DOI:10.13878/j.cnki.jnuist.2021.03.012

孙瑞彬[1]　钱夔[1]　徐伟敏[2]　路红[1]

基于 Tesseract-OCR 的复杂发票自适应识别

摘要

针对复杂发票任意区域下的特定表格内容提取与实时识别问题,提出了一种基于 Tesseract-OCR 引擎的自适应识别方法.首先利用 OpenCV 对发票图像进行预处理滤波、自适应阈值等一系列预处理得到二值图像;然后利用形态学中的开运算提取表格全域线段,进行表格位置提取,并结合表格交点坐标与自定义模板,实现表头与内容自适应适配;最后利用 jTessBoxEditor 对表格区域内容进行字库训练优化,最终实现基于 Tesseract-OCR 的字符识别.实验结果表明该方法具有高准确识别率,支持感兴趣区域自适应识别,具备高可用性.

关键词

发票识别;Tesseract-OCR;OpenCV;字库训练;自适应识别

中图分类号 TP391

文献标志码 A

收稿日期 2021-03-18

资助项目 南京工程学院引进人才科研启动基金(YKJ201918);南京工程学院校级科研基金(CXY201930)

作者简介

孙瑞彬,男,硕士生,研究方向为基于深度学习的机器视觉.s781336445@163.com

钱夔(通信作者),高级工程师,研究方向为机器人与人工智能.kuiqian@njit.edu.cn

1　南京工程学院 自动化学院,南京,211167

2　南京学府睿捷信息科技有限公司,南京,210009

0　引言

随着信息的快速发展,数字时代已然来临.OCR (Optical Character Recognition) 技术即光学字符识别作为计算机视觉领域的一个重要分支,通过具有拍照功能的设备获取文档图片,再利用诸多算法对文档内容进行分析识别.发票识别作为 OCR 领域的重要研究方向,可有效解决票据信息人工录入时效率低、准确率低的问题,提升企业的办事效率.

票据类 OCR 技术的研究吸引了众多学者,产生了许多研究成果,比如:王阳等[1]基于深度学习的 OCR 文字识别方法,解决了银行业对于海量图像处理效率低下的问题;郭剑雄等[2]的英文字符算法研究,有效解决了英文字符识别不准确的问题;刘森等[3]对 Android 图文同步识别系统的研究,改善了 Tesseract-OCR 引擎对模糊图像识别效果不佳的问题.但现有的技术对复杂票据的版面分析并无较好的处理方法,对内容涵盖多种字符的文档图像也做不到高效的精准识别[4].

本文基于 Tesseract-OCR 引擎给出了一套自适应识别方法.利用 OpenCV 函数库对图像进行滤波[5],阈值化处理得到二值图像,然后进行开运算提取发票表格以改善票据版面复杂时难以识别的问题;结合表格交点坐标与自定义模板,完成表头匹配后,再通过 jTessBoxEditor 有针对性地训练字库,优化图文中涵盖中英文、数字及各种符号的识别准确率.最终实现对该类含复杂版面、多语言文本的发票图像感兴趣区域的自适应识别.

1　识别架构

复杂发票的 OCR 识别主要包括图像预处理、表格提取、内容匹配、字符识别 4 个模块,其识别架构如图 1 所示.

1.1　表格提取

预处理即对发票原图进行简单的形态学操作.采用自适应阈值化得到二值图像时,为了获取最佳的二值图像,要求出最佳阈值[6].假设两个高斯分布[7]为

$$F_1(x)=\frac{1}{\sqrt{2}\sigma_1^2}\left(-\frac{(-\mu_1)^2}{2\sigma_1^2}\right), \tag{1}$$

$$F_2(x)=\frac{1}{\sqrt{2}\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}\right), \tag{2}$$

