计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(11):238−244 [doi: 10.15888/j.cnki.csa.007157]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于贪婪算法的文档图像中干扰线的去除①

王　平[1,3,4], 张晓峰[2], 王宜怀[3], 程仁贵[1,4]

[1](武夷学院 数学与计算机学院, 武夷山 354300)
[2](南通大学 信息科学技术学院, 南通 226019)
[3](苏州大学 计算机科学与技术学院, 苏州 215006)
[4](认知计算与智能信息处理福建省高校重点实验室, 武夷山 354300)
通讯作者: 王　平, E-mail: 122742928@qq.com

摘　要: 各种文档中经常包含有各种特殊作用的横线、手划线等, 当这些文档通过扫描等数字化方式存入计算机并需要进一步识别处理成文字编码时, 这些线条却成为 OCR 的干扰因素, 降低了文档内容的识别率. 为此, 本文提出一种新的文档干扰线去除算法, 先将文档图像二值化, 二值化过程考虑了不均匀光照带来的影响; 然后将前景细化为单像素, 减少线条粗细造成的影响; 接着通过一种改进的贪婪算法计算横、竖两个方向线段的权重, 判断权重较高的线段为干扰线; 最后通过与干扰线距离的大小判断图像中每个前景像素的归属, 从而获得一个完整的文档恢复图. 仿真实验表明, 本文提出的算法能够有效去除干扰线, 特别在干扰线与文字粘连的情况下, 去除干扰线的同时较少地影响文档图像的质量, 且具有较高的计算速度和较好的去除效果, 为图像进一步 OCR 识别提供了良好的基础.

关键词: 二值化; 干扰线去除; 贪婪算法; 光学字符识别

引用格式: 王平,张晓峰,王宜怀,程仁贵.基于贪婪算法的文档图像中干扰线的去除.计算机系统应用,2019,28(11):238–244. http://www.c-s-a.org.cn/1003-3254/7157.html

Interferential Line Elimination in Document Image Based on Greedy Algorithm

WANG Ping[1,3,4], ZHANG Xiao-Feng[2], WANG Yi-Huai[3], CHENG Ren-Gui[1,4]

[1](School of Mathematics and Computer Science, Wuyi University, Wuyishan 354300, China)
[2](School of Information Science and Technology, Nantong University, Nantong 226019, China)
[3](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)
[4](Fujian Provincial Key Laboratory of Cognitive Computing and Intelligent Information Processing, Wuyishan 354300, China)

Abstract: Documents often contain horizontal lines, hand lines, etc., which are used for various special functions. When these documents are stored in computers by scanning or the like and need to be further recognized and processed into text codes, these lines become interference factors of OCR, thus the recognition rate of document content is decreased. This study proposes a new document interference line removal algorithm, which first binarizes the document image, and the binarization process takes into account the effects of uneven illumination; then the foreground is refined into single pixels, reducing the thickness of the lines. The effect is then calculated by an improved greedy algorithm to calculate the weights of the horizontal and vertical line segments, and the line segment with higher weight is determined as the interference line; finally, the distance of each foreground pixel in the image is determined by the distance from the interference line. Thereby obtaining a complete document recovery map. The simulation results show that the proposed algorithm can

① 基金项目: 国家自然科学基金 (61672369); 中央引导地方科技发展专项 (2018L3013); 福建省自然科学基金面上项目 (2015J01669,2017J01651);福建省教育厅中青年教师项目 (JA15522)
Foundation item: National Natural Science Foundation of China (61672369); Special Fund of Central Government for Local Science and Technology Development (2018L3013); General Program of Natural Science Foundation of Fujian Province (2015J01669, 2017J01651); Mid-Aged and Young Faculty Program of Education Buearu, Fujian Province (JA15522)
收稿时间: 2019-03-29; 修改时间: 2019-04-26; 采用时间: 2019-05-21; csa 在线出版时间: 2019-11-06

effectively remove the interference lines, especially in the case of interference lines and text adhesion, and remove the interference lines while affecting the quality of document images less, and has a higher computing speed and better removal effect. The removal effect provides a good basis for further OCR recognition of images.

