文档处理中背景字符的去除*)

张重阳[1,2] 杨静宇[1] 李 伟[1] 孙明明[1]

(南京理工大学计算机科学与技术系 南京 210094)[1]

(中创软件工程股份有限公司博士后工作站 济南 250014)[2]

摘 要 识别域图像的提取是文档自动处理系统中一个重要的预处理过程。在实际应用中,用户填写的信息常常与版面中的框线和背景字符存在交叠现象,严重影响了系统的性能。本文提出了基于点边距离分析的背景字符去除算法。首先通过灰度图像匹配的方法精定位背景字符子图像;然后利用形态学方法结合笔画的宽度信息对背景字符子图像进行二值化;最后分析像素点到边界距离的变化确定需要填充的像素位置,并通过形态学方法计算像素的填充值。实验采用了真实票据图像中的日期域,实验结果表明本文的方法获得了基本令人满意的效果,背景字符像素被成功去除。

关键词 图像处理,文档图像分析,图像匹配,二值化,数学形态学

Removing of Preprinted Characters in Document Image Processing

ZHANG Chong-Yang[1,2] YANG Jing-Yu[1] LI Wei[1] SUN Ming-Ming[1]

(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)[1]

(CVIC SE Co. Ltd., Jinan 250014)[2]

Abstract Extraction of recognition item is an important preprocess procedure in a Document image analysis system. In reality, user fill-in data usually cross or touch the preprinted lines and characters, creating tremendous problems for the recognition engines. In this paper, we proposed a practical preprinted character removing method. Image matching algorithm is applied to locate the position of the preprinted character, and then the character image is binarized by mathematical morphology method combing with stroke width information. Last, the preprinted character is removed based on the varying of stroke contours. Experiment results on real-life check images demonstrate the efficient of the proposed method.

Keywords Image processing, Document image analysis, Image matching, Binarization, Mathematical morphology

1 引言

表格文档在人们的日常生活中发挥着重要的作用,实现其自动录入、存储、管理和检索具有极其重要的现实意义。文档自动处理系统中的一个重要处理过程是从图像中提取出用户填写的字符图像。通常表格文档由三个部分组成:版面框线,版面(背景)字符或符号,用户填写的内容。其中前两个部分是预先打印在文档上的。很多情况下,用户填写/打印的内容与这些预打印的版面信息存在粘连或交叠,严重影响了后期字符串的分割与识别。若直接将这些预打印信息擦除,会出现字符笔画断裂、字符结构发生变化等情况,造成系统产生许多单字的误识和拒识。因此如何去除预打印的版面信息同时保持填写内容的完整已成为一个实用 OCR 系统的重要环节之一。

关于直线的检测与去除国内外已有许多报道[1,2]。文[3]中提出了一种基于灰度图像分析的表格框线去除算法,在实际应用中获得了比较理想的效果。完整地提取用户填写的内容是一个非常困难的问题,尤其是在它们与背景字符有交叠时[1]。这方面的研究国内还未曾见报道,国外的相关报道也不多。S Liang 等人[4]提出了基于骨架分析的手写标记符号去除方法,相对于打印字符,标记符号要大得多并且具有比较光滑的曲率,通过骨架分析去除短枝可以得到标记符号的骨架。这一方法运算复杂,只能解决笔画简单并且比背景字符大的多的手写标记符号去除问题。Ye 等人[5]在去字的过程中利用标准版面图像定位背景字符区,然后通过字符的笔画宽度信息去除背景字符,这一算法要求填写字符的笔画比背景字符的笔画宽,图像中所有宽度小于给定阈值的笔画都会被去除。

本文以去除支票图像中用于日期定位的背景字符为例,提出了通过各方向上点边距离分析的去字算法,实验中获得了比较好的效果。第 2 节重点介绍算法的基本原理和实现过程,后两节分别给出了实验结果和结论。

下面的分析假定图像中填写字符和背景字符均为深色(灰度值低),背景为浅色(灰度值高)。

2 算法实现

背景字符的去除分为两个过程:(1)检测图像中背景字符的区域。在没有先验知识的情况下,背景字符的区域很难检测,实际应用中这一问题可以通过版面定义、模板图像匹配的方法解决。(2)去除背景字符像素。背景字符区内的像素不能直接去除,否则在字字交叠的地方会出现目标字符笔画断裂、字符结构发生变化等情况,因此需要进行特殊处理。图 1 为直接去除背景字符的结果,填充色为纯白,字字交叠处的像素被去除造成了字符笔画的断裂。

在字字交叠处,字符笔画的边界发生改变,我们定义点到边界的距离来检测这些发生变化的位置,去除背景字符。本文背景字符去除算法的实现大致分以下几个步骤:

(1)构造背景字符的灰度模板图像并生成相应的掩模图像;

(2)通过模板图像匹配的方法定位处理图像中的背景字

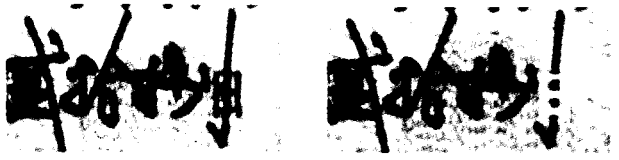
*)基金项目:电子信息产业发展基金(信部运[2003]446 号)。

符子图像；

(3)形态学方法结合笔画的宽度信息对背景字符子图像进行二值化；

(4)根据像素点到边界距离特征区分背景字符中的保留像素和填充像素；

(5)通过形态学方法计算字符像素的局部背景值并填充背景字符像素。



a 原图　　b 直接去除背景字符

图 1　字字交叠图像示例

下面介绍算法的具体实现过程。假设处理图像为 $I=\{I(x,y),(x,y)\in[M\times N]\}$；标准背景字符图像为 $I_b=\{I_b(x,y),(x,y)\in[M_b\times N_b]\}$。

2.1　背景字符检测

图像匹配技术是根据已知的图像模块在另一幅图像中寻找相应或相近模块的过程。这方面已有很多的报道[6,7]，本文采用灰度相关的方法[6]寻找模板字符的左上角坐标，灰度相关归一化公式表示为：

$$R(i,j)=\frac{\sum_{x=0}^{M_b-1}\sum_{y=0}^{N_b-1}[(I(x+i,y+j)-\overline{I(i,j)})\times(I_b(x,y)-\overline{I_b})]}{[\sum_{x=0}^{M_b-1}\sum_{y=0}^{N_b-1}(I(x+i,y+j)-\overline{I(i,j)})^2\times\sum_{x=0}^{M_b-1}\sum_{y=0}^{N_b-1}(I_b(x,y)-\overline{I_b})^2]^{1/2}}$$

其中，$\overline{I(i,j)}=\frac{1}{M_b\times N_b}\sum_{x=0}^{M_b-1}\sum_{y=0}^{N_b-1}I(x+i,y+j)$，$\overline{I_b}=\frac{1}{M_b\times N_b}\sum_{x=0}^{M_b-1}\sum_{y=0}^{N_b-1}I_b(x,y)$

在不考虑倾斜的情况下，匹配模板左上角的坐标(x_0,y_0)对应 $R(i,j)$取最大值时的(i,j)，即：$(x_0,y_0)=\underset{(i,j)\in D}{\operatorname{argmax}}R(i,j)$

其中 D 是搜索的区间。文档处理系统中，前期的版面分析过程可以初步定位背景字符，因此这里的匹配是字符区域的精定位过程，D 可以限定在一个很小的范围内。

背景字符子图像 $I_f=\{I_f(x,y),I_f(x,y)=I(x+x_0,y+y_0)\text{AND}(x,y)\in[M_b\times N_b]\}$。

标准背景字符的不规则区域使用膨胀的掩模图像来描述。首先采用最大方差阈值法[8]对字符模板图像进行二值化，前景像素标记为 1，背景像素标记为 0；然后对二值图像作膨胀运算，膨胀半径取 1，得到掩模图像 $I_b^1=\{I_b^1(x,y),I_b^1(x,y)=1\ \text{AND}(x,y)\in[M_b\times N_b]\}$。

图 2a 是得到的掩模图像，图 2b 是图 1a 中的背景字符子图像。

a 掩模图像；　b 背景字符子图像；　c 图 b 的二值图；

图 2　掩模图像和背景字符子图像

2.2　字符图像二值化

图像中的字符可以看作是笔尖在背景上移动产生的，除笔画交叉或重叠外，一般具有细长型的结构特征。图 3 是笔画和大块物体理想情况下的一维剖面示意图，笔画的宽度较小，其正负两条阶越型边缘之间的距离通常小于大块背景的距离。选取尺度合适(大于 W_1 且小于 W_2)的滤波窗口能够在保留笔画的同时消弱图像中的大块背景和光照不均。数学形态学[10]在提取图像中特定结构的目标上具有较强的优势。假设背景字符的笔画宽度为 SW，则用形态学检测笔画可以表示为：