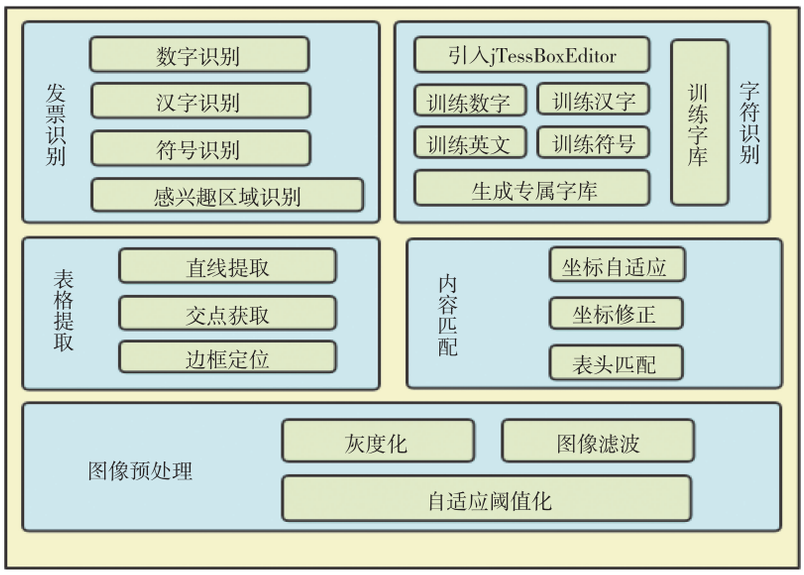


图 1　OCR 总体架构

Fig. 1　General architecture of OCR for invoice recognition

其中 σ_1,σ_2 和 μ_1,μ_2 分别是两个高斯分布的平均值和方差,且假设 $\mu_1 < \mu_2$,最佳阈值 x 需满足:

$$F_1(x) = F_2(x)\ . \tag{3}$$

结合式(1)—(3),得到最佳阈值关于 x 的方程组:

$$\begin{aligned} &ax^2 + bx + c = 0,\\ &a = \sigma_2^2 - \sigma_1^2,\\ &b = 2(\mu_2\sigma_1^2 - \mu_1\sigma_2^2),\\ &c = 2\sigma_1^2\sigma_2^2\ln(\sigma_1/\sigma_2) + \sigma_2^2\mu_1^2 - \sigma_1^2\mu_2^2, \end{aligned} \tag{4}$$

其中 a,b,c 为常数.求解该二次方程,取决于 μ_1 和 μ_2 之间的解即图像的最佳全局阈值 T.由于程序开始阈值和高斯分布的参数均未知,所以运行过程中要给定一个初始阈值去估计高斯分布的参数,再利用高斯分布更新阈值,以此反复直到收敛便可求得全局最佳阈值 T.

接下来,对二值图像进行线段识别,以实现表格提取.表格由水平线和垂直线组成,因此需分别在两个方向上对发票进行线段提取,提取线段的形态学操作就是通过自定义的结构元素,构造对指定形状敏感的形态学运算,再通过膨胀和腐蚀操作处理敏感像素.以提取水平线为例:创建自定义内核形态为竖向矩形,此时的敏感对象是垂直线段,通过开运算腐蚀垂直方向像素,水平线即被保留.提取垂直线段时,只需把内核形态定义为横向矩形.图像所有线段均提取后,对输出结果进行“与”操作以求得交点坐标,发票内容需通过坐标对进行匹配.再对提取出的水平线图、垂直线图做加法合并,即可得到完整的表格框线图.

1.2　内容匹配

发票内容为多行多列文本,整体识别效果较差,因此对发票先分割再识别[8],把含有用信息的表格单独切割,每个表格都是一张图像,对于含多行文本的表格,通过算法对其进行再分割,使得到的每张图像都只含一列文本.提取感兴趣区域的公式如下:

$$r_1 = s_1[y_1:y_2,x_1:x_2]\ , \tag{5}$$

其中,r_1 是待识别区域,s_1 为目标图像,x,y 分别为图像的横纵坐标.在交点坐标已知的条件下,将内容与表头进行匹配,再根据字符宽度进行修正,使文本内容与边框分离.坐标 x,y 并非固定数值,而是相对位置[9].

