些矩形框进行 x 轴和 y 轴方向上的投影,再从投影的谷点进行切分,但是这种算

基金项目:国家自然科学基金项目(30300088)
收稿日期:2003-08-26;**改回日期**:2004-12-21

第一作者简介:李艳玲(1978 ~),女。苏州大学通信与信息系统专业硕士研究生,现为内蒙古师范大学讲师。研究方向为数字图像处理。E-mail: liyanling7871397@sohu.com

法对文档的倾斜度要求较高,对于倾斜的文本必须先进行校正,而且文本倾斜的方向只有一致才可以正确分类;另外,它还要求文档中不能存在形状不规则的图片,这些都大大限制了算法的通用性。

Mitchell 等人提出了一种分割与分类方法[6],但是这种方法尽管强调鲁棒性强,能够适合各种版式的页面,但是算法的复杂性以及诸多的人工干预,使得算法的运算时间大大延长,不便于人们使用,而且许多分类规则基本不起作用。

鉴于此,提出一种新的基于模式链分析的分割方法。该方法是对二值化后的图像进行操作。主要使用了数据结构的链表结构,把按照像素存储的图像转换成为一条模式链表,然后再对链表中的各个模式进行分类。通过遍历链表使得分类速度大大加快。

2 模式提取

模式的提取分为两步:第 1 步确定矩形框[6],将图像用一条矩形框链表表示;第 2 步合并相邻的矩形框成为模式,得到模式链表。

2.1 矩形框的确定

众所周知,一个矩形是由上、下、左、右 4 个坐标唯一确定的。这里使用 3×3 像素的模板对图像进行水平和垂直两个方向上的搜索,用矩形框出图像中所有的黑像素。模板的搜索从图像的左上角开始。模板先进行水平搜索,当模板中发现黑像素,则确定矩形框的左上角坐标。模板依据步长 x_{step} 水平向右移动,当模板中不存在黑像素时,开始进行垂直搜索。这时,模板依据步长 y_{step} 垂直移动,直到模板中不存在黑像素时,这个矩形框就被确定下来了。模板中 9 个像素的扫描顺序在水平搜索和垂直搜索过程中是不同的。图 1(a)所示的模板用于水平搜索,模板中像素的扫描顺序是从右向左,依照图中序号的顺序进行;图 1(b)模板用于垂直搜索,模板中像素的扫描顺序是从下到上,也是依照图中序号的顺序进行。

7	4	1
8	5	2
9	6	3

(a)水平搜索时的模板

7	8	9
4	5	6
1	2	3

(b) 垂直搜索时的模板

图 1 模板中像素的扫描顺序

Fig. 1 Order for pixel scanning in a mask

(序号"7"为模板的当前位置)

形成矩形框链表的步骤如下:

(1)首先将二值化后的图像进行标记,白像素用"0"表示,黑像素用"1"表示。将模板的当前位置(每个模板中序号为"7"的位置称作模板的当前位置)放置在图像的左上角,模板中像素的扫描顺序依照图 1(a)所示的顺序进行,设置当前步长 x_{step} 为 3。如果矩形框中没有标记为"1"的黑像素,则模板依据当前步长向右移动。如果在模板中发现标记为"1"的黑像素,则步长的确定就成为关键,其方法如下:

①如果在扫描模板的第 1 列发现黑像素,即在序号为"1","2","3"的位置发现标记为"1"的黑像素,则步长不变;

②如果在第 2 列发现黑像素,即在序号为"4","5","6"的位置发现标记为"1"的黑像素,则步长减 1;

③如果在第 3 列发现黑像素,即在序号为"7","8","9"的位置发现标记为"1"的黑像素,则步长减 2。

(2)创建一个新矩形框,模板的当前位置即为矩形框的左上角坐标。模板继续向右移动,依照步骤 1 的方法继续进行搜索,直到模板中不存在标记为"1"的黑像素为止。此时矩形框的右坐标为模板当前位置的列位置减 1。这时一个矩形框的 4 个坐标已经确定了 3 个。

