计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):317−325 [doi: 10.15888/j.cnki.csa.008268]

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

增值税发票信息结构化识别①

唐　军, 唐　潮

(四川虹微技术有限公司, 成都 610041)
通讯作者: 唐　潮, E-mail: xsb.tangchao@changhong.com

摘　要: 为进一步简化增值税发票识别流程和和提高识别效率, 提出了一种基于 HRNet 和 YOLOv4 的增值税票面信息结构化识别的方法. 首先利用 HRNet 进行增值税发票关键点检测, 进行增值税发票对齐; 其次利用 YOLOv4 进行发票元素的检测; 然后通过 CRNN 对发票元素进行文本识别; 最后形成结构化数据. 在业务数据集中的实验表明, 检测准确率在 0.5 mAP 下达到 75.7, 检测速度达到 12.85 fps, 元素识别率 ECR 达到 69.30%, 实验结果表明算法能有效简化识别流程, 提高识别准确率, 在实时性要求较高和业务噪声复杂的增值税票据识别中有较好适应性和广泛应用前景.

关键词: 增值税发票; 发票识别; HRNet; YOLOv4; CRNN; 结构化识别

引用格式: 唐军,唐潮.增值税发票信息结构化识别.计算机系统应用,2021,30(12):317–325. http://www.c-s-a.org.cn/1003-3254/8268.html

Structural Information Recognition of VAT Invoice

TANG Jun, TANG Chao

(Sichuan Homwee Technology Co. Ltd., Chengdu 610041, China)

Abstract: To simplify the processing steps of VAT invoices and improve recognition accuracy, we propose a method based on HRNet and YOLOv4 to extract structural information of VAT invoices. Firstly, we detect predefined keypoints in the VAT invoice with the HRNet method to align the invoice to a standard template. Then detect the structural information cell in the invoice by YOLOv4. And lastly use CRNN to recognize the cell block image to obtain structural data. The experimental results on real business VAT invoices show that the proposed method gets a detection accuracy of 75.7 at 0.5 mAP, reaches a detection speed at 12.85 fps, and achieves an Element Correct Ratio (ECR) at 69.30%. The results indicate that the proposed method can simplify the process and improve the accuracy of recognition, and it can apply to the scene where requires high real-time performance and needs to deal with complicated noise situation.

Key words: VAT invoice; invoice recognition; HRNet; YOLOv4; CRNN; structural recognition

1　引言

在财务共享中心的日常业务中对纸质发票尤其是国内增值税发票的处理具有重要意义. 大型企业常常因为需要及时处理海量增值税发票而投入大量人力成本, 并且在资金风险管控、报税纳税管理等财务管理中还需维持一定程度的日常资源或管理措施等来强化监督[1].

目前, 大型的财务共享中心都集成了一定程度的自动化处理技术. 一个简化的集成自动化处理技术的系统如图 1 所示.

在实际的财务报销系统票据中, 增值税发票录入工作占据了大部分时间, 因此开发增值税发票识别系统并提高其识别率, 具有重要意义[2]. 当前已有一些研究人员进行了和增值税发票识别相关的研究[3–13], 但都没有涉及增值税发票全票面元素的识别.

① 收稿时间: 2021-02-10; 修改时间: 2021-03-18, 2021-04-02, 2021-04-26; 采用时间: 2021-05-07

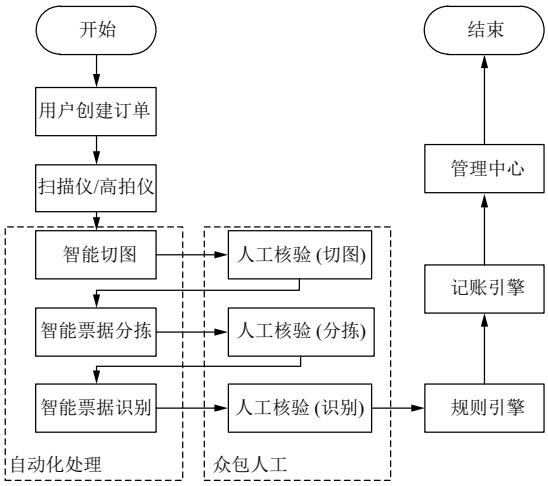

图 1 简化的集成自动化处理技术的财务共享中心

增值税发票识别, 主要由两大任务组成, 一个是光学字符识别 (Optical Character Recognition, OCR), 即, 在图像的什么位置有什么文字, 一个是信息结构化抽取 (Structural Information Extraction, SIE), 可理解为票据版面分析、票面结构化元素识别. 目前, OCR 识别技术相对已比较成熟, 主要可分三大研究方向: 一是文字检测技术, 二是文本识别技术, 三是端到端检测识别技术. 三大研究方向中典型的技术可见表 1.