图 2 是两行两列的表格,已标明横、纵坐标.其中 (x_1,y_1) 并非具体数值如(1,2)、(2,3),x_1 代表第 1 个横坐标,x_2 代表第 2 个横坐标,纵坐标也是同理.取得坐标后将其有序排列,无论图像的位置或大小如何改变,(x_1,y_1)、(x_2,y_2) 两点代表的总是图中左上角的表格,(x_2,y_2)、(x_3,y_3) 代表的总是右下角的表格.由于绝对位置会随图像大小、位置改变而改变,因此本文采用相对位置,实现表头与内容自适应匹配[10].

1.3　字符识别

发票文档由中英文、数字和特殊符号共同组成,Tesseract-OCR 引擎自带的字库识别准确率并不高,

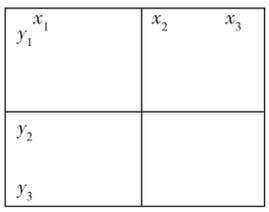


图 2　发票表格式样

Fig. 2　Sample invoice form

因此引入 jTessBoxEditor 来训练专门针对发票识别的字库[11].首先,通过 jTessBoxEditor 将所有要训练的发票图片合并成一个 tif 文件,命名为 name.tif,文件名任意命名即可.系统路径导入到该 name.tif 文件所在目录后,训练步骤如表 1 所示.

2　实验结果分析

作为本文实验对象的复杂发票含大小不一的表格 110 余项,部分表格内含多列文本,本方法将每一列信息都切割为一张图像,因此有用信息共 146 项,发票原图如图 3 所示.识别前首先完成发票表格提取,提取出的表格图像如图 4 所示.

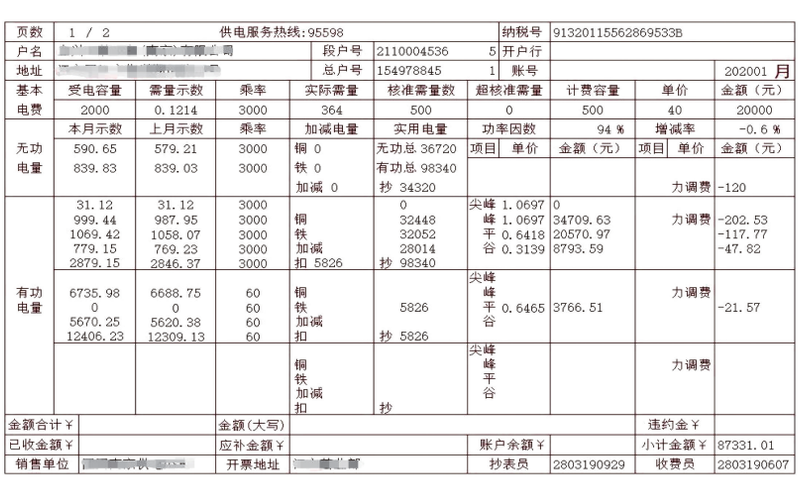
其次进行内容匹配,通过修正坐标,将内容与表格边框分隔开,使表头与内容精准匹配.通过 jTessBoxEditor 软件训练字库,对生成的 100 多张 box 文件逐一修改,其中背景模糊的数字图像,box 文件会出现无法检测、识别错误的问题,如图 5a,使用原生的 jTessBoxEditor 软件只出现了 5 个识别边框,正确识别的数据仅 2 个;优化后,8 个数据均被边框检测到且准确识别.图 5b 展示了训练汉字优化前后的对比效果[11].

本方法与腾讯云 OCR、百度云 OCR 以及原生的 Tesseract-OCR 引擎展开对比,对多张发票的实验数据进行分析,将所得结果列于表 2.

