基于反馈合并的中英文混排版面 OCR 技术研究

任荣梓,高 航
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:迄今,光学字符识别(OCR)技术已普遍应用于社会生活的方方面面,单一字符集 OCR 技术领域已经取得重大突破。但由于中文和英文版面分析之间存在的明显差异,现有中英文混排 OCR 技术的表现均不尽如人意。针对传统 OCR 方法实现方式的缺点和不足,在研究中英文混合版面分析切分技术难点的基础上,提出了一种改进的基于反馈合并的中英文混合版面分析切分方法。该方法在综合应用 Canny 算子的图像二值化方法和中值滤波法进行滤波预处理的基础上,采用投影法两次分割字符区域,并对具体切分技巧进行了较为深入的研究。对比验证实验结果表明,所提出的版面分析切分方法可成功分离中英文混合文档中的中文、英文和数字字符,正确率比传统方法高出约 8 个百分点,可达到 97%,较好地解决了传统方法对粘连字符处理效果不佳的问题。

关键词:文字识别;中英混排;版面分析;分离
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)03-0039-05
doi:10.3969/j.issn.1673-629X.2017.03.008

Investigation on Layout Analysis Technology of Chinese and English Mixed OCR Based on Feedback Merging

REN Rong-zi, GAO Hang
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: So far, Optical Character Recognition (OCR) technology has been widely applied in all aspects of social life, and a single character set OCR has made a major breakthrough in the technology field. However, due to the obvious differences between Chinese and English layout analysis, the performance of the existing English and Chinese mixed OCR technology is not satisfactory. According to the shortcomings and deficiencies of traditional OCR method, on the basis of the analysis of the segmentation technique difficulties in the study of Chinese and English mixed layout, an improved segmentation method of Chinese and English mixed layout OCR analysis based on feedback merging is proposed. Based on the comprehensive utilization of the Canny operator image binary method and median filter method for filter preprocessing, this method segments the character region twice by projection method, and has conducted the thorough research to the specific segmentation techniques. Experiment results show that the proposed method can be successfully separated in mixed document in Chinese, English and numeric characters. The correct rate is higher than the traditional method about 8 percentage points, which can reach 97%, effectively solving the problem of ineffective adhesion character for the traditional methods.

Key words: character recognition; English and Chinese mixed; layout analysis; separation

0 引 言

近年来,关于 OCR(光学字符识别)技术的研究蓬勃发展,优秀的 OCR 算法更是层出不穷。例如,由南开大学机器智能研究所研究的英文 OCR 技术在 OCR 英文核心技术评测中获得世界第一,而由北京信息工程学院研究的中文 OCR 核心技术在 UNLV(美国内华达大学拉斯维加斯分校)的一次中文评测中获得最佳。其他比较著名的 OCR 技术包括 Tesseract-OCR、汉王等。

上述 OCR 技术虽然在各自单纯语种环境下表现优异,但是均不能保证对中文和英文及标点符号混排的图片进行有效识别。绝大部分针对中英文混合图片的现行 OCR 技术都是先采用版面分析技术,即先实行中英文的分割,再运用两种不同的算法分别进行识别,

收稿日期:2016-04-13 **修回日期**:2016-08-10 **网络出版时间**:2017-01-10
基金项目:江苏省科技成果转化专项资金(BA2012023)
作者简介:任荣梓(1993-),男,硕士研究生,研究方向为图像处理;高 航,副教授,硕士生导师,研究方向为图像处理、嵌入式应用。

由此可见版面分析过程就显得尤为重要。目前常用的版面分析算法分为三种:自顶向下法、自下而上法和综合法。

自顶向下法重视全局信息,从页面的整体入手,先利用图像处理的常用方法将文本图像划分成若干区域,再根据文本结构信息将第一次划分出来的区域进行二次划分。此类方法包括投影二分法[1]、循环 $X-Y$ 切分法[2]等,但该类方法对于信息内容复杂的版面分割精度并不理想。