通过国内外研究可知, 增值税发票识别虽然可分为两大任务, 但是由于版面较为复杂, 因此在综合 OCR 和结构化元素处理两方面时, 研究思路也多种多样, 但大体上可分为 3 类, 第一类是先进行常规 OCR, 然后进行版面分析; 第二类是先进行版面分析, 再进行 OCR; 最后一类是直接端到端, OCR 和版面分析同时进行. 典型研究工作见表 2.

表 1 OCR 识别技术三大研究方向典型工作

研究方向	典型方案/研究成果
文字检测	EAST[14]、CTPN[15]、RRPN[16]、PSE[17]、DB[18]、SAST[19]
文本识别	CRNN[20]、STAR-Net[21]、RARE[22]、SRN[23]
端到端检测识别	Rosetta[24]、CRNN[25]、FOTS[26]、STN-OCR[27]、TextBoxes++[28]

表 2 增值税发票识别

思路	细分方法	典型方案/研究成果
先进行OCR, 后进行版面分析	先进行OCR, 后通过人为规则进行版面分析	Receipts2go[29]、文献[30]、文献[31]
	先进行OCR, 后通过对OCR结果(文字识别结果)进行建模	文献[32]、文献[33]、文献[34]、文献[35]、文献[36]、文献[37]
	先进行OCR, 后通过对OCR结果(文字识别结果及其对应的文字框位置)进行建模	CUTIE[38]、g-DICE[39]、文献[40]、文献[41]、文献[42]、文献[43]、文献[44]、文献[45]
	先进行OCR, 然后结合OCR文本结果加上发票图片, 然后输出结构化数据	文献[46]
先确定版面结构, 后进行OCR	先进行模板对齐, 然后OCR	文献[2]
	直接ROI模板对齐, 然后OCR	文献[47]、文献[48]、文献[49]
	通过表格或表格线来进行版面分析, 然后OCR	文献[50]、文献[51]
	首先通过目标检测、语义分割技术获得版面信息; 然后进行OCR	Chargrid[52]
OCR与版面分析同时进行	直接端到端方案, 输入是图片, 输出是结构化信息	EATEN[53]、TRIE[30]

表 2 中所列各种方法均只识别增值税中的部分字段, 并未识别增值税发票全票面要素, 且流程上都比较复杂, 不容易推广至其它发票的结构化识别, 且通过表 2 中的相关研究工作可知, 由于增值税版面复杂, 且真实采集的业务数据通常伴有阴影、污染、褶皱、印章、LOGO 等噪声干扰, 因此开发实际针对业务场景的增值税发票识别系统, 应尽量减少人为规则的介入. 针对已有研究在扩展性、易用性等方面的缺点, 在考虑开发算法效率、准确性的前提下, 本文提出采用业内广泛用于目标检测的 YOLO 系列[54–57] 算法中兼具检测速度和检测精度的 YOLOv4[57] 来实现增值税发票的结构化信息抽取.

2 算法概述

本文采用如图 2 所示各子流程进行增值税发票的结构化信息抽取.

从图 2 至图 9 中可知, 增值税发票结构化信息抽取主要包括以下几个步骤: S1: 关键点检测并进行模板对齐 (图 2 为原图, 图 3 为关键点检测结果, 图 4 为对齐模板, 图 5 为利用检测的关键点进行对模板齐后的

结果); S2: 票据结构化元素检测 (图 6 为元素结构化检测结果); S3: 结构化元素识别 (图 7 为结构化元素识别结果); S4: 后处理形成结构化数据 (图 8 为后处理示意图, 图 9 为最终处理结果). 下面分别介绍各个步骤的实现原理.