Key words: binarization; interferential line elimination; greedy algorithm; OCR

纸质文档的内容数字化处理工作, 在各行各业中广泛应用, 目前较为典型、高效的方式是通过将纸质等重要文档通过扫描、照相等方法获得其数字化文档图像. 这些包含大量文字内容信息的数字化图像资料的好处在于, 一方面这些资料是数字化的文件, 很容易保存在计算机的存储器中, 方便存储和管理; 另一方面, 这些文档可以进一步使用 OCR 软件进行识别, 能够快速地获得文档图像中的内容, 避免了繁琐的文字输入工作. 目前, 如果文档仅仅是只包含文字的文档图像, 尤其是印刷体文字, 其 OCR 的识别率非常高, 能达到 99% 以上, 已经在各种领域中得到应用. 然而, 大多数文档中经常包含各种干扰信息, 如各种干扰线, 例如当人们在原始文档上留下横线等来标记文档中的重要内容, 或者文档本身就存在各种横线表示需要填写信息或者其他提醒时, 则文档图像的 OCR 的识别率会急剧下降. 因此, 如何去除文档图像中的干扰线成为文档图像 OCR 前的一个重要的预处理问题.

1 研究概况

文档图像去除干扰线的以往工作可以分为两类, 一类是规则线段, 另一类是不规则的手划线. 规则线段一般表现为印刷的下划线、表格的边缘线等. 对于规则线段的处理, Bai 等人[1]通过连通分量分析和下边缘分析策略获得干扰线的位置, 并去除, 但是该方法只能处理文档图像中标准的下划线去除; Shi 等人[2]为了去除手写阿拉伯数字中的规则的线段使用了一种 directional local profiling 方法, 但是该方法只能检测和去除预打印规则行线段; Alipour 等人[3]利用了规则横线的特征去除手写文档中的横线, 重点考虑了线的边缘检测; Imtiaz 等人[4]使用滑动窗口中的熵来判断当前区域中是否包含干扰线, 以便达到去除水平规则线和垂直边缘线的目的. 而对于不规则的干扰线, 比如手划线的表现较规则的线段去除复杂得多, 各种文献中也出现了多种方法, Cheng 等人[5]使用超图来检测图像中的干扰线, 采用主曲线方法、改进的最短路径法和方向偏移算法实现, 整个方法稍显复杂, 且没有说明当干扰线与字符笔画重合时如何只去掉干扰线, 而保留字符; Kaur 等人[6]基于连通元的 FCM 聚类、分类方法找到干扰线的区域并去除, 但其要求是取定类型的标注和下划线; Banerjee 等人[7]也使用连通元检测干扰线区域, 并对与文字粘连的干扰线特殊处理; Rehman 等人[8]首先处理了粘连的字符和干扰线, 然后判断连通元; Pratihar 等人[9]利用文档图像中的像素间的几何关系检测干扰线; Das 等人[10]利用 Gabor 滤波器和连通分量分析并检测干扰线; 近年来深度学习发展迅速, 干扰线去除领域也出现了许多使用深度学习网络来实现的方法[11,12].

基于以上文献的处理方法及分析其不足之处, 本文提出一种新的与文字耦合粘连干扰线的去除方法, 该方法处理中首先将文档图像进行二值化, 去除部分噪声像素后得到文档的主要部分; 然后通过细化, 得到单像素的线条; 接着计算这些线条的代价, 若代价超过预先设定的阈值, 则认为其是存在的一条干扰线; 最后, 通过前景像素与干扰线的距离判断其归属, 达到文字文档干扰线去除的目的. 此所提出的方法使用了贪婪算法计算线条代价, 整个算法速度较快, 并将提出的算法在多种类型的文档图像的测试中, 获得了较好的效果.

2 图像预处理

文档图像的预处理是为了提取出文字的主要特征部分, 减少噪声的干扰. 本文方法的预处理包含两个步骤, 图像二值化和图像细化, 其中图像二值化是为了提取出文档图像的前景内容部分, 而图像细化目的则是将前景部分的内容描述为中心线的形式呈现, 以减少线条的粗细对干扰线检测的影响.

