2025 年 第 25 卷 第 9 期
2025, 25(9):03788-07

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2402410
引用格式:贺锋, 张威, 杨玉燕, 等. 增值税发票全票面结构化识别[J]. 科学技术与工程, 2025, 25(9): 3788-3794.
He Feng, Zhang Wei, Yang Yuyan, et al. Full-ticket structural recognition of VAT invoice[J]. Science Technology and Engineering, 2025, 25(9): 3788-3794.

增值税发票全票面结构化识别

贺锋[1], 张威[1], 杨玉燕[2], 陈博扬[1], 王建松[2]
(1. 华中科技大学电子信息与通信学院, 武汉 430074; 2. 广东烟草梅州市有限公司, 梅州 514000)

摘 要 增值税发票商品明细部分的项目名称、规格型号等的格式和内容非常灵活复杂,且缺乏完整表格线对各信息字段进行分隔,现有方法对增值税发票进行全票面信息结构化识别还存在元素识别率低、计算复杂度过高等问题,提出一种基于计算机形态学的全票面信息结构化识别方法。该方法采用形态学操作检测发票表格线,对发票不同区域裁切并识别文字;再利用增值税发票商品明细区域版面排布隐含规则,结合计算机形态学操作获得的文字连通区域,构建完整表格结构;最后基于文本检测神经网络(text detection neural network with differentiable binarization, DBNet)和卷积递归神经网络(convolutional recurrent neural network,CRNN)实现文本的检测和识别。提出的方法在 3 种版式共 49 张增值税发票数据集上测试,结果表明,元素识别率分别达到 99.9%、97.4%和 98.8%,单张平均运行时间分别为 0.90、0.47 和 0.82 s,全票面结构化识别性能超过多个对照表格识别模型以及文献方法。

关键词 增值税发票; 表格检测; 形态学操作; 结构化识别; 倾斜校正; 红章消除
中图法分类号 TP391.1; **文献标志码** A

Full-ticket Structural Recognition of VAT Invoice

HE Feng[1], ZHANG Wei[1], YANG Yu-yan[2], CHEN Bo-yang[1], WANG Jian-song[2]
(1. School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China;
2. Meizhou Tobacco Monopoly Bureau (Company), Meizhou 514000, China)

[**Abstract**] The format and content of items such as product names and specifications in the detailed section of VAT invoices are highly flexible and complex, lacking complete gridlines to separate information fields. Existing methods for all-element structural recognition of VAT invoices face issues like low element recognition rates and high computational complexity. A structured recognition method for full face information based on computer morphology was proposed, which uses morphological operations to detect invoice table lines, cuts and recognizes text in different areas of the invoice. Then the implicit rules of the layout of the value-added tax invoice product details area was reused, combined with the text connected areas obtained through computer morphology operations, to construct a complete table structure. Finally, text detection and recognition were achieved using text detection neural network with differentiable binarization (DBNet) and convolutional recurrent neural networks (CRNN). The proposed method was tested on a dataset of 49 value-added tax invoices in three different formats, and the results show that the element recognition rates reached 99.9%, 97.4%, and 98.8%, respectively. The average running time per invoice is 0.90, 0.47, and 0.82 s, respectively. The structural recognition performance of the entire invoice exceeded multiple comparison table recognition models and literature methods.

[**Keywords**] VAT invoice; table detection; morphological operations; structural recognition; tilt correction; seal elimination

增值税发票是记录商品或服务交易经济行为的一种重要凭证。通过光学字符识别技术(optical character recognition, OCR)提取发票关键信息,并根据预定的规则去评判报销的单据,可对出现的风险点及时给出预警提示,避免人为的疏漏,降低风险,提升审核效率[1]。通过提取的数据累积,可以多维度地对报销的单据做分析比较,加强费用的管理,提升财务管理的效能,积聚财务数据资产,为智慧财务的实现提供丰富的数据基础。随着智慧财务的发展,对增值税发票的信息提取需求已经拓展到商品明细清单的完整提取。然而增值税发票商品明细区域的项目名称、规格型号等的格式和内容非常灵活复杂,且缺乏完整表格线对各信息字段进行分隔,因此对发票商品明细的结构化识别具有更大挑战性[2-3]。

已有一些文献提出了不同的增值税发票识别方法。王阳等[4]采用 VGG-16 卷积神经网络对财务票据文字方向进行 0°、90°、180°和 270°四分类,使

收稿日期: 2024-04-03 修订日期: 2024-12-06
基金项目: 梅州市烟草专卖局(公司)科技项目(2023441400240048)
第一作者: 贺锋(1977—),男,汉族,江西永新人,博士,副教授。研究方向:计算机视觉、目标检测与识别。E-mail:hefeng@hust.edu.cn。