$I_f^1=(I_f\cdot B)-I_f$，其中 B 取直径为 SW 的圆形结构元素。

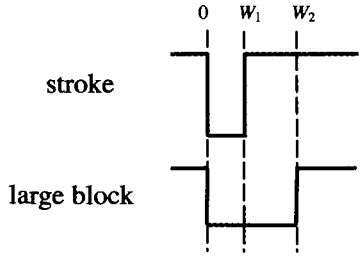


图 3　理想笔画一维剖面示意图

背景字符笔画宽度 SW 的估计。首先对掩模图像作 0，$\pi/4$，$\pi/2$，$3\pi/4$ 四个方向的连续黑像素的游程长度作统计，得到相应的游程长度直方图，它反映了图像中连续黑像素长度的分布情况。由于掩模图像中基本上只含有笔画的信息，因此直方图中最大值对应的游程长度即为估计的平均笔画宽度。

I_f^1 描述了像素局部灰度的变化，将其作为一幅灰度图像二值化，得到 $I_f^2=\{I_f^2(x,y),I_f^1(x,y)>T\ \text{AND}\ (x,y)\in I_f^1\}$，其中 T 为最大方差阈值。背景字符子图像的二值化结果如图 2c。图 4 是本文方法和最大方差法对具有大块背景的字符图像的二值化结果比较。

a 原图；　b 最大方差法；　c 形态学方法

图 4　有大块背景的图像二值化结果

2.3　背景字符去除

点到边界的距离：二值图像中任意一个黑像素到边界的距离表示为向量 $v(x,y)=\{r^d(x,y)\,|\,d=0,1,\cdots,7\}$，其中 r^d 表示 d 方向上点到边界的距离，$d=0,1,\cdots,7$ 代表方向 $d\times\pi/4$，如图 5a 所示。图 5b 是笔画中像素 p 到边界距离的示意图。处理图像和掩模图像的点边距分别表示为 $v_f(x,y)=\{r_f^d(x,y)\,|\,d=0,1,\cdots,7\}$ 和 $v_b(x,y)=\{r_b^d(x,y)\,|\,d=0,1,\cdots,7\}$。

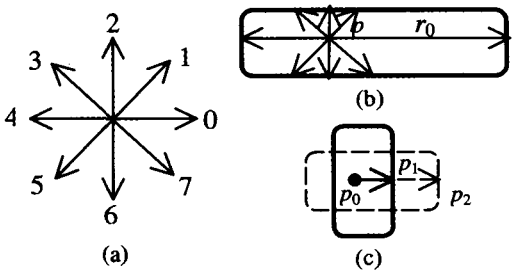


图 5　点到边界的距离

在没有笔画相交时，背景字符区内任意一点到边界的距离应不大于标准模板中对应的距离。在填写字符与背景字符重叠的区域，由于字符的边界发生改变，区域内点到边界的距离在某一个方向上大于标准模板中对应的距离。图 5c 中，竖笔画和横笔画分别用实线和虚线表示。在 0 方向上，重叠区域内的像素点 p_0 到竖笔画边界的距离为 $|\overline{p_0p_1}|$，到整个连通区边界的距离为 $|\overline{p_0p_2}|$，显然 $|\overline{p_0p_2}|>|\overline{p_0p_1}|$。

根据这一原理我们将图像划分为保留区和填充区，保留区内像素的值不变，填充区内像素认为是只属于背景字符，要用背景色填充。填充区像素集合为：$S=\{s(x,y)\mid \forall d,(r_f^d(x,y)-r_b^d(x,y)\leqslant 0 \text{ OR } r_b^d(x,y)>SW) \text{ AND } (x,y)\in I_f^2\cap I_b^1\}$

填充色的计算采用了形态学的方法，分两步：(1)计算掩模图像的边缘点，在边缘点进行腐蚀运算，$I_f^3=I_f\ominus B,(x,y)$ 为边缘点；(2)对掩模像素进行膨胀运算，得到掩模内像素的填充色，$I_f^4=I_f^3\oplus B,(x,y)\in I_b^1$。

图 6　图 1a 的去字后图像

去字后处理图像的像素值为：

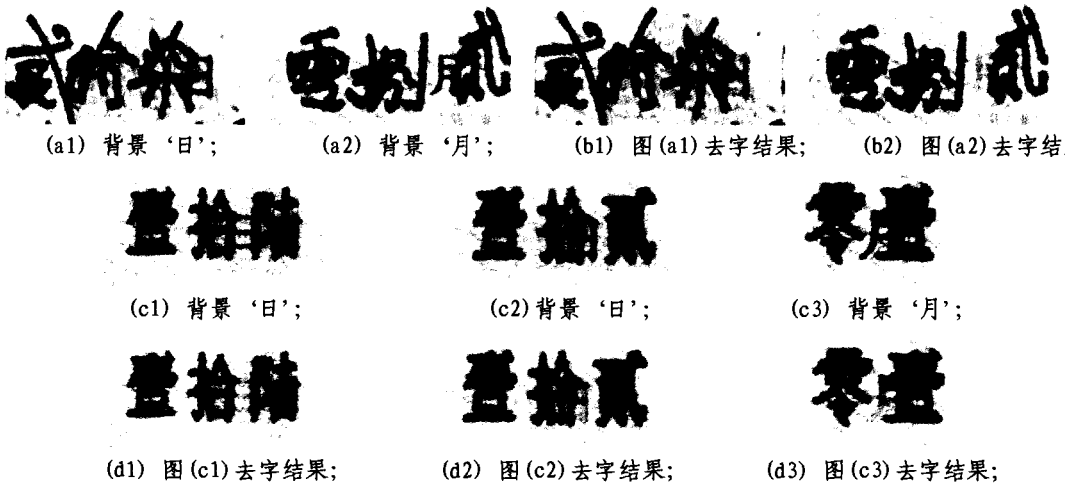
$$I_f^5(x,y)=\begin{cases}I_f(x,y), & \text{if}(x,y)\in\overline{S}\\ I_f^4(x,y), & \text{if}(x,y)\in S\end{cases}$$