与自顶向下法相反,自下而上法重视局部信息,其从图像像素开始,将图像由小区域逐步整合成大区域,最终覆盖整个文本图像。该方法弥补了自顶向下法存在的技术缺陷,包括游程码平滑切分法[3]、K_近邻聚类方法[4]、连通域提取算法切分[5]等,但缺点在于耗时较长。

综合法是文中采用的方法,既汲取了上述两种方法的优点,实现了全局信息与局部信息的融合,又较好地解决了两者存在的技术缺陷,在保证分割精度的前提下兼顾了时间的节省。有代表性的综合法包括基于背景间隔的版面切分算法[6]、基于复杂度的中文版面分析算法[7]等。

形近字是中文字符不同于英文等西方字符的独特之处。现代汉语常用的 3 500 个字符中形近字就不止 500 个,占总数的 14%。此类字符多为左右结构或上下结构,其部首或偏旁又是常见的汉字,给中文字符的识别造成了较大的麻烦。如“明、月”“汪、王”“由、甲”之类的字符通常极易被误混或割裂,严重影响了文本的正确识别。因此在版面分析的切分过程中,对于形近字的识别应充分重视。

1 处理流程

文中介绍的基于反馈合并算法的中英混合版面分析处理流程为:首先进行预处理,对输入的数字图片进行二值化和去噪,预处理完成后利用行分割和字符分割方式对图片进行区域分割,将其分割为中文区域以及英文和数字区域,之后分别采用相对应的方法对两种区域进行二次分割,然后利用评估系数对二次分割结果进行判别,属于粘连字符的情况下则对其再次进行分割,直至检测不到粘连字符时,分割完毕。流程如图 1 所示。

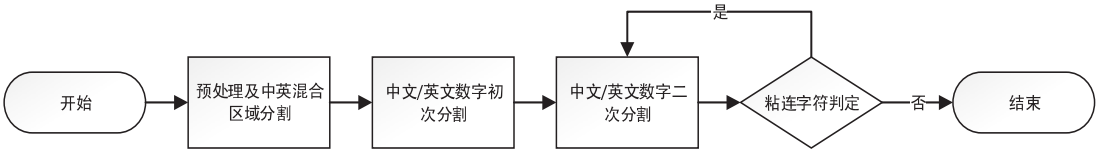

图 1 处理流程图

2 预处理

2.1 二值化

对原始图像进行预处理,包括将图像进行常规初始化操作即二值化[8]和降噪处理。所谓图像二值化即把图像上所有像素点的值进行分化:设置为 0 或 1。二值化的图像具有非常明显的视觉效果即非黑即白,在 OCR 处理中具有极其重要的作用。通常进行二值化的方法是全局二值化阈值法:即设定一个阈值 T,大于等于 T 的所有像素置为 1,小于 T 的像素置为 0。所以选取合适的阈值 T 是关键。目前比较先进的二值化方法有结合 Canny 算子的图像二值化[9]等。文中采用这种二值化方式。

2.2 图像去噪

在二值化完成之后,虽然文本图片已被分割为包含文本信息的前景图片和不包含文本信息的背景图片,可是在前景信息中仍然具有一些零星的噪声点,如果此时不对其进行消除,则对 OCR 后期操作的破坏性影响是很大的。噪声一般分为加性噪声、乘性噪声和量化噪声三种。其中,加性噪声主要是由于摄像机在扫描图像过程中产生的,与信号本身无关。乘性噪声则是图像信号本身所附带的,例如影视图像中产生的雪花点等等。而图像量化中产生的量化误差导致的噪声则称为量化噪声。

目前对去噪方法而言,常用的主要有均值滤波、自适应维纳滤波和中值滤波[10]等。其中,均值滤波主要采用相邻区域像素平均值的均值滤波器,这种方法对于加性噪声的清除效果较为显著,但缺点也十分明显:即因为平均而容易导致图像局部模糊。而自适应维纳滤波是根据图像的局部方差来调整滤波器的输出,克服了图像模糊的问题,但是缺点在于其计算量过大。中值滤波是采用一种较为简单的非线性平滑滤波器,它根据噪声往往都是孤立的特性,把图像中一点的值用其附近有效区域的个点值的中值替代,从而使周围像素差别较大的点得到平均,以此来消除噪声点。因为中值滤波法性能的优异和操作的简便,文中在去噪处理中使用了中值滤波。

