文章编号：2095-2163(2021)11-0101-06 中图分类号：TP391.43 文献标志码：A

基于 OCR 技术的票据识别算法研究

王 兴，郑勇锋，严永兵，刘沿娟，张梦伊
（北京中电普华信息技术有限公司，北京 102200）

摘 要：现有方法在进行票据的识别时，需要特定设备扫描或大量的票据标签才能达到很好的识别效果。为了解决上述问题，提出了采用 OCR 技术进行票据识别算法。该算法是由 OCR 识别和 N 版本程序策略两部分组成。在 OCR 识别阶段，采用了 OCR 进行票据文字的识别，将识别后的非结构化数据转化为结构化数据。在 N 版本程序设计策略中，提出了两种算法：前者进行主关键字的匹配；后者通过选择基准，计算字符大小从而推算出其它字段。选取火车票和发票作为实验数据，广泛的实验结果证明：算法在自然场景下票据识别具有很好的结果。

关键词：票据识别；OCR 识别；N 版本程序设计策略

Research on bill recognition algorithm based on OCR

WANG Xing, ZHENG Yongfeng, YAN Yongbing, LIU Yanjuan, ZHANG Mengyi
(Beijing China-Power Information Technology Co., LTD., Beijing 102200, China)

【Abstract】 In order to achieve good recognition effect, the existing methods need special equipment scanning or a large number of bill labels. In order to solve the above problems, a bill recognition algorithm based on OCR technology is proposed. The algorithm consists of OCR recognition and N version program strategy. In the stage of OCR recognition, OCR is used to recognize the bill text, and the unstructured data after recognition is transformed into structured data. In the program design strategy of N version, two algorithms are proposed: the former to match the main keywords and the latter extrapolates the other fields by selecting a baseline and calculating the character size. Train tickets and invoices are selected as experimental data, and extensive experimental results prove that the algorithm has good results in ticket recognition under natural scenes.

【Key words】 bill recognition; OCR recognition; N version program design strategy

0 引 言

随着财务数字化快速发展，票据识别已成为计算机应用领域的热点研究问题。票总管（http://invoprime.com/）是一款软硬件结合的票据管理系统，通过票据扫描仪对发票进行快速扫描，结合票据 OCR 识别技术，就可以实现票据的识别。但此方法需要特定设备扫描才能进行票据识别，这为票据识别带来极大的不便，无法在自然场景下实现票据识别。文献[1]中把系统分为票据影像自动获取、票据识别、数据自动录入和人机数据审核 4 个模块，使用 CRNN[2]、CTPN[3]、CNN 算法构建了 OCR 识别系统。孟丹丹[4]等人通过对 Visual Geometry Group（VGG）卷积网络和双向长短时记忆网络（Long Short-Term Memory, LSTM）[5]的遗忘门和输入门进行合并改进。晏文仲[6]等人根据银行票据的印刷数字特性，进行字符的提取和分割。经过图像采集、降噪、二值化之后，使用起点直方图法结合步长法进行字符的分割后，使用改进的 LENET[7]卷积神经网络用于提取数字特征，进行分类。闫茹[8]等人利用过分割和组合过分割项得到单个字符后，使用卷积神经网络对单字符进行识别，结合语法自动机校验和模糊字符预测的大写金额文本进行识别。张振宇[9]等人提出了基于 Faster R-CNN[10]的单字检测方法，并针对中文文字排列特点优化了 RPN[10]，设置合理的字符建议区域。提出了利用 BiSRU[11]+CTC[12]网络，对定位到的字符串图像进行识别。以上算法虽然在自然场景下可以取得很好的效果，但这些方法需要大量的训练数据标签。

为了解决上述的问题，本文提出了基于 OCR 技术进行票据识别的算法。该算法由 OCR 识别和 N 版本程序策略[13]两部分组成。具体而言，识别文字阶段采用 PaddleOCR 进行票据文字的识别；N 版本程序设计策略[13]则是将识别后的非结构化数据转化为结构化数据，如图 1 所示。

作者简介：王 兴（1996-），男，硕士，助理工程师，主要研究方向：计算机视觉；郑勇锋（1981-），男，硕士，电力工程师，主要研究方向：信息化规划、建设；严永兵（1986-），男，学士，中级工程师，主要研究方向：信息化规划、建设；刘沿娟（1994-），女，硕士，助理工程师，主要研究方向：多视环境下的场景流估计；张梦伊（1980-），女，学士，高级经济师，主要研究方向：供应商全生命周期管理以及风险管控等。

收稿日期：2021-08-23

图 1 整体网络架构图

Fig. 1 Overall network architecture diagram

1 方法

首先将一张图片送入 PaddleOCR 网络中，识别出图片中的非结构化数据；非结构化数据与票据的主要关键字进行匹配，得到票据的种类；将非结构化数据送入文本匹配当中，识别出最终的票据字段。

本算法主要由 paddleOCR 和 Content Match（内容匹配）两部分组成。paddleOCR 是由百度提供的开源代码，用于识别图片中的汉字，检测出图片中的汉字以及相关的坐标信息；Content Match 实现识别票据的种类和将 PaddleOCR 识别出来的非结构化数据转化为结构化数据。

此算法主要包括文本检测、检测框矫正和文本识别等。其中，文本检测定位文本所在的位置；文本框矫正是将文本框转换成水平矩形框进行后续的文本识别；文本识别使用 CRNN 作为文本识别器。本文在训练模型时，对 3 个模型分别进行了各种技巧和模型压缩。