2.1 图像二值化

图像二值化是一种常规的图像预处理方法, 且二值化的算法有很多, 性能也不尽相同. 对于文字比较清

晰的图像, 用全局阈值的方法就可以获得较好的二值化效果, 但是若获取的文档图像的质量较差, 就需要采用局部阈值方法进行处理. 考虑到实际应用中, 诸如通过扫描、照相等方法获得的文档图像的亮度变化不太均匀, 故而本文使用一种由局部阈值插值产生全部位置阈值的方法.

局部阈值是一个针对局部区域块的阈值, 这种局部区域块范围不能太大, 若太大则设置的阈值可能不适用, 无法有效检测出干扰线, 也不能很小, 若很小则也无法反映该局部区域块的前景和背景的像素强度分布.

本文将一幅大小为$m\times n$的文档图像等分成大小相等的块, 假设等分成$m_1\times n_1$个块 (即列方向等分成m_1份, 行方向等分成n_1份), 则每块中像素的数目为$(m\times n)/(m_1\times n_1)$. 为了保证每块中像素数目不太少, m_1和n_1均不能太大, 实验中它们的取值范围是 5~10 之间的整数. 每个小块中的局部阈值使用大津法获得, 然后使用线性插值法从局部阈值获得每个位置的阈值$T(i,j)$, 当然, 若使用非线性插值也许可以获得更好的效果, 但是消耗的时间开销也会相应增加. 因此, 此处的图像二值化可以描述为:

$$B(i,j)=\begin{cases}0, & \text{if } I(i,j)>T(i,j)\\ 1, & \text{if } I(i,j)\le T(i,j)\end{cases} \tag{1}$$

如图 1 所示, 图 1(a) 是一幅原始的文档图像, 图 1(b) 是其图像二值化的结果. 从图中可以看出, 提出的图像二值化方法在不均匀光照的文档图像中取得了较好的二值化效果.

2.2 图像细化

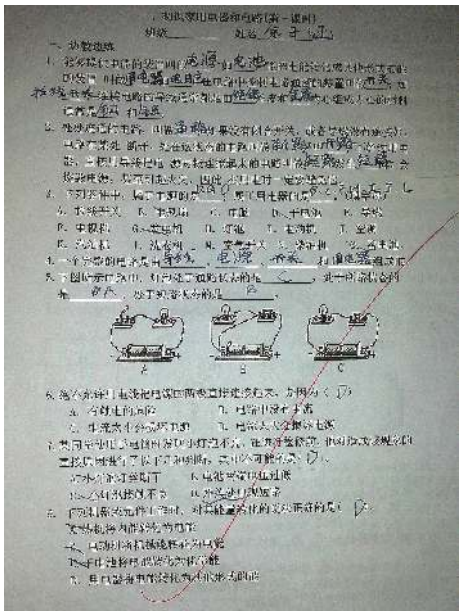
一般情况下, 文档中的文字和相关干扰线都可以通过笔画的中心线来判断, 而且中心线有效地去除了其他像素所引起的干扰因素, 能有效地降低判断的难度. 本文使用 matlab 中常用的图像细化方法, 该方法通过 8 个模板不断地消减二值化图中边缘多余像素, 能得到较好的细化效果. 如图 2 所示, 显示了图像细化的效果, 图 2(a) 是包含干扰线的图, 图 2(b) 是图 2(a) 的细化结果, 可以看出, 该方法细化后的能较好地保留文字的特征轮廓.

3 基于贪婪算法的图像干扰线检测

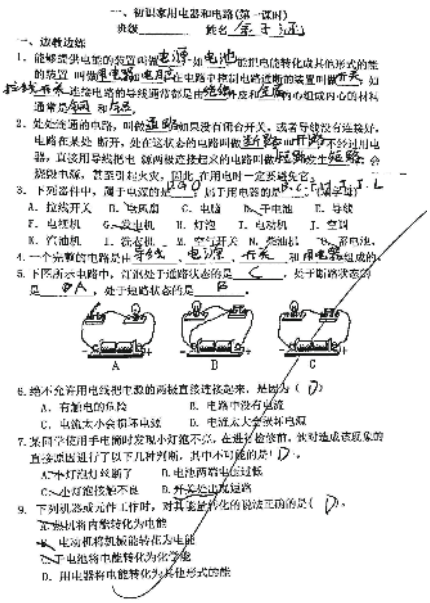
通过上述图像二值化和细化的预处理操作之后, 能获得文档的中心线. 进一步通过观察和分析, 干扰线的中心线具有以下和文字不同的表现特征:

(1) 干扰线一般为横向, 偶尔出现竖方向, 极少出现旋转方向;

(2) 干扰线一般较长, 远远大于文字字体的大小.