3 中英混合区域分割

预处理完成后,利用行分割和字符分割方式对混合区域进行区域分割,之后再分别利用两种不同的方法对确定块中的中文和英文数字块进行分割,直到完

成所有字符的分割工作。对于其中的中文字符块,先行判断它是否是粘连字符,如果是,则对其进行字符再分割。当不再能检测到粘连字符时则证明分割完成。

3.1　行分割

正常的文本图片行与行之间的空格间距是固定的,通常情况下也会小于单行文本的字符高度。因为行与行之间空白的存在,通过检测空白区域,就可以利用它确定一行的首末。可以使用一个固定的比较大的阈值来帮助确定,而这个阈值通常情况下可以使用比二倍于一个字符的宽度略大。对于一个正常文本文件而言,极少有大于两个字符宽度的空格,即使有,因为当一个文本出现大于两个字符宽度的时候,文意已经产生了变化,按照分开行的做法也并不会产生错误。当一行空白的区域高度小于这个阈值时,可以断定它是一个空白。当黑色区域大于某一个阈值时,可以认为它是一行的开始,当黑色区域小于阈值时,可以判断其为一行的结尾。按照这种方法可以把文本按行分割完毕,之后再对字符分割也就更加便捷。

图 2 展示了多行文字的水平投影。

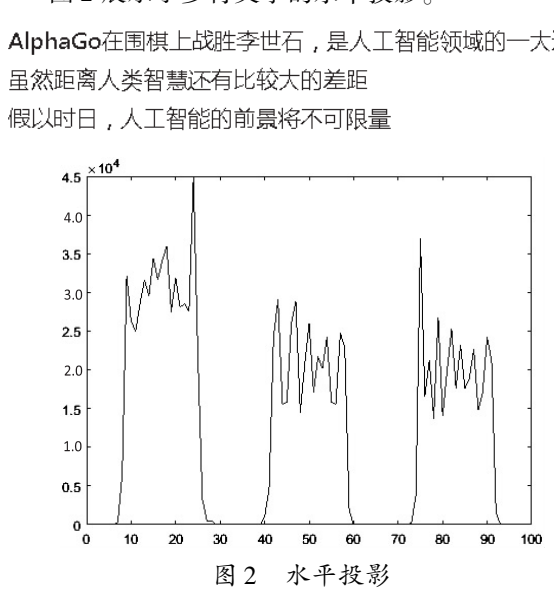

图 2　水平投影

3.2　字符分割

行分割完成之后,就可以进行字符分割。除了行与行之间,同一行以内的字符间也是存在些许空白的,可以利用这些空白把字符分割出来。垂直投影法就是一种不错的方法。把数字图像具体化为一个 $M \times N$ 的矩阵 $g(i,j)$,每一列的垂直投影为:

$$V(j) = \sum_{i=0}^{N} g(i,j), j = 1,2,\cdots,N \tag{1}$$

其中,投影值为 0 的点是字符间的空白。从第一个不为 0 的点 J_a 开始,分割程序从左至右扫描每一行文本,当遇到 $V(j) = 0$ 的点 J_b 时停止,两点之间视为一个字符。使用这种方法循环至一行的结尾。图 3 展示了单行文字的垂直投影。