图 2 待处理增值税发票原图

图 3 HRNet 关键点检测示意图

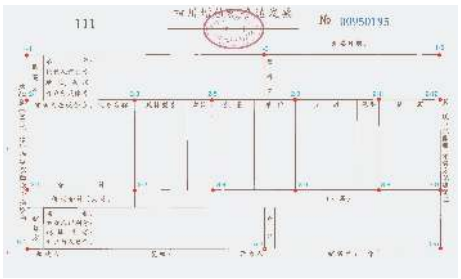

图 4 增值税发票模板图

图 5 增值税发票模板对齐图

图 6 增值税发票元素 YOLOv4 检测结果示意图

图 7 发票元素 CRNN 识别结果示意图

图 8 识别结果后处理示意图

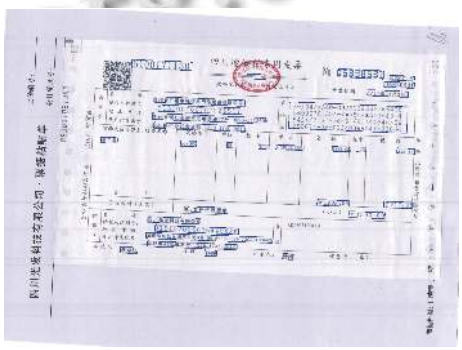

图 9 增值税发票结构化信息抽取框架

2.1 基于 HRNet 的增值税发票模板对齐

因纸质发票从开具到最终采集到系统, 中间会受到各种噪声干扰的影响, 且用户在进行报销的时候, 通常将增值税发票贴于 A4 纸 (或其它背景纸张) 上, 通过扫描仪或高拍仪成像为图像, 然后进行识别处理, 因此通常增值税发票的处理可能面临背景杂乱、摆放角

度倾斜、阴影等干扰. 针对该问题, 现行典型解决方案通常是先进行票据对齐, 先将票据对齐, 然后进行票据主体裁剪得到待处理的增值税发票.

High-Resolution Net (HRNet)[58] 为微软亚洲研究院和中国科学技术大学提出, 其主要特点是在整个过程中特征图始终保持高分辨率, 这与之前主流方法思路上有很大不同. 在 HRNet 之前, 2D 人体姿态估计算法是采用 (Hourglass[59]/CPN[60]/MSPN[61] 等) 将高分辨率特征图下采样至低分辨率, 再从低分辨率特征图恢复至高分辨率的思路 (单次或重复多次), 以此过程实现了多尺度特征提取的一个过程. 具体结构见图 10.

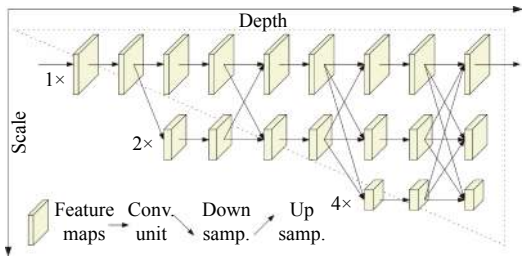

图 10 HRNet 网络结构示意图

HRNet 在整个过程中均保持特征图的高分辨率, 并通过在高分辨率特征图主网络逐渐并行加入低分辨率特征图子网络, 不同网络实现多尺度融合与特征提取. 最终所估计的关键点是在高分辨率主干网络输出.

原 HRNet 论文[58] 实验结果表明, 此网络架构能有效提升关键点检测的准确性. 本文采用 HRNet 来进行增值税表格关键点的检测. 为兼顾标注数据的标注成本, 定义表格关键点为如图 11 所示的 18 个关键点 (例如, 其中“3-4”表示为第 3 行线和第 4 列线的交叉点).

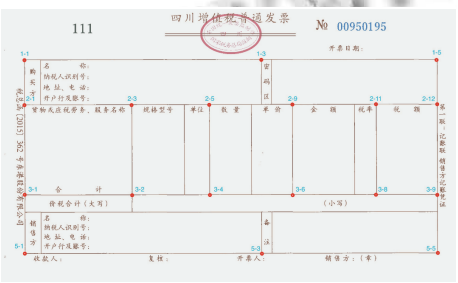

图 11 增值税 18 个关键点示意图

通过 HRNet 训练得到关键点检测网络, 对常规输入增值税图片 I, 首先进行关键点检测, 然后与标准模板中的对应基准关键点对齐, 通过检测关键点和基准关键点的对应关系找到透视变换矩阵 H[62], 最后将输入的增值税图片 I, 变换到与标准模板对应的平面上, 得到 I'. I'即为模板对齐后的增值税发票.

2.2 基于 YOLOv4 的结构化元素检测

目前, 工业界常用目标检测算法, 有 Faster-RCNN 系列[63,64], SSD 系列[65,66], YOLO 系列[54–57] 等. 其中, YOLO 系列在兼顾准确性的同时, 极大考虑了检测的实时性, 成为工业应用的首选方案. YOLOv4 为 YOLO 系列最新研究, 其网络结构的 backbone、neck、head 集成了最新的深度学习技巧, 并对数据增强模块、BN 模块、SAM 模块等进行了必要的改进, 使整个 YOLOv4 目标检测方案达到较好的性能.