表 1　训练步骤

Table 1　Training steps

步骤	方法描述	具体操作/指令	预期结果
S1	生成最小化识别边框	tesseract name.tif name batch.nochop makebox	得到 box 文件
S2	校正识别边框数据	打开 Box Editor,调整识别边框的位置数据	得到新的 box 文件
S3	生成特征训练文件	tesseract name.tif name nobatch box.train	得到 tr 文件
S4	生成字符集文件	unicharset_extractor name.box	得到 unicharset 文件
S5	创建字符特征文件	输入 name 0 0 0 0 0 后保存	name 为字库文件名,5 个 0 表示子集中无粗体斜体等特殊字体
S6	生成数据字典	mftraining-F-file-U unicharset name.tr	得到 shapetable inttemp pffmtable 3 个文件,file 为 S5 创建的文件名
S7	生成符号字典	cntraining name.tr	得到 norproto 文件
S8	合并数据文件,生成字典	combine_tessdata name.	得到 name.traineddata 文件
S9	保存字典文件	将 name.traineddata 文件复制到 tessdata 文件目录下	训练完成



页数	1 / 2	供电服务热线:95598			纳税号	91320115562869533B		
户名	[redacted]		段户号	2110004536 5	开户行			
地址	[redacted]		总户号	154978845 1	账号			202001 月
基本电费	受电容量	需量示数	乘率	实际需量	核准需量数	超核准需量	计费容量	单价	金额(元)
	2000	0.1214	3000	364	500	0	500	40	20000
	本月示数	上月示数	乘率	加减电量	实用电量	功率因数	94 %	增减率	-0.6 %
无功电量	590.65	579.21	3000	铜 0	无功总 36720	项目 单价	金额(元)	项目 单价	金额(元)
	839.83	839.03	3000	铁 0	有功总 98340				
				加减 0	抄 34320			力调费	-120
	31.12	31.12	3000		0	尖 1.0697	0		
	999.44	987.95	3000	铜	32448	峰 1.0697	34709.63	力调费	-202.53
	1069.42	1058.07	3000	铁	32052	平 0.6418	20570.97		-117.77
	779.15	769.23	3000	加减	28014	谷 0.3139	8793.59		-47.82
	2879.15	2846.37	3000	扣 5826	抄 98340				
有功电量	6735.98	6688.75	60	铜		尖			
	0	0	60	铁	5826	峰		力调费	
	5670.25	5620.38	60	加减		平 0.6465	3766.51		-21.57
	12406.23	12309.13	60	扣	抄 5826	谷			
				铜		尖			
				铁		峰			
				加减		平		力调费	
				扣	抄	谷			
金额合计￥		金额(大写)						违约金￥	
已收金额￥		应补金额￥			账户余额￥			小计金额￥	87331.01
销售单位	[redacted]	开票地址	[redacted]		抄表员	2803190929		收费员	2803190607

图 3　发票原图

Fig. 3　Original invoice

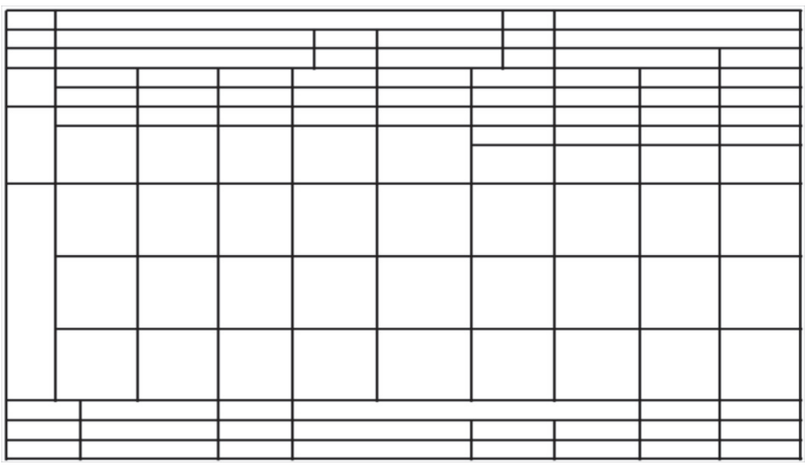


图 4　表格提取图

Fig. 4　Table extracted from invoice of Fig. 3

训练前

	Char	X	Y	Width	Hei...
1	i	8	13	45	33
2	g	62	13	24	33
3	o	87	13	64	34
4	4	164	14	20	33
5	3	189	14	24	33