(3)使用图 1(b)所示的模板在垂直方向进行搜索。模板先回到矩形框的左上角。搜索之前,需要检测这个已经确定了左右边界的矩形框中,在 3 行内存在黑像素的最下边一行,以此来确定步长 y_{step},之后,模板的垂直搜索方法和水平搜索类似。模板顺着矩形框的左边界向下搜索,一直进行到模板中不存在标记为"1"的黑像素为止,则矩形框的下坐标为模板当前位置的行位置减 1。这时一个矩形框就被确定下来了。同时将该矩形框中所有标记为"1"的像素重新标记为"2",以区别该像素是否已经被矩形框框定。

(4)下一个矩形框的确定从上一个矩形框的右边界开始,继续先向右,再向下进行搜索。只要发现标记为"1"的黑像素就创建新的矩形框。模板向右搜索时,每次只扫描图像的 3 行,这个过程一直进行到图像的右边界为止。然后再开始新一轮的向右扫描。

使用该方法确定矩形框的优点是:不需要扫描图像中的每个黑像素,即可确定图像中的所有黑色区域。其原因举例说明如下:从算法的描述中可以看到,在水平搜索过程中,如图 1(a)所示,如果模板中的 9 个像素全部是黑像素,那么首先在序号为"1"的位置发现了黑像素,这时创建矩形框,模板以

步长 3 向右移动,其余 8 个像素没有被扫描就进入了矩形框。由此可见,只要找到模板中最右边的黑像素,其左边的黑像素未经扫描就进入了矩形框而被提取出来,从而大大地减少了计算量。垂直搜索也具有这样的特点。搜索完毕后,整个图像就用一条框有黑像素的矩形框链表表示了,而且图像中的每个黑像素至少包含在一个矩形框中。

2.2 模式的形成

模式由矩形框形成。这时不必再对原图像进行操作,因为矩形框链表已经包含了图像的所有信息。所谓模式就是由一组相邻或相交的矩形框组成的尺寸更大的矩形,同时每个模式中还含有形成该模式的矩形框的信息。文献[7]中给出了一种合并算法。根据该算法,当两个矩形框的边界相交或者相邻,即边界距离相差至多一个像素时,就合并它们。图 2 给出了两个矩形 a 和 b 位置关系的示意图,其中

$$a_{width} = a_{right} - a_{left} + 1 \quad (1)$$

$$a_{height} = a_{bottom} - a_{top} + 1 \quad (2)$$

$$b_{width} = b_{right} - b_{left} + 1 \quad (3)$$

$$b_{height} = b_{bottom} - b_{top} + 1 \quad (4)$$

$$h = a_{width} + b_{width} - [\max(a_{right}, b_{right}) - \min(a_{left}, b_{left})] \quad (5)$$

$$v = a_{height} + b_{height} - [\max(a_{bottom}, b_{bottom}) - \min(a_{top}, b_{top})] \quad (6)$$

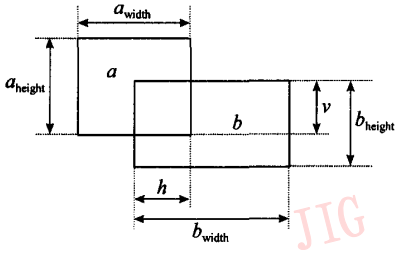
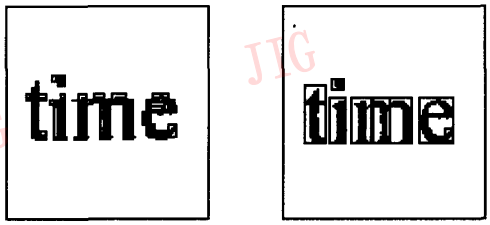