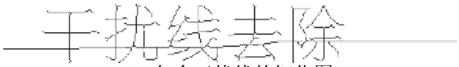
(a) 原图



(b) 二值化结果

图 1 图像二值化

(a) 包含干扰线的图



(b) 包含干扰线的细化图

图 2 图像细化

因此本文方法中考虑去除的干扰线具有以下特征: 单一方向 (横向或者竖向), 且大于一定长度. 这两种特征既符合干扰线的特性, 又极大地降低了检测干扰线的难度. 其中“单一方向”通过扫描图像时只遵循“从下往上, 从右到左”的原则来保障, 而“大于一定长度”通过线段的权值V_{li}大于阈值T_l保证, 即:

$$V_{li} > T_l \tag{2}$$

其中, 计算每条中心线的权值V_{li}的算法, 如下所述:

(1) 初始化, 设$V_{li} = 0\ (i = 1, \cdots, n)$, 其中$n$是图像中包含中心线的数目;

(2) 扫描前方的像素点, 并加上相应的权值;

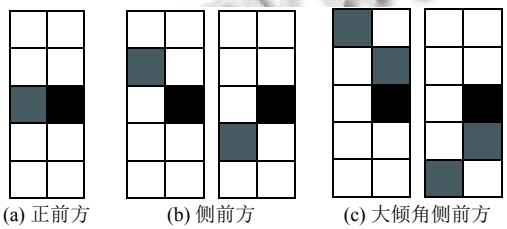
(3) 循环步骤 (2), 直到遍历了细化图像中每个像素点.

为了让T_l的设置具有自适应性, 其取值如下:

$$T_l = 3V_{lm} \tag{3}$$

其中, V_{lm}是所有V_{li}的中值. 这样就可以根据当前文档图像的情况, 获得阈值.

如图 3 所示, 其显示了上述步骤 (2) 中处理的 3 类情况, 黑色的为当前像素点, 灰色的是前方像素点, 这些都是中心线上的像素点. 其中, 当前像素点的前方像素点分 3 种情况: 图 3(a) 表示正前方有像素点, 图 3(b) 表示侧前方有像素点, 而图 3(c) 的倾斜的角度更大一些. 图 3(a) 情况的权值为 3, 图 3(b) 为 2, 图 3(c) 为 1. 对于一个位置, 只能属于这 3 种情况其中的一种, 并且优先属于权重较大的. 比如一个位置既满足图 3(a) 和图 3(b) 时, 它只属于图 3(a), 其他情况依次类推. 由于在一个位置只取了权值的最大值, 因此该提出的算法属于贪婪算法的一种.



(a) 正前方 (b) 侧前方 (c) 大倾角侧前方

图 3 当前像素的前方像素 3 类情况

另外, 上述算法过程中主要考虑的是横向的干扰线, 对于竖向的干扰线的检测, 则有两种方案:旋转模板或者旋转图像, 这两种方法都是可行的, 而斜向的干扰线在横向和竖向的检测过程中都有兼顾, 因此不需要单独列出来.

由此, 通过采用本节基于贪婪算法的干扰线检测算法, 可计算如图 2 所示的中心线权值, 得到干扰线的中心线Cg, 如图 4 所示. 去除掉干扰线的中心线Cg部分余下的则是文字中心线Cw, 下一节给出相应的方法.

图 4 检测出的干扰线部分

4 图像干扰线去除

通过上面的方法, 检测出干扰线的具体位置后, 需要将二值化图像中的干扰线去除, 只留下文字部分. 一般情况下, 某个前景像素到文字中心线Cw和干扰线中心线Cg哪个距离近, 就可以认为它属于距离近的那个部分. 即:

$$I(x,y) \in \begin{cases} I_f \text{ if } DCw(i,j) \le DCg(i,j) \\ I_b \text{ if } DCw(i,j) > DCg(i,j) \end{cases} \tag{4}$$

其中, D是像素$I(x,y)$到Cw或Cg的距离, I_b为干扰线像素集合, I_f为文字像素集合.