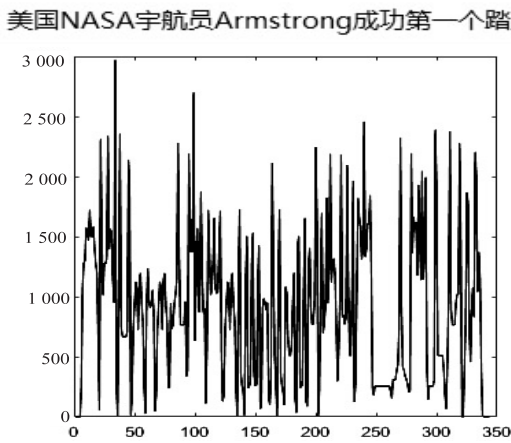

图 3　垂直投影

3.3　区域分割

把中文字符的中心投影映射到同一水平线上,可以发现中文字符的水平间隔是均匀且固定的,而且根据反复观察得知该特性并不受字体或者样式的改变影响,例如华文仿宋和加粗等。这意味着可以用水平间隔的恒定性来判断一些字符是否为中文。与之类似,英文字符也具有类似的性质,只是每隔一段时间就会出现大的跨越;但是当中文、英文以及数字混合出现在一行的时候,间隔将会变得混乱。根据这种差异性可以由此来进行中英区域的分割。

因为英文字符和数字字符有类似的宽度和周期,并且在一行之中都会有一些固定的单词之间的空格。英文以及数字字符的长和宽远远小于中文字符。所以,可以利用监测字符的长和宽来分离出中文字符区域。之后在剩下的英文区域和数字区域中,依据相同的方法,鉴于标点符号的长宽比相对来说要小很多,可以依此迅速分离出标点符号。由此,中文和英文数字得以分离开来。图 4 展示了分离结果。

图 4　区域分离结果

4　中英字符二次分割

4.1　中文字符二次分割

相关文献中提出了很多分离中文的算法,例如基于可见性的中文字符分离[11]、基于中文笔画结合的手写中文字符分离[12]、基于多信息融合的中文字符分离[13]、基于单元合并的中文字符分离[14]以及基于反馈的中文字符分割算法[15]等等。以上方法在一定程度上可以较好地分离中文字符,但是对部分较为生僻且结构特殊的中文字符都会出现不同程度的错误,而且单纯经过分割算法,虽然绝大部分中文字符可以被分离出来,但是在已经分理出的字符中仍有一部分粘连字符,例如常见的“日”和“月”就容易粘连为“明”。

为保证最终的识别结果正确,需要对已分离出的结果进行粘连字符的检测和二次分离。所以文中将已有的算法进行综合改进,提出了一种反馈合并算法用来分离中文字符。具体过程如下:

(1)设立评估系数。

(L,U),(R,D) 表示 i 的位置,其中 (L,U) 和 (R,D) 分别是该单元左上角坐标和该单元右下角坐标;

待评估字符的宽度 P_w、高度 P_h、行间距 P_l 和所占空间 P_s;

$(a_1,a_2,a_3,a_4,a_5,a_6)$ 是根据先验知识确定的系数;

W_i 和 H_i 代表正在合并的某个特定单元的字符的宽和高,H_{ij} 表示合并后的高度;

M 表示总单元个数,N 表示剩余单元个数。

$$\alpha = \frac{N}{M}\left|\min(R_i - R_j) - \max(L_i,L_j)\right| \tag{2}$$

$$\beta = \frac{N}{M}\left|\min(D_i - D_j) - \max(U_i,U_j)\right| \tag{3}$$

(2)设立一个评估标志位 Flag 并将其置为 FALSE。

对任意单元 i,遍历单元集合,当存在单元 j 满足下面所述条件时,则将单元 i 和单元 j 合并为一个字符。

$$\begin{cases} a_1 * W_i \geqslant H_{ij} \\ a_1 * H_i \geqslant H_{ij} \\ c_1 * P_l \geqslant P_h \\ c_2 * P_s \geqslant P_w \end{cases} \tag{4}$$

(3)记录在第二步中进行合并之后的单元的合并信息,并将它们的标志位 Flag 置为 TRUE,当有一个字符通过评估时,将合并为它的字符记为“通过单元”。全部结束之后将会有一部分单元被保留下来而没有被合并,此时就需要对剩余未标记为“通过单元”的所有单元使用第二次反馈因子进行二次合并。即当剩余非“通过单元”满足以下条件时,对其进行合并。

$$\begin{cases} c_4 * w > \alpha > c_1 * h \\ c_5 * \alpha < \beta \leqslant c_6 * \alpha \\ \beta \geqslant c_3 * \min((D_i - U_i),(D_j - U_j)) * P_h \\ c_2 * P_1 \geqslant P_h \\ c_4 * P_h > \beta > c_1 * P_h \end{cases} \tag{5}$$