用 YOLOv3 检测文本行,最后用 CRNN 网络进行文字识别。何鎏一等[5]对光照不均匀增值税发票图像进行了图像增强后采用 CRNN 模型进行文字识别。谢阳等[6]提出基于形态学方法检测增值税发票中的表格线,从而实现对发票各部分的裁切和文字识别。王兴等[7]基于百度 PaddleOCR 对增值税发票进行文本检测和识别。尹潇伟等[8]结合中文票据文本的特点,提出了改进的文本检测和文本识别模型。以上发票信息提取方案的技术路线都可以总结为:先对发票进行文本检测,再对检测到的文本进行文字识别,最后通过关键字匹配提取发票信息字段,但这些文献的方法都没有涉及发票商品明细部分隐式表格结构的重建,只能用于商品明细比较简单的发票或不需要提取明细的场景。

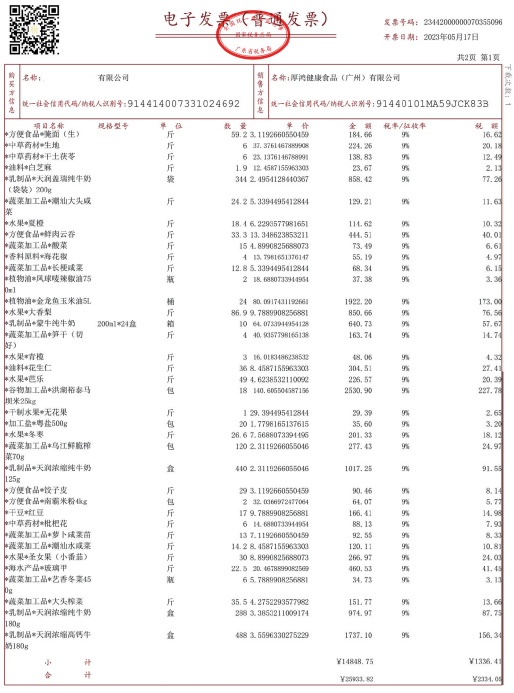
时瑞等[9]提出模板与内容分离的发票识别方法,利用票据模板与内容的颜色差别进行分离,采用孪生神经网络将输入图像中分离出的模板与模板数据库模板进行匹配从而提取票据结构,最后用百度 PaddleOCR 完成文字检测与识别,然而发票与模板的匹配准确率低于 80%。唐军等[2]提出一种增值税发票全票面结构化识别方案,采用 HRNet 进行发票关键点检测,再用 YOLOv4 进行发票元素检测,最后用 CRNN 进行文本识别,整个方案非常复杂,需要训练多个深度学习模型,但发票全票面信息字段识别准确率只有 69.3%。以上文献方案技术路线可以总结为:先定位发票各区域版面排布,再识别其中的文字,但也没有对发票商品明细部分隐式表格结构进行重建。

综上,现有文献给出的增值税发票识别方案均未涉及商品明细部分隐式表格结构的重建,并且对这部分的识别准确率难以达到应用需求。考虑到基于深度学习的表格识别存在数据收集和标注困难、结果可解释性差、增值税发票版面结构复杂、排版规则隐含等因素,因此现提出一种增值税发票全票面结构化识别方法,采用可解释性强、鲁棒性好的形态学方法,利用增值税发票版面排布的先验规则,无需训练即可实现对发票各信息元素进行版面分割,进而对商品明细区域基于形态学操作构建隐含的表格线,完成发票表格完整结构的构建,最终实现高效全票面结构化识别。

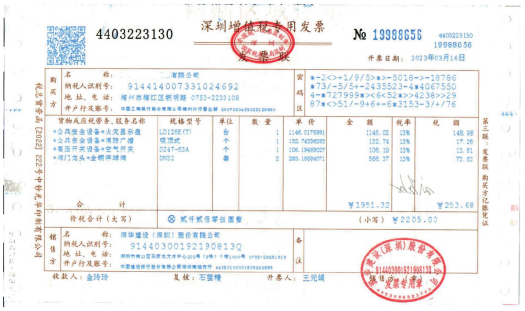
1　增值税发票结构化识别

1.1　结构化识别总体流程

根据增值税发票版面排布的差异,增值税发票(含普票、专票、电子票)及其附属明细清单主要涉及图 1 所示的 3 种版式。图 1(a)为类型Ⅰ的电子



(a)类型Ⅰ



(b)类型Ⅱ

(c)类型Ⅲ

图 1　增值税发票示例

Fig. 1　Sample VAT invoices

发票,商品明细部分各字段之间没有表格横竖线;图 1(b)为类型Ⅱ增值税发票的商品明细部分有竖线但没有横线,当明细内容较多时发票另附的清单如图 1(c)为类型Ⅲ,同样没有横线。