直接求解距离的计算量比较大, 因此可以使用以下方案 (这里以求到文字中心线Cw的距离为例):

(1) 初始化距离矩阵 (与图像大小相同) 中所有的位置为一个极大值max(实验中可取值 10 000);

(2) 设置Cw中所有的像素对应位置的距离为 0;

(3) 设置所有距离为max且与距离 0 相邻的位置距离为 1;

(4) 循环步骤 (3), 设置所有距离为max且与距离i相邻的位置距离为$i+1$.

通过本节图像干扰线去除算法的处理, 将如图 2 所示的内容去除干扰线之后的效果如图 5 所示, 可以看出, 本方法对文字内容的轮廓、文字线条的连续性等都具有较好的保留.

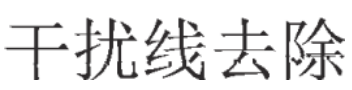


图 5 干扰线去除效果图

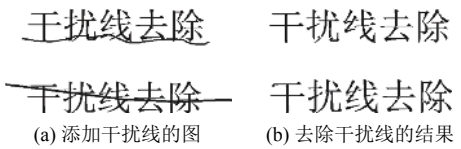
5 实验分析

本文所提的方法, 主要针对文档图像 OCR 处理之前的相关干扰线的去除预处理, 采用了原始文档图像二值化、细化获得中心线, 再采用贪婪算法的方式对

所有中心线中的相关干扰线进行检测和去除, 实验分别在人造文档图像和真实扫描文档图像上进行.

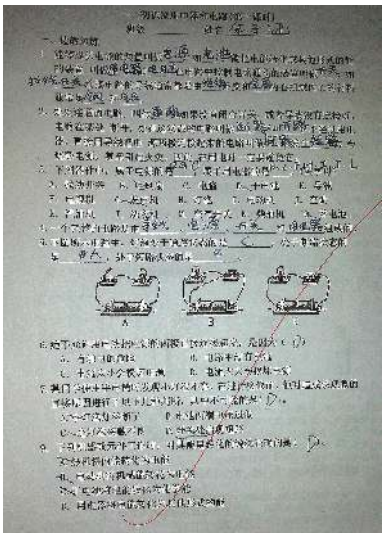
首先, 本文所提出的方法在一组人为制造的图像上进行测试, 人为制造的文档局部图像如图 6(a) 所示. 由于该人为制造出的图像中的干扰线与文字内容较为粘连耦合, 这增加了干扰线去除的精度, 对方法的要求较高, 因此能够考验本文方法在这种极端条件下的性能. 如图 6(b) 所示为本文方法去除干扰线效果, 其中, 在第一组中, 由于人为干扰线与文字的粘连耦合度很高, 故而造成干扰线去除时将个别文字的一些笔画也去除掉, 但文字上下文内容仍然保留精确, 后续的 OCR 处理仍然基本是有效的, 实际情况处理中, 这种极度粘连耦合的情况应该是极少的; 在第二组中, 人为干扰线与文字的粘连耦合度一般, 可以看出干扰线去除效果很好, 极大地保留了文字内容的完整性和可读性, 经进一步的清华紫光 OCR 软件测试能 100% 正确获取该局部图像的文字内容.

然后, 本文所提方法再在一组真实文档的扫描图像上进行实验测试. 原始文档图像如图 7(a) 至图 7(c) 所示, 这组图像是扫描现实文档获得的, 并希望通过 OCR 获取其中的文字. 这些图像中均包含了较多的干扰线, 如规则的长横线、不规则的划线等. 由于干扰线的判断具有一定主观性, 而本文干扰线去除方法的目的重点在于保留文档的文字内容, 以便 OCR 能以高精度获取文字内容, 因此本方法主要实现将规则的长横线、不规则的划线等理论上都识别成相关干扰线, 并根据自适应获取的阈值进行相关干扰线的判断检测, 进而予以去除, 结果如图 8 所示. 从图中可以看出, 本文所提方法对长度超过阈值的干扰线, 包括原文中的长横线、划线等都能有效去除, 较为清晰的保留了文字内容的完整性, 而一些短划线的残留分两种情况, 一种是短划线与文字的粘连耦合度高, 如与文字笔画重合, 影响了阈值判断, 另一种是本身其长度低于了阈值.