(4)检测粘连字符:设置一个 W_a 变量来表示平均字符的宽度,令 W_i 表示待测字符的宽度,T 是用先验知识设置的阈值。如果 $|W_a - W_i > T|$,则该字符 P_i 为一个粘连字符。T 的设置通常与上文中单一单元的宽度相近。

(5)分离粘连字符:上一步寻找到粘连字符之后,需要首先确定一个正确的分割点。对于中文字符的粘连字符,它的宽度可以根据所有中文字符的平均高度来确定,因为中文字符独特的矩形结构,它们的高宽比在 1.05 ~ 1.15 之间,因此可以利用投影法来确定字符的边界:即从左至右扫描每一行以确定全部的粘连字符的区域并标识出全部的分割点,分割完成。

4.2 英文和数字的二次分割

对英文和数字字符的再分离可以利用字符图像背景的上下凹区域进行再切分[16]。通过计算图像的背景域,提取出上下凹区域,再采用相邻匹配原则和最小面积选择原则确定切分域,从而提取出切分线进行切分。文中采用该方法进行英文和数字的二次分割,达到了较为理想的效果。

5 中英文及数字混合字符分割实验结果

实验选取了包含中英混合文字的报纸、书刊、网页快照作为测试文件,字体主要是由宋体标准和加黑组成,英文为主要正常字体包含部分斜体,字号大小均有差异,扫描分辨率为 300 ~ 400 dpi。将混合材料分成三组,数量分别控制在 1 500、3 000 和 5 000 数量级(因为材料本身限制会有一些浮动),三组材料中文所占比例大致均衡,约为 44%(668)、56%(1 650)、48%(2 502)。结果见表 1。

表 1 实验结果

方法	待测字符数量	错误数量	正确率/%
传统单元合并方法	1 521	187	87.7
	2 948	344	88.3
	5 214	601	88.4
传统反馈评估方法	1 521	97	93.6
	2 948	203	93.1
	5 214	285	94.5
反馈合并方法	1 521	53	96.5
	2 948	112	96.2
	5 214	149	97.1

表 1 记录了测试样本的数量及错误数量,其中错误数量按照原本正确的材料中的字符中未出现的来计数,即两个字符粘连成一个字符记为两次错误,而一个字切分成两个的情况则记为一次错误。最终的结果表明,文中方式对字符的切割效率较好,比使用传统单元合并的版面分析法提高约 8%,比使用传统反馈评估的方法提高约 3%。

6 结束语

针对传统 OCR 方法实现方式的缺点和不足,在研究中英文混合版面分析切分方法技术难点的基础上,

提出了一种改进的基于反馈合并的中英文混合版面分析切分方法。该方法在综合应用 Canny 算子的图像二值化方法和中值滤波法进行滤波预处理的基础上,采用投影法两次分割字符区域,并对具体切分技巧进行了较为深入的研究。对比验证实验结果表明,所提出的版面分析切分方法可成功分离中英文混合文档中的中文、英文和数字字符,且具有普遍高于传统方法的正确率,较好地解决了传统方法对粘连字符处理效果不佳的问题。

参考文献:

[1] 王　丹,刘　江. 基于投影直方图的文档图像快速匹配研究[J]. 计算机技术与发展,2011,21(7):129-131.

[2] Mao S,Kanungo T. Empircal perforinanoce evaluation of page segmentation algorithms[C]//Proceeding of SPIE conference on document recognition and retrieval. [s. l.]:[s. n.],2000:303-312.

[3] 张　利,朱　颖,吴国威. 基于游程平滑算法的英文版面分割[J]. 电子学报,1999,27(7):102-104.