增值税发票版面复杂,商品明细部分信息字段密集存于隐式表格单元中,项目名称、规格型号部分的单元格存在跨上下多个文本行的情况。增值税发票信息结构化提取总体流程如图 2 所示。读取图像后先检测表格获得发票的表格线坐标,并根据表格外围轮廓信息完成表格倾斜校正。根据检测到的表格线,分区域分别进行信息提取并存入 Excel 表格。发票全票面主要分 3 个部分信息。

(1)表格头,包含发票名称、发票号码、开票日期等,通过匹配关键字方式逐个提取。

(2)购买方/销售方信息,根据发票名称及表格线坐标确定其区域,OCR 后匹配关键字逐个提取。

(3)商品明细信息,包含项目名称、规格型号、金额等 8 个字段,根据发票类型对该区域进行隐含的表格线构建,再对该区域进行 OCR,将检测并识别的文本根据其中心坐标查找其所属单元格位置后存入 Excel 表格。

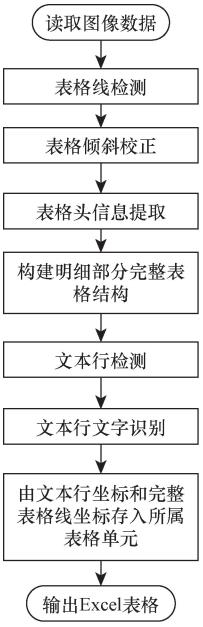


图 2 增值税发票结构化识别流程

Fig. 2 Flowchart of VAT invoice structural recognition

1.2 发票表格线检测

采用形态学方法对输入图像首先检测表格横竖线,进而检测表格边界轮廓,相关步骤如下。

步骤 1 检测横竖线。

(1)将输入图像转为灰度图,对灰度图取反后进行自适应阈值二值化。

(2)构造一条行为 1、列为图像宽度 1/30 的横线矩阵模板,用横线矩阵模板对二值图进行形态学开操作,得到只保留了二值图中所有横线的横线图。

(3)构造一条行为图像高度 1/30、列为 1 的竖线矩阵模板,用竖线矩阵模板对二值图进行开操作,得到只保留了二值图中所有竖线的竖线图。

横、竖线矩阵模板的宽度和高度根据实际运行效果设置,过大会增加表格线断裂概率,过小会保留一些不属于表格线的结构单元。

步骤 2 检测表格。

(1)将横线图和竖线图相加得到表格掩码图(如图 3 所示)。

(2)对表格掩码图检测面积最大的最外层轮廓。

(3)将提取到的轮廓拟合为多边形并计算多边形的最小矩形包围框。

(4)计算轮廓的最小矩形包围框的倾斜角度并反向旋转实现倾斜校正,输出表格坐标信息。

步骤 3 检测表格横、竖线坐标。

(1)在竖线图中取穿过表格中心位置的一条长度为图像宽度的水平检测线;查找线上所有在表格宽度范围内且像素值非零的点的横坐标,并从这些连续非零的点集中取中心点的横坐标为某条竖线的横坐标,依次获得竖线图中从左到右的竖线的横坐标 $x_1, x_2, \cdots$。

(2)在横线图中取穿过表格中心位置的一条长度为图像高度的竖直检测线;查找线上所有在表格高度范围内且像素值非零的点,并从这些连续非零的点集中取中心点的纵坐标为某条横线的纵坐标,依次获得横线图中从上到下的横线的纵坐标 y_1, $y_2, \cdots$。

采用步骤 3 的方法,根据表格的类型和结构,只需改变检测线的位置还可以得到表格中其他位置的横竖线坐标。

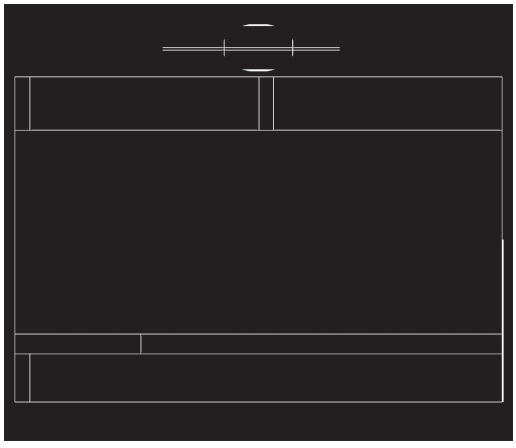


图 3 表格掩码图

Fig. 3 Mask bitmap of table

1.3　发票商品明细区域隐式表格构建

根据图1所示含隐式表格的增值税发票商品明细区域信息字段排布规则,结合形态学操作将文字连通为文字块用以定位待添加横竖线的位置。从图1(a)可以看出,此类增值税发票中间需添加7条竖线,依次取“规格型号”“单位”的左边缘和“单位”“数量”“单价”“金额”“税率/征收率”的右边缘为竖线位置。发票加竖线算法描述如下。