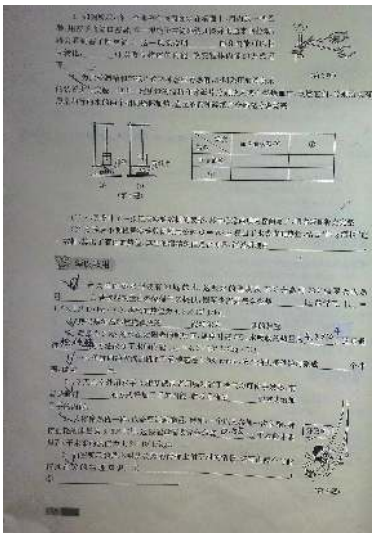


(a) 添加干扰线的图 (b) 去除干扰线的结果

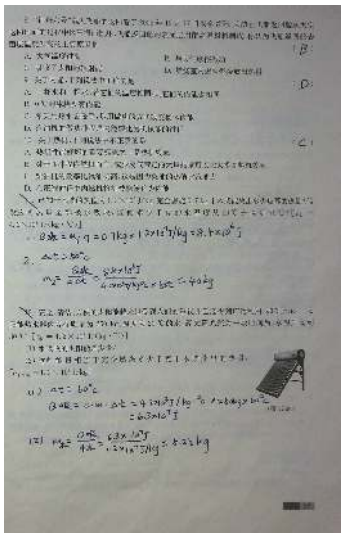
图 6 人造图像的实验结果



(a) 原始文档图1



(b) 原始文档图2

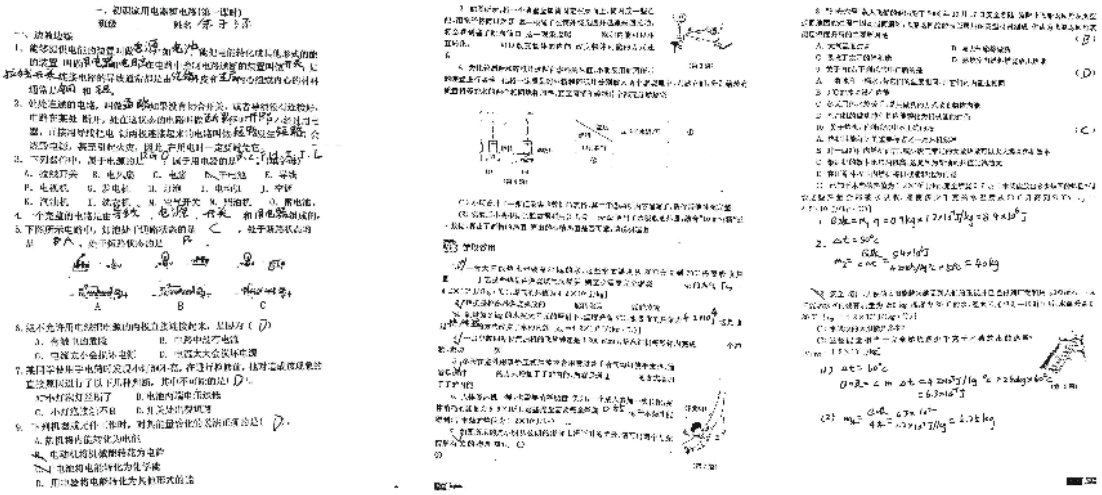


(c) 原始文档图3

图 7 真实扫描的文档图像

为了测试与类似算法的性能, 原始文档图像经本文方法去除相关干扰线之后, 经过清华紫光 OCR 软件测试, 如表 1 所示, 能正确获取文字内容的占比率大幅提高, 表明本文方法在去除原始文档干扰线预处理中是有效的. 另外与类似算法的比较中, 本文提出的算法, 虽然在正确率上没有绝对的优势, 但由于算法步骤少, 在速度上超越了原有的一些方法, 如表 2 所示.