通过 YOLOv4 进行结构化元素检测, 即通过 YOLOv4 对增值税发票中的元素进行检测, 得到目标框位置. 本文定义增值税发票元素如图 12 所示.

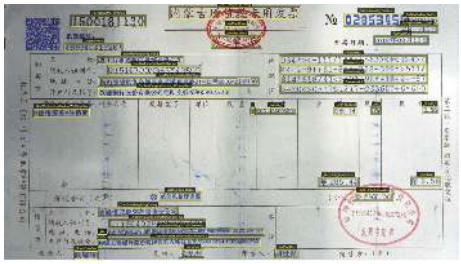

图 12 增值税票据结构化元素定义示意图

2.3 基于 CRNN 的结构化元素识别

CRNN[20] 为业界广泛使用的文本识别算法, 2015 年提出后, 迅速成为序列图片 OCR 识别的首选. CRNN[20] 的识别原理如图 13 所示.

本文采用 CRNN 识别 YOLOv4 检测的增值税发票元素.

2.4 后处理形成结构化数据

将 CRNN 识别后的数据后处理形成结构化的数据. 后处理主要是指删除部分干扰项, 合并明细项的条目, 坐标转换等, 以方便实际业务对接系统使用.

3 实验结果与分析

3.1 数据集

增值税发票涉及数据敏感性问题, 目前已知开源文献或社区中, 并未发现存在大规模真实场景的增值税发票数据集. 大规模的真实标注增值税发票的结构化元素需投入大量人力物力.

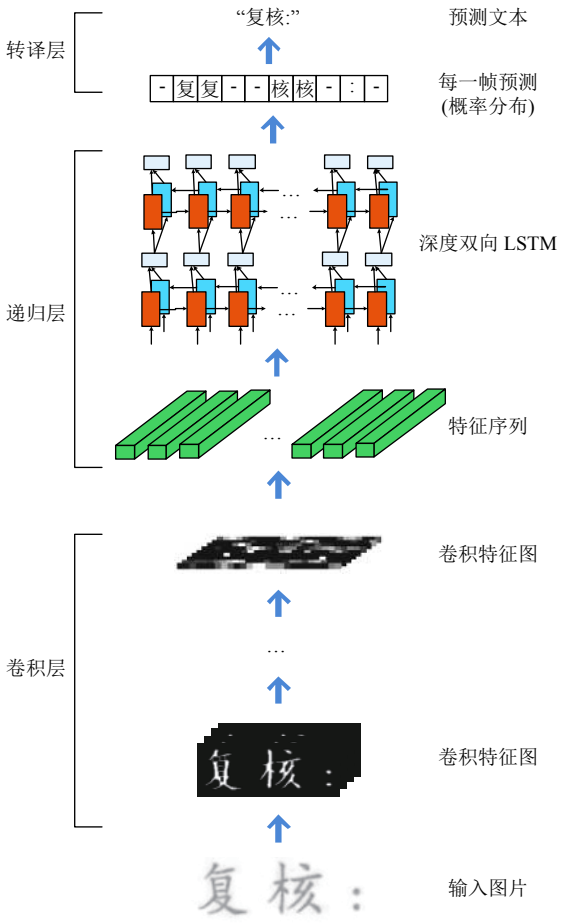

图 13 CRNN 识别架构

(1) 关键点检测数据

针对业务增值税发票, 共进行人工核对标注 3347 张 (标注样例见图 11 所示), 其中 2678 张作为训练样本, 668 张作为验证样本.

(2) 结构化元素检测数据

本文首先通过业务结构化数据 (已完成记账或报销发票的结构化存档结果, 不含位置标注信息, 只含元素文本信息), 通过人工规则及随机扰动, 合成 51 842 张增值税发票, 然后训练 YOLOv4 得到预训练模型, 基于预训练模型, 针对业务增值税发票进行检测, 形成预标注结果, 然后人工核对标注, 共标注 500 张增值税发票.

合成的 51 842 张发票, 其中 50 000 张进行训练, 1842 张进行验证. 500 张业务标注票据, 其中 400 张作为训练样本, 100 张作为验证样本. 为本文后续表述方便, 记合成的 51 482 张发票 (50 000 张训练, 1482 张验证) 为 DataA, 500 张业务数据 (400 张训练, 100 张验证) 记为 DataB.

合成增值税发票数据示意图如图 14、图 15 所示.