1.3.1　算法1 发票商品明细区域加单元格竖线

输入:裁切的尺寸为 $H \times W$ 的商品明细区域图像 I_a。

输出:表格内部竖线横坐标列表 $X=[x_1,x_2,\cdots,x_7]$。

(1) 将 I_a 转为灰度图并进行自适应二值化得到 I_b。

(2) 使用像素值为1尺寸为 $\frac{W}{120}\times\frac{W}{35}$(垂直×水平)的矩形结构元素 S_1 对 I_b 进行开操作,$I_{\text{open}}=I_b\circ S_1$。得到二值图像的文字区域连通图 I_{open}(图4)。

(3) 在 $I_{\text{open}}(5,5)$ 坐标处取通过该点线长为W-10、线宽为1的横线 L_{h1},遍历 L_{h1},将像素值从0到255或从255到0转变的位置横坐标存入列表,取列表的第3、5、6、8、10、13号成员添加至列表得 $X=[x_1,x_2,x_3,x_4,x_5,x_7]$。

(4) 在 $I_{\text{open}}(5,\text{H}-5)$ 坐标处取通过该点线长为W、线宽为1的横线 L_{h2},遍历横线,将像素值从0到255或从255到0转变的位置横坐标存入列表,取列表的第6号成员添加至列表得 $X=[x_1,x_2,x_3,x_4,x_5,x_6,x_7]$,返回该竖线横坐标列表 X。

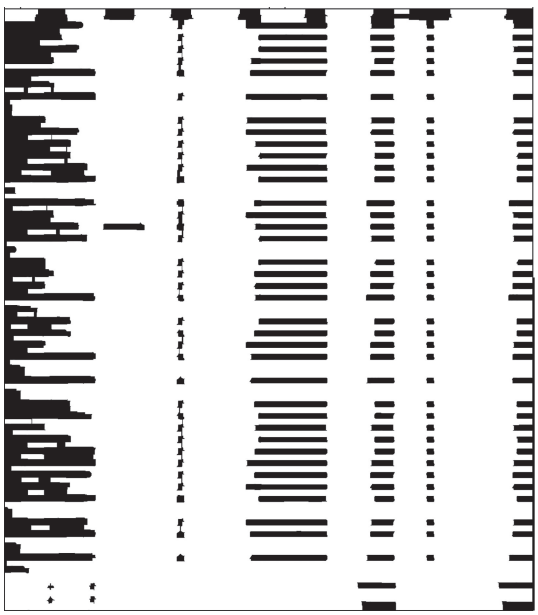


图4　加竖线算法得到的文字连通图

Fig. 4　Connected text obtained by algorithm of adding vertical line

增值税发票商品明细区域还需要添加单元格横线。从图1可以看出,“金额”列不存在同一表格单元多个文本行情况,取“金额”列的每个文本行上移一个小的固定偏置即为合适的单元格横线位置。发票加横线算法描述如下。

1.3.2　算法2 发票商品明细区域加单元格横线

输入:裁切的尺寸为 $H_2\times W_2$ 发票“金额”列图像 I_a。

输出:商品明细区域横线纵坐标列表 $Y=[y_1,y_2,\cdots,y_m]$。

(1) 将 I_a 转为灰度图并进行自适应阈值二值化得到 I_b。

(2) 使用像素值为1,尺寸为 $\frac{W}{150}\times\frac{W}{48}$ 的矩形结构元素 S_1 对 I_b 进行开操作,$I_{\text{open}}=I_b\circ S_1$。得到二值图像的文字区域连通图 I_{open}(图5)。

(3) 在 $I_{\text{open}}(W_2-5,5)$ 坐标处取通过该点线长为 H_2-10、线宽为1的竖线 L_{v},遍历 L_{v},将像素值从255到0转变(白到黑)的位置纵坐标添加到列表 Y。

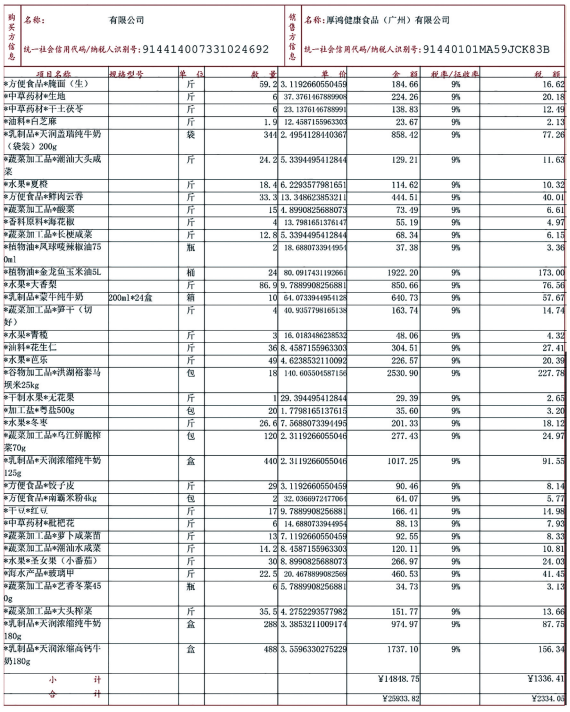
(4) 如果 L_{v} 上第一个点像素值为0,遍历 L_{v},找到第一个从0到255转变的位置 y_1',列表 Y 最后一个元素为 y_{last},字符高度 $h_{\text{char}}=H_2-y_{\text{last}}$,$y_1=y_1'-h_{\text{char}}$,将 y_1 添加到列表 Y。如果“金额”与第一行数字粘连,补第一条横线。