图 6 是图 1a 的去字后图像。

3　实验结果

为检验本文提出的背景字符去除算法的性能，我们对真实支票图像中的日期域进行实验。图像由清分机扫描获得，人工选取字字交叠的图像，图像的分辨率为 200DPI，灰度级为 256 级。

图 7 是背景字符去除的部分实验结果，背景字符为‘日’或‘月’。图 a1，a2 的填写字符为手写体，实验结果如图 b1，b2，图 c1-3 的填写字符为打印体，实验结果如图 d1-3。实验结果可以看出，本文提出的背景字符图像去除算法获得了基本令人满意的效果，背景字符大部分被成功去除。由于填充像素的灰度值是根据局部背景的灰度值选取的，因此笔画相交复杂的区域填充值一般比较低。本文算法要求字字相交时，背景字符连通区的边界有较大的改变。对于图 a2 中存在笔画相切的情况(‘月’右竖线)，算法也获得了较好的实验结果(图 b2)。但是当笔画相切并且重叠比例很大时，如图 c1 中与‘日’的竖笔画重叠区域，本文的方法会将部分填写字符像素作为背景字符像素来处理，如图 d1 所示。



(a1) 背景 ‘日’；　(a2) 背景 ‘月’；　(b1) 图(a1)去字结果；　(b2) 图(a2)去字结果；

(c1) 背景 ‘日’；　(c2) 背景 ‘日’；　(c3) 背景 ‘月’；

(d1) 图(c1)去字结果；　(d2) 图(c2)去字结果；　(d3) 图(c3)去字结果；

图 7　背景字符去除实验结果

结论　随着字符识别技术和计算机技术的发展，能够自动处理的文档也越来越多。在一些文档图像中，用户填写的信息常常与预打印的背景字符存在粘连或交叠，严重影响了系统后期字符串的分割与识别。目前，国内外关于如何去除这些背景字符的研究报道还很少。本文提出了基于点边分析的方法检测字字交叠并去除背景字符。同时提出了形态学方法结合笔画宽度信息的字符图像二值化方法。在真实票据图像上获得了较好的实验效果。但是本文方法对于笔画相切甚至是互相掩埋的情况处理不好，这是结构方法的局限，这一问题的解决可能需要结合灰度信息或复杂的识别反馈过程。

参 考 文 献

1　Bin Yu, Jain A K. A generic system for form dropout. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1996, 18(11): 1127～1134

2　Tseng Yi-Hong, Lee Hsi-Jian. Interfered-character recognition by removing interfering-lines and adjusting feature weights. In: Proc. Fourteenth Int. Conf. on Pattern Recognition, Brisbane, Qld. Australia, 1998. 1865～1867

3　张重阳，陈强，娄震，杨静宇. 基于灰度图像的表格框线去除算法. 计算机研究与发展，2005，4(42)：635～639

4　Liang S, Ahmadi M, Shridhar M. Segmentation of handwritten interference marks using multiple directional stroke planes and reformalized morphological approach. IEEE Trans Image Process, 1997, 6(8): 1195～1202

5　Ye Xiangyun, Cheriet M, Suen C Y. A generic method of cleaning and enhancing handwritten data from business forms. Document Analysis and Recognition, 2001, 4(2): 84～96

6　罗钟铉，刘成明. 灰度图像匹配的快速算法. 计算机辅助设计与图形学学报，2005，17(5)：966～970

7　孙远，周刚慧，赵立初，等. 灰度图像匹配的快速算法. 上海交通大学学报，2000(34)5：702～704

8　Otsu N. A threshold selection method from grey-level histograms. IEEE Trans. Sys., Man, Cybern, 1978, 8: 62～66

9　崔屹. 图像处理与分析-数学形态学方法及应用. 北京：科学出版社，2000，4