图 14 人工规则合成增值税发票示例 1

图 15 人工规则合成增值税发票示例 2

(3) 文本识别

针对文本识别任务, 本文并未进行单独训练, 采用百度开源模型[67] 进行发票结构化元素的识别, 若想获得更好的识别结果, 可对业务数据标注, 并对开源模型进行微调训练.

(4) 数据增强

真实采集的增值税发票通常会受到各种干扰, 因此针对关键点检测任务的训练数据, 做了裁剪、仿射变换、随机角度旋转、高斯 (均值、中值) 平滑、运动模糊、锐化、颜色反转、阴影等一系列数据增强[68], 真实数据与增强数据的比例为 1:4.

因发票元素检测任务, 是在对齐后的图片中进行的, 故 YOLOv4 中原有数据增强方式并不适合发票元素检测, 所以针对元素结构化检测这一任务, 本文调整原有数据增强为高斯 (均值、中值) 平滑、运动模糊、锐化、颜色反转、对比度拉伸、灰度化等数据增强[68].

数据增强能有效提升训练模型的泛化性能, 使其更适应真实的业务数据.

3.2 训练策略

针对关键点检测任务, 采用 HRNet-W32 网络结构, Adam 优化器、BatchSize 为 3 进行训练 (因增值税发票的关键点检测属于增值税发票结构化识别的上游

任务, 因此对其准确性要求较高, 故采用 $H\times W$=1152×1536 大分辨率图像作为网络输入的样本尺寸), 以 JointsMSELoss 作为损失函数进行训练. 损失函数如式 (1) 所示.

$$loss=\frac{1}{M\times N}\sum_{c=1}^{\#joints}\sum_{i=0,j=0}^{M,N}\left(I(i,j)-\bar{I}(i,j)\right)^2 \quad (1)$$

其中, $M\times N$代表图片的大小, $\bar{I}(i,j)$代表网络在(i,j)位置的预测输出, $I(i,j)$代表在(i,j)位置的真实标签. $\#joints$ 代表预测的关键点个数.

针对结构化元素检测任务, 首先在生成的数据中进行预训练. 然后在真实的业务数据中以预训练模型参数为网络初始参数进行微调, 得到最终模型. 预训练时采用 SGD 优化器、BatchSize 为 8 进行训练, 因为对检测结果精度要求较高, 网络模型采用 $H\times W$=1024×1024 大分辨率图像作为网络输入的样本尺寸. 其中 HRNet-W32 的训练在 3 张 Tesla-P100 16 GB 显卡上训练完成, YOLOv4 在 Tesla-P100 16 GB 单卡完成.

关于两个任务中, 其余未说明的参数 (如网络结构、学习率、迭代次数等) 设置均保持同原论文一致, 不再赘述.

3.3 实验结果

(1) 关键点检测

关键点检测的评价指标, HRNet[58] 中人体关键点检测采用 PCKh 得分 (head-normalized Probability of Corrrect Keypoint), 针对增值税发票的关键点检测已不适合. 本文对其进行适当调整为:

$$p@\alpha=\frac{\#correct}{\#total\_joints}\times 100\% \quad (2)$$

其中, $p@\alpha$ 表示关键点检测的准确率 (α=0.125); $\#correct$ 表示检测正确的关键点的数量 (同 HRNet[58] 中定义 l, l 等于关键点 1-1 到关键点 2-1 的距离, 当检测点坐标与标签对应点坐标的像素距离小于 l 时, 认为检测正确); $\#total\_joints$ 表示待检测票据中可见的总的关键点个数. 实验结果见表 3.

表 3 中, 668 张测试样本中有 13 张样本因票据残缺, 标注文件中没有 1-1 或 1-2 关键点, 所以以上测试结果为 655 张测试样本的统计结果. 根据实际经验, 当 α 小于 0.02 时, 比较接近人工标注误差; 当 α 小于等于 0.1 时, 对发票模板对齐结果不会产生影响.

表 3 关键点检测实验结果

α	$p@\alpha$ (%)	α	$p@\alpha$ (%)	α	$p@\alpha$ (%)
0.01	27.64	0.05	96.61	0.15	98.52
0.02	55.53	0.075	98.28	0.175	98.53
0.03	80.78	0.1	98.44	0.2	98.53
0.04	92.40	0.125	98.49	0.4	98.57

(2) 元素结构化检测

结构化检测实质上是一个目标检测任务, 故采用 Precision、Recall、FPS、mAP 等作为评价指标[57,69]. 结果见表 4.