(5) 返回列表 Y。

图5　加横线算法得到的文字连通图

Fig. 5　Connected text obtained by algorithm of adding horizontal line

图6为使用算法1和算法2得到的添加了单元格横竖线的发票,可以看出,添加的表格线将隐式单元格全部正确分隔开。



购买方信息	名称: 有限公司				销售方信息	名称:厚鸿健康食品(广州)有限公司		
	统一社会信用代码/纳税人识别号:91441400733102469 2					统一社会信用代码/纳税人识别号:91440101MA59JCK83B		
项目名称	规格型号	单位	数量	单价	金额	税率/征收率	税额	
*方便食品*腌面(生)		斤	59.2	3.1192660550459	184.66	9%	16.62	
*中草药材*生地		斤	6	37.376146788990	224.26	9%	20.18	
*中草药材*干土茯苓		斤	6	23.137614678899	138.83	9%	12.49	
*油料*白芝麻		斤	1.9	12.458715596330	23.67	9%	2.13	
*乳制品*天润盖瑞纯牛奶(袋装)200g		袋	344	2.4954128440367	858.42	9%	77.26	
*蔬菜加工品*潮汕大头菜咸菜		斤	24.2	5.3394495412844	129.21	9%	11.63	
*水果*夏橙		斤	18.4	6.2293577981651	114.62	9%	10.32	
*方便食品*鲜肉云吞		斤	33.3	13.348623853211	444.51	9%	40.01	
*蔬菜加工品*酸菜		斤	15	4.8990825688073	73.49	9%	6.61	
*香料原料*海花椒		斤	4	13.798165137614	55.19	9%	4.97	
*蔬菜加工品*长梗咸菜		斤	12.8	5.3394495412844	68.34	9%	6.15	
*植物油*风球唛辣椒油75ml		瓶	2	18.688073394495	37.38	9%	3.36	
*植物油*金龙鱼玉米油5L		桶	24	80.091743119266	1922.20	9%	173.00	
*水果*大香梨		斤	86.9	9.7889908256881	850.66	9%	76.56	
*乳制品*爱牛纯牛奶	200ml*24盒	箱	10	64.073394495412	640.73	9%	57.67	
*蔬菜加工品*笋干(切好)		斤	4	40.935779816513	163.74	9%	14.74	
*水果*青梅		斤	3	16.018348623853	48.06	9%	4.32	
*油料*花生仁		斤	36	8.4587155963303	304.51	9%	27.41	
*水果*芭乐		斤	49	4.6238532110092	226.57	9%	20.39	
*冷冻加工品*洪湖裕泰马蹄米25kg		包	18	140.60550458716	2530.90	9%	227.78	
*干制水果*无花果		斤	1	29.394495412844	29.39	9%	2.65	
*加工品*粤盐500g		包	20	1.7798165137615	35.60	9%	3.20	
*水果*冬枣		斤	26.6	7.3688073394495	201.33	9%	18.12	
*蔬菜加工品*乌江鲜脆榨菜70g		包	120	2.3119266055046	277.43	9%	24.97	
*乳制品*天润浓缩纯牛奶125g		盒	440	2.3119266055046	1017.25	9%	91.55	
*方便食品*饺子皮		斤	29	3.1192660550459	90.46	9%	8.14	
*方便食品*南畹米粉4kg		包	2	32.036697247706	64.07	9%	5.77	
*干货*红豆		斤	17	9.7889908256881	166.41	9%	14.98	
*中草药材*鱼腥花		斤	6	14.680733944954	88.13	9%	7.93	
*蔬菜加工品*萝卜咸菜苗		斤	13	7.1192660550459	92.55	9%	8.33	
*蔬菜加工品*潮汕水咸菜		斤	14.2	8.4587155963303	120.11	9%	10.81	
*水果*圣女果(小番茄)		斤	30	8.8990825688073	266.97	9%	24.03	
*海水产品*玻璃甲		斤	22.5	20.467889908256	460.53	9%	41.45	
*蔬菜加工品*艺香冬菜450g		瓶	6	5.7889908256881	34.73	9%	3.13	
*蔬菜加工品*大头榨菜		斤	35.5	4.2752293577982	151.77	9%	13.66	
*乳制品*天润浓缩纯牛奶180g		盒	288	3.3853211009174	974.97	9%	87.75	
*乳制品*天润浓缩高钙牛奶180g		盒	488	3.5596330275229	1737.10	9%	156.34	
小 计					¥14848.75		¥1336.41	
合 计					¥25933.82		¥2334.08	