图 2　两个矩形的位置关系

Fig. 2　Relation between two rectangles

若同时满足 $h \geq -1$ 及 $v \geq -1$,则合并矩形 a 和 b。其中,a,b 的下角 left,top,right,bottom 分别表示矩形框的左、上、右、下 4 个坐标。图 3(b)所示为在图 3(a)的基础上合并矩形框形成模式的结果。

模式链表形成的步骤如下:先新建一个模式,取出矩形框链表中的第 1 个矩形框并存入新建模式的子链表中(该模式就是模式链表的第一个模式)。再取出下一个矩形框,判断与上一个矩形框是否满足合并条件,如果满足就存入当前模式的子链表中,否则新建模式,把这个矩形框存入这个新建模式的子链表中,并把这个模式存入模式链表。继续抽取



(a)形成矩形框　　(b)形成模式

图 3　矩形框和模式的形成

Fig. 3　Results of the rectangles and patterns formation

矩形框,若它与模式链表中各模式子链表内的任何一个矩形框满足合并条件,就把它存入这个模式的子链表中,否则新建模式并将其存入该新建模式的子链表中,再将该模式存入模式链表。模式链表的示意图,如图 4 所示。最后整个图像就由这条模式链表来表示,而每个模式内又含有一条子链表,即形成该模式的矩形框链表。

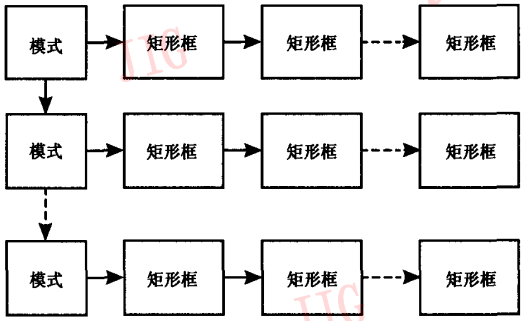


图 4　模式链表的示意图

Fig. 4　Schematic illustration of the pattern list formation

3　模式分类

3.1　按模式的统计特征分类

文本页面图像中,文字区域的模式通常为一个字符或几个字符,而非文字区域的模式却可能是形状各异、尺寸不一的矩形。用于图像分类的统计特征有很多,当然也可用来作为模式的特征。文献[6]中,在按照模式的统计特征进行分类时,使用了大量的统计特征,这些特征不仅计算起来繁琐,而且许多特征在分类过程中没有起到任何作用。模式的统计特征有:模式中黑像素的个数、模式的大小、黑游程的个数、黑白像素比等等。通过观察文字与图片,这里取文字和图片最本质的差别作为分类依据,即模式的最大黑游程(MAXBRL),就可以达到较好

的效果。对于一个汉字模式来说，如果模式大小为 30×30 像素，则其最大黑游程不可能大于 30。于是，只要存在某个模式的最大黑游程大于这个值，就将其分类为图片。通常使用的黑游程是水平方向上的。但是考虑到图片的无方向性，这里加入了垂直方向的黑游程。即两个方向中任一方向上的最大黑游程大于 MAXBRL 都将分类为图片，这样就可以将文本图像中尺寸较大的图片模式挑选出来，并标记为 I 类（即图片模式），其余皆为 T 类（即文字模式）。通过对大量的文本资料进行统计得知，对于分辨率为 300dpi 的扫描图像，一般的字体大小不会超过 35×35 像素，而对于分辨率小于 300dpi 的图像，其字体大小也不会大于这个值，因此实验中的 MAXBRL 取值为 35。对于一般图片区和文字区形状较规则或者图片区和文字区分开的文本页面图像，通过这个分类过程就可以得到较好的效果了。

3.2 按模式上下文分类

由于文本页面的复杂性，有可能存在某些图片模式，其特征与文字模式特征近似，仅使用模式的统计特征无法将图片与文字完全分离开来。这时考虑到每个模式并不是孤立存在于文本页面中，对于一幅图片来说，其通常是由几个大小不同的相互交叉或相邻的模式构成，成行成段的文字亦是如此。每个模式都与其邻近模式的分类存在某种关系，模式的上下文就是利用了这种关系。