表 4 元素结构化检测实验结果

数据	模型	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
DataA	YOLOv4-s	0.693	0.635	0.763	0.604	24.69
	YOLOv4-m	0.83	0.612	0.813	0.659	27.55
	YOLOv4-l	0.739	0.693	0.813	0.671	15.70
DataB	YOLOv4-s	0.585	0.66	0.67	0.458	48.78
	YOLOv4-m	0.608	0.773	0.755	0.52	32.47
	YOLOv4-l	0.494	0.807	0.757	0.558	12.85

表 4 中, mAP 是在 NMS iou-thresh 等于 0.5 情况下测得, FPS 是在 1024×1024 图片下测得.

(3) 结构化识别

结构化识别主要目的是对各个字段进行识别, 因财务记账不允许任何的字符识别错误, 因此本文不采用通用 OCR 采用的字符识别率指标, 而采用元素识别率 (element correct ratio, ecr) 作为评价指标.

$$ecr=\frac{\#correct\_elements}{\#total\_elements}\times 100\% \quad (3)$$

其中, elelemnts 表示票面上的结构化数据中的任意一个元素; $\#correct\_elements$ 表示识别正确的元素数量 (该元素的所有字符均识别正确, 则该元素识别正确); $\#total\_elements$ 表示总的元素数量. 结果见表 5.

表 5 元素结构化识别实验结果

模型	ecr (%)
YOLOv4-s	68.45
YOLOv4-m	67.95
YOLOv4-l	69.30

因结构化识别元素标注耗时长, 故本次识别率测试样本为从业务系统中随机抽选的 39 张图片. 从上述结果可以看出, 在未进行识别模型训练的条件下, 元素识别率可以达到 69.30% 的准确率, 通过对 CRNN 识别模型进行训练可以达到更好的结果

4 结论与展望

本文主要提出了一种增值税发票结构化信息识别

的方法, 为企业财务系统或记账系统提供自动化的发票识别方法, 有效减少企业运营成本.

本文采用基于 HRNet 关键点检测的方法, 进行发票模板对齐; 采用 YOLOv4 进行结构化元素的检测; 采用 CRNN 进行结构化元素的识别. 在业务数据集中的实验表明, 检测准确率在 0.5mAP 下达到 75.7, 检测速度达到 12.85 fps, 元素识别率 ecr 达到 69.30%, 本文方法在增值税发票结构化信息识别中具有较好效果. 整个系统高效简介, 人工规则介入较少, 极易拓展至相关图像结构化信息抽取领域.

本文通过 HRNet 关键点对齐降低检测难度, 通过 YOLOv4 检测发票元素, 通过 CRNN 识别发票元素的一套方法, 确实能有效简化增值税发票的识别流程, 为增值税发票结构化信息抽取提供一种简单的思路. 但本文研究尚存在一定缺陷, 一方面, 本文并未对关键点检测方面的算法进行横向对比, 且目标检测算法又分单阶段、双阶段和多阶段算法, 本文也仅从中选择比较有代表性的单阶段 SSD、FCOS 系列和双阶段 Faster RCNN 作为对比, 未能进行广泛的对比实验, 值得进一步拓展研究; 另一方面, 本文的检测在训练过程中损失函数曲线发生多次抖动, 需要进一步从模型层面、模型训练层面进行研究改进; 最后在实用时, 通常面临褶皱、印章、模糊等各种各样的干扰噪声, 如何提升模型的鲁棒性也是进一步研究的重要课题.

参考文献

1 谢志钢. 面向增值税发票的图像自动处理技术研究 [硕士学位论文]. 上海: 上海交通大学, 2015.

2 Yin Y, Wang Y, Jiang Y, *et al*. The image preprocessing and check of amount for VAT invoices. In: Liang QL, Liu X, Na ZY, *et al*. eds. Communications, Signal Processing, and Systems. Singapore: Springer, 2019. 44–51.

3 胡泽枫. 基于 OCR 的批量发票识别系统研究与实现 [硕士学位论文]. 广州: 广东工业大学, 2019.

4 胡泽枫, 张学习, 黎贤钊. 基于卷积神经网络的批量发票识别系统研究. 工业控制计算机, 2019, 32(5): 104–105, 107. [doi: 10.3969/j.issn.1001-182X.2019.05.043]

5 蒋璎. 基于深度学习的发票识别系统 [硕士学位论文]. 南京: 南京邮电大学, 2019.

6 刘欢. 基于深度学习的发票图像文本检测与识别 [硕士学位论文]. 武汉: 华中科技大学, 2019.

7 黄志文. 基于深度学习的发票自动识别系统的设计与实现 [硕士学位论文]. 广州: 广东工业大学, 2018.