图6 算法1和算法2构建的完整表格结构

Fig. 6 Complete table structure constructed using algorithm 1 and 2

1.4 发票表格检测与表格构建计算复杂度

发票的表格线检测需对输入图像进行两次形态学开操作、轮廓查找、图像旋转等操作,而商品明细部分隐式表格的构建需要对相应区域进行一次形态学开操作。对于一个 $M \times N$ 大小的图像和一个 $k \times 1$ 大小的结构元素,不论膨胀还是腐蚀,对于每个像素,结构元素需要在其周围滑动 $k \times 1$ 次,因此,总的操作次数大约为 $M \times N \times k \times 1$。所以,一次形态学开操作计算复杂度为 $O(2MNkl)$。由于 OpenCV 库对形态学操作已经高度优化,实际计算时间相对较短。使用 Intel i9-10940X CPU @ 3.30 GHz,对于图6所示 1 697×1 199 的图像进行表格线检测需 38.2 ms,画表格竖线用时 16.2 ms,画表格横线用时 6.4 ms。

1.5 文本检测与识别

文本检测模型采用 DBNet[10],输入文本图像,输出检测到的文本行包围框的4个角的坐标。文本识别选用 OCR 领域里常用的 CRNN[11]。对于文档字符的识别,CRNN 具有轻量且准确率高的优点。CRNN 模型对输入文本行的灰度图像用类似 VGG 架构的7层卷积网络提取特征,然后对特征图序列化,再采用双层双向 LSTM 对特征图序列进行解码输出字符序列。CRNN 对文本行图像直接进行序列字符识别,无需先进行字符分割,避免了文本中由于字符间隔小、图像模糊等带来的分割难的问题。

2 增值税发票结构化识别实验

2.1 测试数据集

对于含隐式表格的6种增值税发票(含发票附属明细清单),包括:电子发票(普通发票)、电子发票(增值税专用发票)、增值税专用发票、增值税普通发票、增值税电子普通发票、增值税专用发票附属销售货物清单,已经涵盖了企业财务报销票据种类中除交通费外的绝大部分。

上述6种增值税发票的版式可归并为图1所示3种类型。收集了图1(a)所示无单元格横竖线的电子发票24张(类型Ⅰ);图1(b)所示无单元格横线的增值税发票(类型Ⅱ)18张,其中10张为纸质扫描图像;图1(c)所示增值税专用发票附属商品明细清单(类型Ⅲ)7张作为测试数据集。类型Ⅰ发票总字符数22 486,总单元格数5 116;类型Ⅱ发票总字符数5 715,总单元格数536;类型Ⅲ发票总字符数5 538,总单元格数981。

2.2 实验评估指标

实验使用元素识别率(element correct ratio, ECR)、字符识别率(char correction ratio, CCR)评估信息提取性能。定义的增值税发票元素包括发票名称、发票代码、发票号码、开票日期、购买方信息、销售方信息以及发票明细部分的项目名称、规格型号、单位、单价、数量、金额、税率、税额。

$$\mathrm{ECR} = \frac{\#\mathrm{correct\_elements}}{\#\mathrm{total\_elements}} \times 100\% \tag{1}$$

式(1)中:elements 为票面上待提取的结构化数据中的任意一个元素;#correct_elements 为正确识别的元素数量(该元素所有字符均识别正确并且其表格位置被正确识别);#total_elements 为票面上待提取的结构化数据中的所有元素。

$$\mathrm{CCR} = \frac{\#\mathrm{correct\_chars}}{\#\mathrm{total\_chars}} \times 100\% \tag{2}$$

式(2)中:chars 表示票面待提取元素中的任意一个字符;#correct_chars 表示正确识别的字符数量(这些正确识别的字符必须处于正确的表格位置);#total_chars 为票面待提取元素中的所有字符。

2.3 实验结果

分别使用本文方法、商用的百度云表格文字识别 v2、海康威视提出的 LGPMA[12] 表格识别模型和 PaddleOCR 表格识别模型 SLANet[13] 进行测试。百度云表格文字识别 v2 代表当前最新的商用表格识别能力,LGPMA 和 SLANet 为具有代表性的基于深度学习的端到端表格识别模型。测试结果如表1~表3所示。