训练后

	Char	X	Y	Width	Hei...
1	-	8	13	22	33
2	1	34	13	22	33
3	2	62	13	21	33
4	8	87	13	23	34
5	2	114	13	21	34
6	.	138	35	12	12
7	4	164	14	20	33
8	3	189	14	24	33

a. 数字训练对比

训练前

	Char	X	Y	Width	Hei...
1	#	7	7	34	37
2	-	45	23	35	21
3	~	86	11	30	1
4	E	11	48	16	0
5	I	31	48	3	0
6	~	86	9	3	35
7	~	113	9	3	35
8	~	86	39	30	3
9	F	122	9	34	35

训练后

	Char	X	Y	Width	Hei...
1	科	7	7	34	37
2	学	46	7	34	37
3	园	84	7	34	37
4	天	122	7	34	37
5	元	161	7	34	37
6	东	199	7	34	37
7	路	236	7	34	37
8	1	275	7	13	37
9	8	295	7	16	37

b. 汉字训练对比

图 5　字库训练对比效果

Fig. 5　Comparison of character training effect on number (a) and Chinese character (b) recognition

表 2 实验结果对比

Table 2 Performance comparison of the proposed method and other OCR technologies

OCR 技术	汉字准确率/%	数字准确率/%	符号准确率/%	所需时间/s	支持感兴趣区域自适应识别
Tesseract-OCR	28	61	46	约 15	不支持
腾讯云 OCR	100	92	100	约 4	不支持
百度云 OCR	100	96	94	约 2	不支持
本实验	92	100	94	约 1	支持

分析表 2 可知,改进后的 Tesseract-OCR 引擎对该类复杂发票的识别性能已有了很大提升,具体表现在以下方面:

1)识别时间上,本方法识别一张 1 020×770 像素的发票图像,用时约 1 s,而百度云 OCR 用时 2 s,腾讯云 OCR 用时 4 s.

2)识别准确率方面:本方法对发票的关键信息即各类数字数据的识别准确率可达 100%,其余两款 OCR 引擎尚未做到;对于汉字和符号的识别准确率,由表 2 可看出,相比较原生的 Tesseract-OCR 引擎,经本方法优化后的准确率已有大幅提升,但由于可训练的汉字对象较少,因此仍存在少量的识别错误.

3)自适应性方面,本方法对图像采用先分割后识别,使表头与内容自适应适配,从而实现任意区域下对特定表格进行内容提取,并高效精准识别.而百度云 OCR 与腾讯云 OCR 都是对发票做整体识别,并不具备自适应性.

因此,本方法能够对发票的特定区域实现高效精准的自适应识别,具有良好的工程可用性.

3 结语

本文以复杂发票为对象,针对其任意区域下的特定表格内容提取与实时识别问题,提出了一种基于 Tesseract-OCR 开源引擎的识别方法.利用 Python 的第三方库 OpenCV 对发票原图进行形态学处理以得到二值图像.在此基础上完成表格位置提取,并结合交点坐标与自定义模板完成了表头与内容适配.再通过 jTessBoxEditor 对所有发票模板的表格内容进行字库训练,使识别更具针对性.实验结果表明,本方法能够对发票的感兴趣区域实现精准高效的自适应识别.下一个研究方向为污迹发票实时识别,以期实现模糊不规则表格内容识别.