8 潘妍. 票据结构化识别方法研究 [硕士学位论文]. 杭州: 浙江大学, 2020.

9 刘峰. 一种改进的自适应增值税发票字符识别方法研究 [硕士学位论文]. 湘潭: 湘潭大学, 2014.

10 武军亮. 增值税发票中有效信息的识别算法研究与实现 [硕士学位论文]. 青岛: 青岛科技大学, 2018.

11 廖玉钦. 增值税发票自动识别算法研究 [硕士学位论文]. 大连: 大连海事大学, 2018.

12 冯炳. 智能识别技术在企业信息化系统中的应用. 电子技术与软件工程, 2019, (3): 60.

13 蒋冲宇, 鲁统伟, 闵峰, 等. 基于神经网络的发票文字检测与识别方法. 武汉工程大学学报, 2019, 41(6): 586–590.

14 Zhou XY, Yao C, Wen H, *et al*. East: An efficient and accurate scene text detector. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2642–2651.

15 Tian Z, Huang WL, He T, *et al*. Detecting text in natural image with connectionist text proposal network. European Conference on Computer Vision. Amsterdam: Springer, 2016. 56–72.

16 Ma JQ, Shao WY, Ye H, *et al*. Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia, 2018, 20(11): 3111–3122. [doi: 10.1109/TMM.2018.2818020]

17 Li X, Wang WH, Hou WB, *et al*. Shape robust text detection with progressive scale expansion network. arXiv: 1806.02559, 2018.

18 Liao MH, Wan ZY, Yao C, *et al*. Real-time scene text detection with differentiable binarization. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11474–11481. [doi: 10.1609/aaai.v34i07.6812]

19 Wang PF, Zhang CQ, Qi F, *et al*. A single-shot arbitrarily-shaped text detector based on context attended multi-task learning. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 1277–1285.

20 Shi BG, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298–2304. [doi: 10.1109/TPAMI.2016.2646371]

21 Liu W, Chen CF, Wong KYK, *et al*. STAR-Net: A SpaTial attention residue network for scene text recognition. Proceedings of the British Machine Vision Conference. York: BMVC, 2016.

22 Shi BG, Wang XG, Lv PY, *et al*. Robust scene text recognition with automatic rectification. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR). Las Vegas: IEEE, 2016. 4168–4176.
23 Yu DL, Li X, Zhang CQ, *et al*. Towards accurate scene text recognition with semantic reasoning networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 12110–12119.
24 Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 71–79.
25 Li H, Wang P, Shen CH. Towards end-to-end text spotting with convolutional recurrent neural networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 5248–5256.
26 Liu XB, Liang D, Yan S, *et al*. FOTS: Fast oriented text spotting with a unified network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5676–5685.
27 Bartz C, Yang H, Meinel C. STN-OCR: A single neural network for text detection and text recognition. Computer Vision & Pattern Recognition. IEEE, 2017.
28 Liao MH, Shi BG, Bai X. TextBoxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing, 2018, 27(8): 3676–3690. [doi: 10.1109/TIP.2018.2825107]
29 Janssen B, Saund E, Bier E, *et al*. Receipts2Go: The big world of small documents. Proceedings of the 2012 ACM Symposium on Document Engineering. Limerick: ACM, 2012. 121–124.
30 Zhang P, Xu YL, Cheng ZZ, *et al*. TRIE: End-to-end text reading and information extraction for document understanding. arXiv: 2005.13118, 2020.
31 Tuganbaev D, Pakhchanian A, Deryagin D. Universal data capture technology from semi-structured forms. Eighth International Conference on Document Analysis and Recognition (ICDAR'05). Seoul: IEEE, 2005. 458–462.
32 Aslan E, Karakaya T, Unver E, *et al*. An optimization approach for invoice image analysis. 2015 23nd Signal Processing and Communications Applications Conference (SIU). Malatya: IEEE, 2015. 1130–1133.
33 Aslan E, Karakaya T, Unver E, *et al*. A part based modeling approach for invoice parsing. Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Rome: SciTePress, 2016. 390–397.
34 Minagawa A, Fujii Y, Takebe H, *et al*. Logical structure analysis for form images with arbitrary layout by belief propagation. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Curitiba: IEEE, 2007. 714–718.
35 Ha HT, Medved' M, Nevěřilová Z, *et al*. Recognition of ocr invoice metadata block types. Proceedings of the 21st International Conference on Text, Speech, and Dialogue. Brno: Springer, 2018. 304–312.
36 Nguyen MT, Phan VA, Linh LT, *et al*. Transfer learning for information extraction with limited data. Proceedings of the 16th International Conference of the Pacific Association for Computational Linguistics. Hanoi: Springer, 2019. 469–482.
37 Huang Z, Chen K, He JH, *et al*. ICDAR2019 competition on scanned receipt OCR and information extraction. 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney: IEEE, 2019. 1516–1520.
38 Zhao XH, Niu ED, Wu Z, *et al*. CUTIE: Learning to understand documents with convolutional universal text information extractor. arXiv: 1903.12363, 2019.
39 Santosh KC. g-DICE: Graph mining-based document information content exploitation. International Journal on Document Analysis and Recognition (IJDAR), 2015, 18(4): 337–355. [doi: 10.1007/s10032-015-0253-z]
40 Yi F, Zhao YF, Sheng GQ, *et al*. Dual model medical invoices recognition. Sensors, 2019, 19(20): 4370. [doi: 10.3390/s19204370]
41 Liu XJ, Gao FY, Zhang Q, *et al*. Graph convolution for multimodal information extraction from visually rich documents. arXiv: 1903.11279, 2019.
42 Sunder V, Srinivasan A, Vig L, *et al*. One-shot information extraction from document images using neuro-deductive program synthesis. arXiv: 1906.02427, 2019.
43 Palm RB, Winther O, Laws F. CloudScan-a configuration-free invoice analysis system using recurrent neural networks. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017. 406–413.
44 Bart E, Sarkar P. Information extraction by finding repeated structure. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. Massachusetts: ACM, 2010. 175–182.
45 Schulz F, Ebbecke M, Gillmann M, *et al*. Seizing the treasure: Transferring knowledge in invoice analysis. 2009 10th International Conference on Document Analysis and Recognition. Barcelona: IEEE, 2009. 848–852.
46 Palm RB, Laws F, Winther O. Attend, copy, parse end-to-end information extraction from documents. 2019