表 1　类型 I 增值税发票识别结果

Table 1　Recognition results of type I VAT invoice

方法	ECR/%	CCR/%	运行时间/s
百度云表格	—	—	—
LGPMA	—	—	5.11
SLANet	—	—	2.06
本文方法	99.9	99.9	0.900

注:“—”表示因表格识别错误太多而未统计该项数据。

表 2　类型 II 增值税发票识别结果

Table 2　Recognition results of type II VAT invoice

方法	ECR/%	CCR/%	运行时间/s
百度云表格	—	—	1.65
LGPMA	—	—	2.80
SLANet	—	—	0.868
文献[2]	69.3*	—	—
本文方法	97.4	98.9	0.467

注:“—”表示因表格识别错误太多而未统计该项数据;* 本文与文献[2]使用的增值税发票测试数据集不同,其未公开数据集。

表 3　类型 III 增值税发票识别结果

Table 3　Recognition results of type III VAT invoice

方法	ECR/%	CCR/%	运行时间/s
百度云表格	96.85	99.91	3.40
LGPMA	—	—	4.10
SLANet	—	—	1.35
本文方法	98.8	99.37	0.816

注:“—”表示因表格识别错误太多而未统计该项数据。

分析表 1 ~ 表 3 可知,本文方法对隐式单元格发票的表格识别性能具有明显优势,具体表现在以下方面。

从表格识别性能上,商用的百度云表格识别 v2 不支持类型Ⅰ发票识别,返回了错误;对类型Ⅱ的发票返回的 Excel 文件中没有识别商品明细部分的隐式表格;对类型Ⅲ表格因单元格识别错位和单元格漏检产生一些错误,元素识别率为 96.85%,低于本文方法。LGPMA 和 SLANet 表格识别模型对于 3 种类型的发票表格返回的 Excel 中存在大量表格单元错位;没有建立完整表格结构。文献[2]对类型Ⅱ的增值税发票元素识别率仅为 69.3% 明显低于本文方法的 97.4%。本文方法对 3 种类型发票表格元素识别率分别达到 99.9%、97.4% 和 98.8%,字符识别率分别达到 99.9%、98.9% 和 99.37%,其中元素识别错误主要原因包括:印章干扰、文本漏检、单元格错位、文字识别错误等。

从运行时间上,本文方法运行速度比 3 种对照方法都更快,相对速度第二的 SLANet 在 3 个类型发票上处理速度分别高了 129%、86% 和 65%。本文方法进行发票识别所需时间主要包括表格检测、画表格横竖线、文字检测和识别 3 个部分。表格检测时间与图像像素数成正比,一般费时在 15 ~ 34 ms;画表格横竖线只需对部分区域进行形态学操作,一般在 22 ms 以内;文字的检测与识别占据剩余的时间。由于本文方法在表格线检测和隐式表格构建方面均采用底层高度优化的形态学操作,实际运行效率高。

2.4　增值税发票识别应用实例分析

基于 A 企业财务报销合规性检查和财务数据分析的需要,应用所提出的增值税发票全票面结构化识别方法将每张增值税发票(含发票附属明细清单)进行结构化提取后保存为 Excel 文件,提取内容为 2.2 节所述发票元素。输入为一次报销单的所有票据图像文件夹,逐一读取图像并采用图 2 所示流程进行全票面结构化识别。检测发票表格后,统一将发票表格部分 resize 为宽 1 200 像素。形态学操作、自适应二值化、获取图像轮廓、获取图像倾斜角度、图像旋转等均调用 OpenCV 相应 API 实现。文本检测和识别基于 PaddleOCR 框架,直接调用已经训练好的 DBNet 和 CRNN 模型。

表格头信息提取时根据商品明细部分的表格竖线坐标,将表格头大致分为左中右 3 个部分,分别截取后进行文字识别,这样提高了识别后处理的便利。发票名称部分存在严重的红章干扰,极易导致名称识别错误。将发票名称区域的图像转到 HSV 空间后,提取 S 通道并归一化后直接进行文字识别能够很大程度缓解这个问题。发票购买方和销售方信息提取也采用以上裁切再识别提取的方法。最后对 Excel 各单元进行后处理,包括去除多余空格、金额的核验、必备字段的检查等,后处理发现的问题打印提示信息以便人工核查。

本文方法在 A 企业实际财务票据信息提取任务中能适应绝大部分情况。金额、税率等数字部分的信息提取在票面无污染或干扰的情况下出错概率极低,项目名称字段的字符可能非常多,因而是文字识别错漏发生概率最高的字段,但该字段信息冗余大,即使出现很少量的错漏基本不影响关键信息。应用中也发现一些人为干扰问题,例如发票背面手写签名、发票粘贴纸上印的说明文字透过发票票面等,但这些人为干扰可以采取简单措施予以规避。