确定模式上下文的方法：首先需要确定一个参数，即模式的邻域半径，用 R_{NEI} 表示。对于给定的模式 a，判断另外一个模式 b 是否属于 a 的上下文模式，判决规则如下：

$$b_{right} > a_{left} - R_{NEI} \tag{7}$$

$$b_{left} < a_{right} + R_{NEI} \tag{8}$$

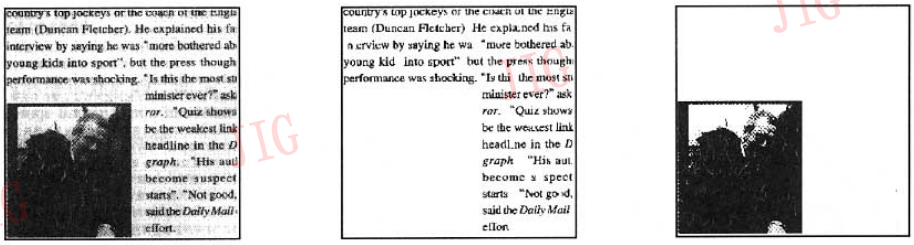
$$b_{bottom} > a_{top} - R_{NEI} \tag{9}$$

$$b_{top} < a_{bottom} + R_{NEI} \tag{10}$$

如果上述的 4 个条件全部满足，则认为模式 b 为模式 a 的上下文，本实验使用的 R_{NEI} 为 20。按模式的统计特征分类完毕后，尺寸较大的图片模式被标记为 I 类。由于存在不规则的图片，尺寸较大的 I 类模式必定与其他图片、文字模式相交或相邻，而文本页面中的文字必定成行或成段出现，因此文字模式的上下文模式中必定存在文字模式。通过这种模式上下文的定义，可以把所有与大图片模式交叠的文字模式挑选出来。本文提出的这种按照模式上下文的分类方法仅对 T 类模式进行操作。遍历模式链表，先把和大图片模式（即 I 类模式）相交或相邻的 T 类模式重新标记为 O 类，表示模式的分类未定。这里的相邻指两个矩形框在水平和垂直方向相差的像素数不多于 3 个，如图 2 所示，为两个模式的位置关系，用算式表示为 $h \geq -3$ 且 $v \geq -3$，然后对这样的模式查看其上下文，如果存在有文字模式（即 T 类模式），则该模式就被判为 T 类，该过程需要遍历链表几次，就可以把所有的文字正确地分离出来，剩余的 O 类模式则被判为 I 类。最后将 T 类模式存入文字链表中，I 类模式存入图片链表中，分别输出，就得到了分类好的两幅图。

4 实验结果

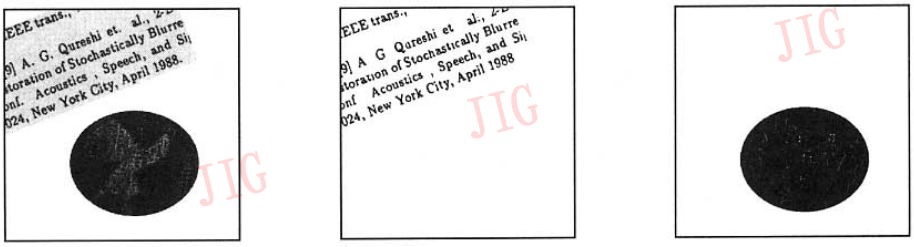
为了验证本文提出算法的有效性，使用 3 幅灰度图像，其中，一幅为通常意义下的文字区和图片区都较规则的图像（图 5）；一幅为文字倾斜的图像（图 6）；另外一幅为形状较复杂的图片镶嵌于文字当中的图像（图 7）。这些实验图像均来源于杂志和文献资料。进行分类之前，需将图像进行二值化。与传统的分割与分类方法相比，这种方法适用于多种版式的文本页面图像，而且对于图片形状复杂的