International Conference on Document Analysis and Recognition (ICDAR). Sydney: IEEE, 2019. 329–336.

47 Chien PH, Lee GC. A template-based method for identifying input regions in survey forms. Pattern Recognition and Image Analysis, 2011, 21(3): 469–472. [doi: 10.1134/S1054661811020210]

48 Peng HC, Long FH, Chi ZR. Document image recognition based on template matching of component block projections. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(9): 1188–1192. [doi: 10.1109/TPAMI.2003.1227996]

49 Sun YY, Mao XF, Hong S, *et al*. Template matching-based method for intelligent invoice information identification. IEEE Access, 2019, 7: 28392–28401. [doi: 10.1109/ACCESS.2019.2901943]

50 Tseng LY, Chen RC. Recognition and data extraction of form documents based on three types of line segments. Pattern Recognition, 1998, 31(10): 1525–1540. [doi: 10.1016/S0031-3203(98)00007-7]

51 Tanaka H, Takebe H, Hotta Y. Robust cell extraction method for form documents based on intersection searching and global optimization. 2011 International Conference on Document Analysis and Recognition. Beijing: IEEE, 2011. 354–358.

52 Katti AR, Reisswig C, Guder C, *et al*. Chargrid: Towards understanding 2D documents. arXiv: 1809.08799, 2018.

53 Guo H, Qin XM, Liu JM, *et al*. EATEN: Entity-aware attention for single shot visual text extraction. 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney: IEEE, 2019. 254–259.

54 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 779–788.

55 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6517–6525.

56 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

57 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.

58 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 5686–5696.

59 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 483–499.

60 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112.

61 Li WB, Wang ZC, Yin BY, *et al*. Rethinking on multi-stage networks for human pose estimation. arXiv: 1901.00148, 2019.

62 Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In: Fischler MA, Firschein O, eds. Readings in Computer Vision. Amsterdam: Elsevier, 1987. 726–740.

63 Girshick R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440–1448.

64 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

65 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single Shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

66 Fu CY, Liu W, Ranga A, *et al*. DSSD: Deconvolutional Single Shot Detector. Computer Vision & Pattern Recognition. IEEE, 2017.

67 Du YN, Li CX, Guo RY, *et al*. PP-OCR: A practical ultra lightweight OCR system. Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

68 Jung A, Wada B, *et al*. Corvette111/imgaug. https://github.com/aleju/imgaug. (2020-06-01).

69 Jocher G, Stoken A, Borovec J, *et al*. Ultralytics/YOLOV5: V3.0. https://github.com/ultralytics/yolov5. (2020-08-13).