[4] 周国兵,吴建鑫,周　嵩. 一种基于近邻表示的聚类方法[J]. 软件学报,2015,26(11):2847-2855.

[5] 陈　艳,孙羽菲,张玉志. 基于连通域的汉字切分技术研究[J]. 计算机应用研究,2005,22(6):246-248.

[6] 杨　宁. 基于背景间隔的中文版面分析系统[D]. 南京:南京理工大学,2002.

[7] 范玉凤. 基于复杂度的自适应中文版面分析方法研究[D]. 青岛:中国海洋大学,2011.

[8] Zhang Y L,Zhang S C. Image rotation and binaryzation based on . Net[C]//7th international conference on electronic measurement and instruments. Beijing:[s. n.],2005:406-408.

[9] 陈　强,朱立新,夏德深. 结合 Canny 算子的图像二值化[J]. 计算机辅助设计与图形学学报,2005,17(6):1302-1306.

[10] 张　恒,雷志辉,丁晓华. 一种改进的中值滤波算法[J]. 中国图象图形学报,2004,9(4):408-411.

[11] Xu Liang,Yin Fei,Wang Qiufeng,et al. Touching character separation in Chinese handwriting using visibility-based foreground analysis[C]//11th international conference on document analysis and recognition. Los Alamitos,CA,USA:IEEE Computer Society,2011:859-863.

[12] 赵姝岩,郭　捷,施鹏飞. 基于笔画分析和背景细化的粘连手写汉字切分[J]. 上海交通大学学报,2003,37(9):1434-1437.

[13] 付　强,丁晓青,蒋　焰. 基于多信息融合的中文手写地址字符串切分与识别[J]. 电子与信息学报,2008,30(12):2916-2920.

[14] Liu Mingzhu,Suo Yuxiu,Ding Yinan. Research on optimization segmentation algorithm for Chinese/English mixed character image in OCR[C]//4th international conference on instrumentation and measurement,computer,communication and control. New York,NY,USA:IEEE,2014:764-769.

[15] 安艳辉,董五洲. 基于识别反馈的粘连字符切分方法研究[J]. 河北省科学院学报,2008,25(2):32-35.

[16] 罗　佳. 一种对粘连英文字符串的快速切分算法研究[J]. 计算机技术与发展,2014,24(8):59-62.

(上接第 38 页)

[4] 孙延奎. 小波分析及其应用[M]. 北京:机械工业出版社,2005.

[5] 衡　彤. 小波分析及其应用研究[D]. 成都:四川大学,2003.

[6] Mallat S G. Multiresolution representation and wavelets[D]. Philadelphia,PA:University of Pennsylvania,1988.

[7] Mallat S G. Multiresolution approximations and wavelet orthonormal bases of L2(R)[J]. Transactions of the American Mathematical Society,1989,315(1):69-87.

[8] Graves A,Mohamed A R,Hinton G. Speech recognition with deep recurrent neural networks[C]//International conference on acoustics,speech and signal processing. [s. l.]:IEEE,2013:6645-6649.

[9] Hermans M,Schrauwen B. Training and analysing deep recurrent neural networks[C]//Advances in neural information processing systems. [s. l.]:[s. n.],2013:190-198.

[10] Pascanu R,Gulcehre C,Cho K,et al. How to construct deep recurrent neural networks[C]//Proceedings of the 2014 international conference on learning representations. [s. l.]:[s. n.],2014.

[11] 桑燕芳,王　栋,吴吉春,等. 水文时间序列小波互相关分析方法[J]. 水利学报,2010(11):1272-1279.

[12] 左其亭,高　峰. 水文时间序列周期叠加预测模型及 3 种改进模型[J]. 郑州大学学报:工学版,2004,25(4):67-73.

[13] 朱跃龙,李士进,范青松,等. 基于小波神经网络的水文时间序列预测[J]. 山东大学学报:工学版,2011,41(4):119-124.

[14] 余宇峰,万定生. Benford 法则在水文数据质量挖掘中的应用研究[J]. 微电子学与计算机,2011,28(8):180-183.