3　结论与展望

设计了一种增值税发票全票面结构化识别方法,得到以下结论。

(1)采用形态学操作对图像提取表格线和表格

外边框轮廓后进而得到表格最小矩形包围框后即可精确计算出表格倾斜角度完成小角度的倾斜校正,避免了常用的霍夫变换或神经网络检测倾斜角度带来的计算量大和精度不足问题。

(2)对发票商品明细部分的隐式表格单元进行形态学操作后提取隐含的表格横竖线,可完美构建完整表格结构。

(3)在实际场景采集的49张包含电子发票、扫描版和非扫描版增值税发票、增值税专用发票附属销售清单的三个数据集上,本文方法对于增值税发票全票面元素提取得到的元素识别率分别达到99.9%、97.4%和98.8%,运行时间分别为0.900、0.467和0.816 s。发票识别性能超过多个对照表格识别模型以及文献方法。

本文方法已经经过一定数量的增值税发票测试表明了较好的识别性能和鲁棒性,然而当前版本还存在一些不足。目前使用的文本检测模型DBNet对一些小目标例如孤立的数字"1"存在漏检现象需要针对性调优;提出的红章去除方法对于发票名称部分的印刷红章有较好效果,然而对备注区域的油印红章效果欠佳;调用的PaddleOCR框架下的文本识别模型CRNN对于个别特定字符容易识别错误需要针对性微调训练。

参 考 文 献

[1] 周俊亭, 席彦群, 周媛媛, 等. 大数据、人工智能与财税服务创新[J]. 中国软科学, 2020(8): 69-77.
Zhou Juting, Xi Yanqun, Zhou Yuanyuan, et al. Big data, artificial intelligence and innovation of fiscal and taxation service[J]. China Soft Science, 2020(8): 69-77.

[2] 唐军,唐潮. 增值税发票信息结构化识别[J]. 计算机系统应用, 2021, 30(12): 317-325.
Tang Jun, Tang Chao. Structural information recognition of VAT invoice[J]. Computer Systems & Applications, 2021, 30(12): 317-325.

[3] 苗峰. 基于影像识别的多系统集成的增值税发票精细管理实践[J]. 中国管理信息化, 2020, 23(23): 95-98.
Miao Feng. Refined management practice of value-added tax invoices through multi-system integration based on image recognition[J]. China Management Informationization, 2020, 23(23): 95-98.

[4] 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(S2): 375-379.
Wang Yang, Li Zhendong, Yang Guanci. Application research of OCR character recognition based on deep learning in banking industry[J]. Application Research of Computers, 2020, 37(S2): 375-379.

[5] 何鎏一,杨国为. 基于深度学习的光照不均匀文本图像的识别系统[J]. 计算机应用与软件, 2020, 37(6): 184-190.
He Liuyi, Yang Guowei. Recognition system of illumination uneven text image based on deep learning[J]. Computer Applications and Software, 2020, 37(6): 184-190.

[6] 谢阳,程艳云. 基于OpenCV形态学的发票定位研究[J]. 计算机与数字工程, 2021, 49(4): 809-812.
Xie Yang, Cheng Yanyun. Invoice location research based on OpenCV morphology[J]. Computer & Digital Engineering, 2021, 49(4): 809-812.

[7] 王兴,郑勇锋, 严永兵, 等. 基于OCR技术的票据识别算法研究[J]. 智能计算机与应用, 2021, 11(11): 101-106.
Wang Xing, Zheng Yongfeng, Yan Yongbing, et al. Research on bill recognition algorithm based on OCR[J]. Intelligent Computer and Applications, 2021, 11(11): 101-106.

[8] 尹潇伟,孙仁诚, 王霄鹏, 等. 基于深度学习的中文票据文本检测与识别方法[J]. 青岛大学学报(自然科学版), 2022, 35(4): 1-7.
Yin Xiaowei, Sun Rencheng, Wang Xiaopeng, et al. Chinese invoice text detection and recognition method based on deep learning[J]. Journal of Qindao University (Natural Science Edition), 2022, 35(4): 1-7.

[9] 时瑞,蒋三新. 基于模板与内容分离的票据识别方法[J]. 电子测量技术, 2023, 46(6): 122-128.
Shi Rui, Jiang Sanxin. Bill recognition method based on template and content separation[J]. Electronic Measurement Technology, 2023, 46(6): 122-128.

[10] Liao M, Zou Z, Wan Z, et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2023, 45(1): 919-931.

[11] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(11): 2298-2304.

[12] Qiao L, Li Z, Cheng Z, et al. LGPMA: complicated table structure recognition with local and global pyramid mask alignment[J]. ArXiv, 2021: 2105.06224.

[13] Inc Baidu. PaddleOCR table recognition [EB/OL]. (2022-10-24)[2024-04-03]. https://github.com/PaddlePaddle/PaddleOCR/blob/main/ppstructure/table/README_ch.md.