(a) 原始图像 (256×256)　(b) 文字区域　(c) 图片区域

图 5 图片区和文字区形状都较规则的页面图像

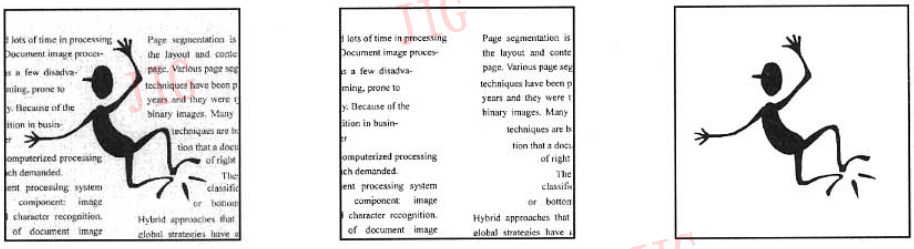
Fig. 5 Document image with regular halftone and text regions



(a)原始图像 (500 × 500)　(b)文字区域　(c) 图片区域

图 6　文字区倾斜的页面图像

Fig. 6　Document image with skewed text region



(a)原始图像 (1 024 × 1 024)　(b)文字区域　(c) 图片区域

图 7　含有不规则图片镶嵌在文字中的页面图像

Fig. 7　Document image with irregular halftone regions enchased in the text regions

面以及含有倾斜文字的页面都能取得很好的分割效果。从结果可以看到,文字被很好地提取了出来。传统的分割算法,如基于投影轮廓切分的 RXYC 算法[5],由于投影轮廓中不存在明显的谷点,因此将无法对图 7(a)这样版式复杂的图像页面进行正确的分割。

5　结　论

在文献[6]的用模式链表表示文本页面图像的基础上提出了一种新的分割与分类算法。由于文献[6]中的模式分类过程不仅运算复杂、不易实现,而且规则繁琐、不易理解。因此本文对模式分类进行了改进:首先在独立模式分类中减少了分类特征的个数,只使用了最大黑游程一个特征,就取得了很好的分割效果;其次在上下文分类中,不用对所有的模式操作,而仅对大图片模式周围个别不确定的模式进行二次分类。这样更进一步提高了算法的运算速度。最后的结果分两类输出,其中的文字区可以直接进入 OCR 系统,图片区可进入图片库保存。而且本算法不受文字的倾斜,图片形状复杂以及图文混排图像的影响。

参考文献(References)

1　Abele L, Wahl F, Scherl W. Procedures for an automated segmentation of text, graphic and halftone regions in documents[A]. In: Proceedings of the 2nd Seandinavian Conference on Image Analysis[C], Hellsinkii, 1981: 177 ~ 182.

2　Nagy G, Seth S C. Document analysis with an expert system[A]. In: Pattern Recognition Practice[M], Gelsema E S, Kanal L N Editors, North Holland: Elsevier Science Publishers B. V., 1986: 149 ~ 159.

3　Strouthopoulos C, Papamarkos N. PLA using RLSA and a neural network[J]. Engineering Applications of Artificial Intelligence, 1999, **12**(2): 119 ~ 138.

4　Shulan Deng, Shahram Latifi, Emma Regetova. Document segmentation using polynomial spline wavelets[J]. Pattern Recognition, 2001, **34**(12): 2533 ~ 2545.

5　Jaekyu Ha, Robert M Haralick, Ihsin T. Philips recursive X-Y cut using bounding boxes of connected components[A]. In: Proceedings of Third International Conference on Document Analysis and Recognition[C], Montréal, Canada, 1995: 952 ~ 955.

6　Mitchell Phillip E, Yan Hong. Document page segmentation based on pattern spread analysis[J]. Optical Engineer, 2000, **39**(3): 724 ~ 734.

7　Tseng Lin Yu, Chen Rung Ching. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming[J]. Pattern Recognition Letters, 1998, **19**(10): 963 ~ 973.