

基于形态学的二值文档图像背景归一化问题研究

杨德强¹ 施心陵²

(云南师范大学计信学院 昆明 650092)¹ (云南大学信息学院 昆明 650091)²

摘要 对于包含白底黑字和黑底白字两种背景的二值文档图像,要进行后续的文字切分或 OCR,首先必须进行背景色的归一化。本文提出了一种采用数学形态学的方法对两种背景进行形态滤波,取出图像的黑色背景区域,再把黑色背景区域在原图中的对应区域反色,实现图像背景归一化。

关键词 数学形态学,形态滤波,背景归一化

1 引言

文档图像处理系统的目的是在数字图像中识别文字、图形和图片,并按照人们的需要提取出所需要的信息。其中的文字处理包括段落、行和单字的切分及随后进行的 OCR^[1] (Optical Character Recognition, 光学字符识别)。但是如果一幅文档图像不仅仅是单一的白底黑字(或黑底白字),而是包含了白底黑字和黑底白字两种背景的图像,显然这是目前 OCR 系统不能处理的。必须把该图像转换成背景一致的白底黑字图像才能进行后续的文字处理。

在传统的二值文档图像背景归一化算法中,直接以字为目标,对图像进行特征提取,从而实现背景归一化。这些算法对不同背景上的汉字,一般要求具有一致的字体、字号,且字体、字号要预知。当一个汉字同时在两种背景中时,往往无法提取。而且这些算法会造成汉字失真较大,需要对提取出的汉字进行轮廓跟踪^[2],修正产生的失真,算法复杂,运算量大。

对此本文提出了一种基于数学形态学^[3]方法的文档图像黑白背景归一化的方法。要进行背景归一化,其实就是要将图像中黑底白字的部分转换为白底黑字(假定背景归一化的目标是实现白底黑字)。如果能够取出文档图像中的黑底白字区域,把该区域进行反色,再并入原图即可实现文档背景的归一化,本文就是基于这种设想进行的。利用数学形态学的开闭运算所具有的滤波^[4]功能,去掉文档图像上的白底黑字和黑底白字,得到只含有白底和黑底但不含有文字的一些黑白区域,再把黑色区域在原图中的对应区域进行反色变换,即可实现文档图像的背景归一化。

2 基于形态学的文档图像背景归一化

2.1 数学形态学的形态滤波

形态学中定义了两个基本变换(运算),分别是膨胀(dilate operator)和腐蚀(erode operator),把这两个基本变换按一定的顺序进行组合便可构成一些具有特殊功能的运算,其中最重要的便是开运算和闭运算。这些运算的定义如下:

膨胀: $A \oplus B = \{x: (-B + x) \cap A \neq \emptyset\}$

腐蚀: $A \ominus B = \{x: B + x \subseteq A\}$

开运算: $A \circ B = (A \ominus B) \oplus B$

闭运算: $A \cdot B = (A \oplus B) \ominus B$

其中开运算可以磨光外边界,把比结构元素小的孤立点,突刺滤掉,闭运算可以磨光内边界,把比结构元素小的缺口或孔填充上,而且这两种运算都能保持图像总的位置不变。

把包含白底黑字和黑底白字的文档图像上的黑底看作是目标图像,而把白底上的黑字和黑底上的白字都看作是图像上的并噪声和差噪声,这样我们要取出黑色背景区域的操作,变成为对图像的去噪。

我们采用形态滤波器(开闭滤波器)来完成对图像的去噪,开运算去除白底上的黑字,而闭运算可以填充黑底上的小孔,即把黑底上的白字去掉,从而达到提出黑色底纹区域的目的。

2.2 文档图像背景归一化

2.2.1 提取黑色底纹 当然,要准确地提取出黑色底纹区域,干净地去除白底上的黑字,做一次形态滤波显然不够,我们采用交替序列滤波器,滤波器的构造如下:

$S = (S \cdot B) \ominus B$ 或 $S = (S \ominus B) \cdot B$

其中, S 为目标图像, B 为结构元素。

初始时用一个很小的结构元素开始滤波,例如,采用 3×3 (像素)的结构元素开始滤波,然后逐渐增大结构元素的尺寸,逐步消除图像上的孤立点和孔洞,即消除图像上的白底黑字和黑底白字,最后图像上只剩下无字的黑底和白底,从而取出图像上的黑

底。

2.2.2 文档背景归一化 通过上述形态滤波后,提取出图像的黑色底纹,只要把原图上与黑色底纹对应区域作反色变换,即可实现文档背景归一化。

算法如下:

(s1,s2,s3 分别表示图 1 中的各图像)

i. 扫描经过滤波后的图像 S3

如果 s3 中的当前点(point)的像素值为 0,说明是黑底;

则把在 s2 中的对应区域反色。即把 s2 中的当前点(point),白点变为黑点,黑点变为白点。

继续检查下一像素点(point = point + 1);

ii. 输出经过上述步骤处理后的图像 s2,即为背景归一化后的图像。

3 背景归一化算法流程图

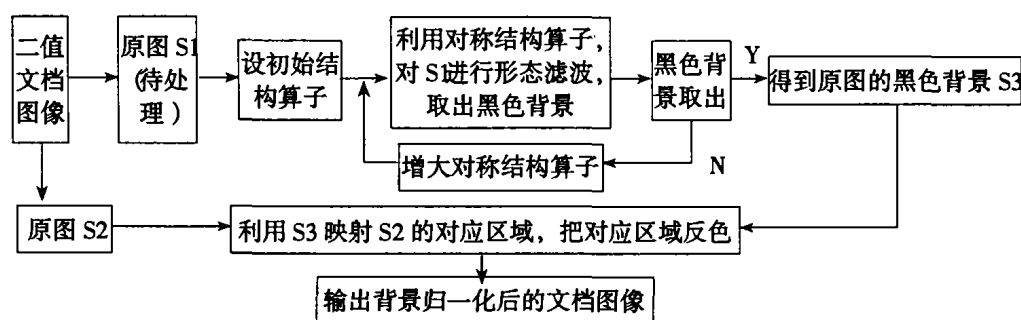
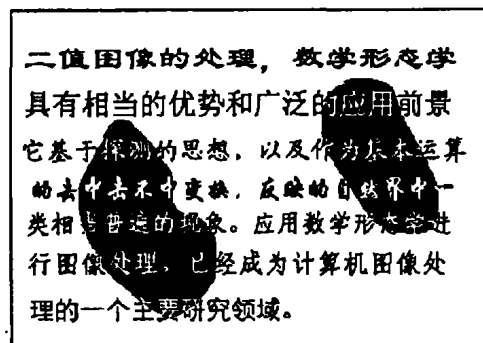
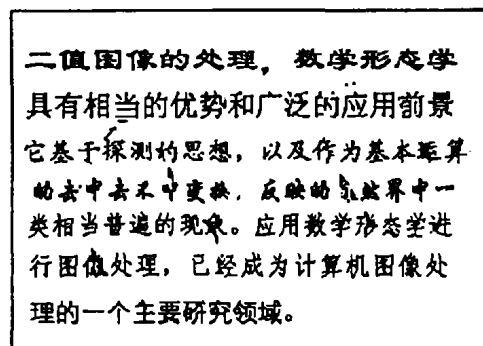


图 1 背景归一化流程图

4 实验结果



(a) 包含两种背景的原图像



(b) 用本算法处理后的图像

图 2

文中提出的算法已通过 Visual C++ 6.0 编程实现, 所得实验结果如图 2 所示。图 2(a)有几种不同

字体字号的两种背景的文字, 而且有同一个字处在不同背景中的情况。图 2(b)是采用本算法进行背景归一化, 将黑底白字转变成白底黑字的结果。由于文中的算法是以底纹作为目标, 而把汉字作为滤波的对象, 因此汉字的字体、字号对滤波产生的影响很小。该算法充分利用二值文档图像的特点, 通过对背景的操作, 达到图像特征提取, 克服了传统算法要求汉字字体、字号一致且已知的缺点, 而且还能对一个汉字同时处于不同背景中的情况进行背景归一化, 使得算法具有更广的通用性。

结论 本文对包含两种背景的二值文档图像的背景归一化问题进行了研究, 提出了基于数学形态学的形态滤波, 取出黑色背景区域, 再对原图进行反色, 从而实现文档背景的归一化。实验表明, 该算法可实现文档图像较小失真的背景归一化。

参考文献

- 1 Lu Yi. Machine Printed Character Segmentation An Overview[J]. Pattern Recognition, 1995, 28(1): 67 ~ 80
- 2 孙家广, 杨长贵. 计算机图形学[M]. 北京: 清华大学出版社, 1994
- 3 Serra J. The 'Centre De Morphologic Mathematique' An interview. In Mathematical Morphology and its Applications to Image Processing, Ed., Serra, J., and Soille, P., Kluwer Academic Publishers, 1994. 369 ~ 374
- 4 Dougherty E R, Astola J. An Introduction to Nonlinear Image Processing. SPIE Press, 1994