参考文献

References

[1] 王阳,李振东,杨观赐.基于深度学习的 OCR 文字识别在银行业的应用研究[J].计算机应用研究,2020,37(增 2):375-379
WANG Yang,LI Zhendong,YANG Guanci.Research and application of OCR based on deep learning in banks[J].Application Research of Computers,2020,37(sup2):375-379

[2] 郭剑雄,杨力华.一种基于衬线去除的英文印刷体多字号字符分割算法[J].模式识别与人工智能,2006,19(6):702-707
GUO Jianxiong,YANG Lihua.Approach to segment multi-size machine printed characters by removing serifs[J].Pattern Recognition and Artificial Intelligence,2006,19(6):702-707

[3] 刘淼,杨镇豪,谢韵玲,等.Android 图文同步识别系统的设计和实现[J].计算机工程与设计,2014,35(6):2207-2213
LIU Miao,YANG Zhenhao,XIE Yunling,et al.Android synchronized-OCR system design and realization[J].Computer Engineering and Design,2014,35(6):2207-2213

[4] 谢亚琴,王超.基于不同邻近标签数选择的 LANDMARC 定位算法研究[J].南京信息工程大学学报(自然科学版),2019,11(5):621-624
XIE Yaqin,WANG Chao.An improved LANDMARC method based on different neighbor reference nodes[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2019,11(5):621-624

[5] 李毅荣,郭磊,张漫杨.基于 Tesseract-OCR 的快递单中手机号码识别应用的实现[J].电子测试,2018(22):23-28
LI Yirong,GUO Lei,ZHANG Manyang.Implementation of telephone number recognition in express list based on Tesseract-OCR[J].Electronic Test,2018(22):23-28

[6] 张永宏,吴鑫.BP 神经网络在图像字符识别中的改进和应用[J].南京信息工程大学学报(自然科学版),2012,4(6):526-529
ZHANG Yonghong,WU Xin.Improvement and application of BP neural network in image character recognition[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2012,4(6):526-529

[7] 申彤,庄建军,黎文斯,等.基于 HOG 特征提取和支持向量机的东巴文识别[J].南京大学学报(自然科学),2020,56(6):870-876
SHEN Tong,ZHUANG Jianjun,LI Wensi,et al.Research on recognition of Dongba script by a combination of HOG

feature extraction and support vector machine[J].Journal of Nanjing University (Natural Science),2020,56(6):870-876
[8] Smith R W.History of the Tesseract OCR engine:what worked and what didn't:how to build a world-class OCR engine in less than 20 years[R].Proceedings of SPIE,2013,8658:865802-1-865802-20
[9] Nor D M,Wahab M H A,Jenu M Z M,et al.A new visual signature for content-based indexing of low resolution documents [J].Journal of Information Retrieval and Knowledge Management,2012,12(2):88-95
[10] Ma Y W,Wang B,Hu H T.Hybrid model for Chinese character recognition based on Tesseract-OCR[J].International Journal of Internet Protocol Technology,2020,13(2):102-108
[11] 石煌雄,胡洋,蒋作,等.基于深度学习的电气铭牌可变区域识别方法的研究[J].云南民族大学学报(自然科学版),2020,29(4):350-355
SHI Huangxiong, HU Yang, JIANG Zuo, et al. A recognition method for the variable region of electrical nameplate based on deep learning[J].Journal of Yunnan Minzu University (Natural Sciences Edition),2020,29(4):350-355

Adaptive recognition of complex invoices based on Tesseract-OCR

SUN Ruibin[1] QIAN Kui[1] XU Weimin[2] LU Hong[1]
1 School of Automation,Nanjing Institute of Technology,Nanjing 211167
2 Nanjing Xuefu Ruijie Information Technology Company Limited,Nanjing 210009

Abstract An adaptive recognition method based on Tesseract-OCR engine is proposed to solve the problem of extracting and real-time recognition of specific table items in any region of complex invoices.First,the invoice image is preprocessed by OpenCV for filtering,adaptive threshold,etc.,to get a binary image.Then,the open operation in morphology is used to extract the global line segments and position of the table.The coordinates of the intersection points of the table is combined with the custom template to realize the adaptive adaptation between the table header and the content.Then the jTessBoxEditor is used to train and optimize the content of the table items,and finally the character recognition based on Tesseract-OCR is realized.The experimental results show that this method has high accurate recognition rate,supports the adaptive recognition of ROI (Region of Interest),and is highly available.

Key words invoice recognition;Tesseract-OCR;OpenCV;character training;adaptive recognition