通过上述实验测试, 本文提出的算法能够有效地去除文档图像中的相关干扰线且处理速度快, 特别是对于和文字粘连在一起的干扰线去除也能有较好的效果, 自适应长度阈值的处理方法使得本文方法可以针对各种各样的文档图像的相关干扰线进行检测和去除.



(a) 文档图1去除干扰线的结果 (b) 文档图2去除干扰线的结果 (c) 文档图3去除干扰线的结果

图 8 真实扫描图像去除干扰线的实验结果

表 1 与原始文档图的测试正确率 OCR 识别性能对比 (%)

方法	文档图 1	文档图 2	文档图 3
原始图像	78.9	83.5	81.7
文献[7]	94.1	95.9	93.6
文献[9]	96.2	97.4	97.7
本文方法	95.7	97.2	98.1

表 2 与文献[7]、文献[9]的速度对比 (毫秒)

方法	文档图 1	文档图 2	文档图 3
文献[7]	1137	1352	988
文献[9]	1263	1439	1127
本文方法	796	867	621

6 结论与展望

本文提出了一种干扰线去除的方法, 该方法先通过预处理得到每个文字或线条的中心线, 然后利用贪婪算法计算每条中心线的权值, 并认为权值大于阈值的中心线处存在干扰线, 最后结合形态学操作去除干扰线并尽可能保留文字. 提出的算法能较为有效地去除干扰线部分, 同时对规则的长横线、竖线以及不规则的划线也能一并消除, 从而降低了文档图像中的 OCR 处理的干扰因素, 经实验测试表明该方法是有效的. 另外, 由于采用的细化算法会产生毛刺, 从而影响最终的去除效果, 因此下一步工作中需要优化细化算法.

参考文献

1 Bai ZL, Huo Q. Underline detection and removal in a document image using multiple strategies. Proceedings of the 17th International Conference on Pattern Recognition. Cambridge, UK. 2004. 578–581.

2 Shi ZX, Setlur S, Govindaraju V. Removing rule-lines from binary handwritten arabic document images using directional local profile. Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey. 2010. 1916–1919.

3 Alipour F, Faez K, Seifzadeh S. Ruling lines removal in handwritten documents. Proceedings of the 8th Iranian Conference on Machine Vision and Image Processing (MVIP). Zanjan, Iran. 2013. 766–769.

4 Imtiaz S, Nagabhushan P, Gowda SD. Rule line detection and removal in handwritten text images. Proceedings of the 5th International Conference on Signal and Image Processing. Bangalore, India. 2014. 310–315.

5 Cheng ZG, Liu YC. Removal of interferential curve from text image. Proceedings of 2005 International Conference on Information Technology: Coding and Computing. Las Vegas, NV, USA. 2005. 154–159.

6 Kaur T, Mittal R. Hand-drawn annotation and underline detection and removal in scanned documents using artificial neural network & fuzzy C-means clustering. European Journal of Advances in Engineering and Technology, 2016, 3(1): 12–20.

7 Banerjee T, Biswas S. Hand-drawn line removal from Bangla printed document images. Proceedings of 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS). Bhubaneswar, India. 2015. 116–121.

8 Rehman A, Kurniawan F, Saba T. An automatic approach for line detection and removal without smash-up characters. The

Imaging Science Journal, 2011, 59(3): 177–182. [doi: 10.1179/136821910X12863758415649]

9 Pratihar S, Bhowmick P, Sural S, *et al*. Removal of hand-drawn annotation lines from document images by digital-geometric analysis and inpainting. Proceedings of the 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). Jodhpur, India. 2013. 1–4.

10 Das S, Banerjee P. Gabor filter based hand-drawn underline removal in printed documents. Proceedings of the 1st International Conference on Automation, Control, Energy and Systems (ACES). Hooghy, India. 2014. 1–4.

11 Calvo-Zaragoza J, Pertusa A, Oncina J. Staff-line detection and removal using a convolutional neural network. Machine Vision and Applications, 2017, 28(5-6): 665–674. [doi: 10.1007/s00138-017-0844-4]

12 Konwer A, Bhunia AK, Bhowmick A, *et al*. Staff line removal using generative adversarial networks. Proceedings of the 24th International Conference on Pattern Recognition (ICPR). Beijing, China. 2018. 1103–1108.