（1）票据的种类识别：首先选择票据中的关键词（能代表票据种类的唯一标识），然后与 PaddleOCR[2] 识别出来字段进行匹配，最后确定相应票据的种类。

（2）将非结构化数据转化为结构化数据：通过 PaddleOCR（https://github.com/PaddlePaddle/PaddleOCR）识别出来的文字是非结构化的，如图 2 所示。这些非结构化数据没有具体的语义信息，需要将非结构化数据转化为结构化数据，为解决这一问题，本文采用 N 版本程序设计策略[13]，如图 3 所示。

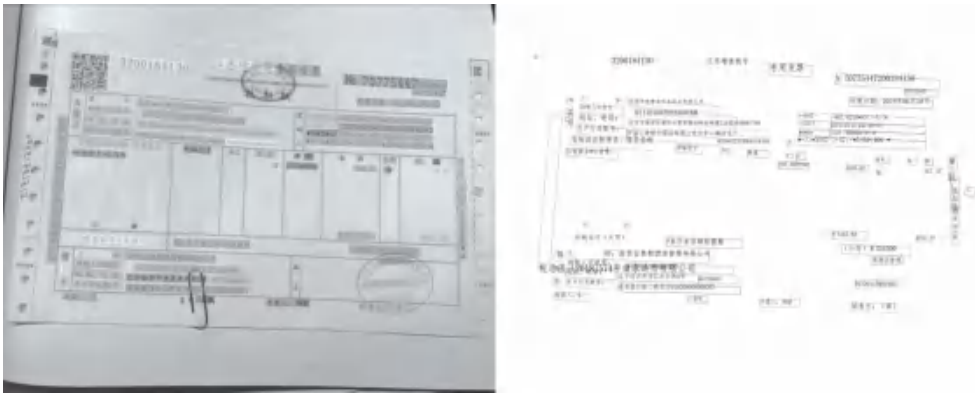


图 2 PaddleOCR[2] 识别出的非结构化数据

Fig. 2 Unstructured data identified by PaddleOCR[2]

N 版本程序设计策略采用多个弱的算法，将各自预测的结果通过表决器进行判断，最后组合成一个相对较强的算法。本文提供了两种算法：

算法一 主关键字匹配（以发票为例）

当检测的文本里面有开票日期，并且后面是包含年月日的字符串，就可以认为这些文本是开票日期。同样的，当检测的文本中，有“￥”字符时，本文就认为其可能是税额、金额或价税总额。但这种算

法存在一些问题：不同的照片由于拍摄角度的不同，光照等相关环境的影响，开票日期和含有年月日的字段不一定在同一个识别框中，这就无法通过匹配的方式得到。当匹配销售方和购买方名称时，无法判别哪个是购买方的名称，哪个是销售方的名称。因此，此算法虽然实现起来相对简单，但容易产生二义性，转化精度不高。

算法二　引入预测坐标

通过选择基准，计算字符大小和推算其它字段，最终得到识别的结果。图像坐标系建立如图 4 所示。

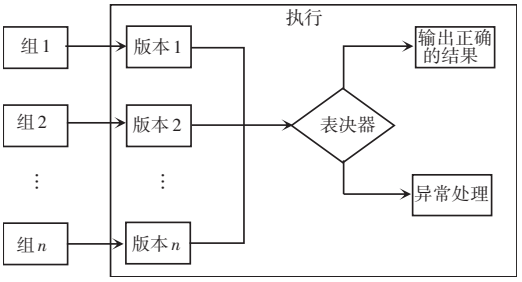


图 3　N 版本程序设计算法策略

Fig. 3　N version program design algorithm strategy

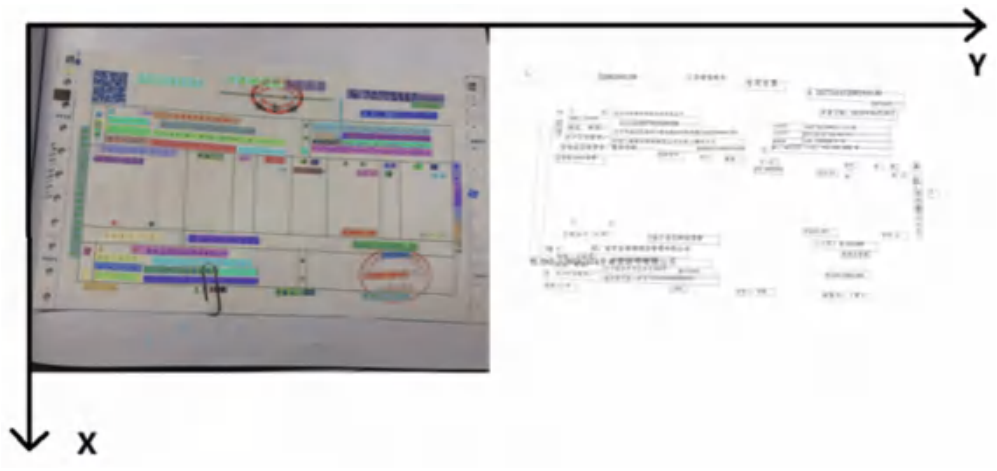


图 4　图像坐标系建立图

Fig. 4　Image coordinate system establishment diagram

算法步骤：

(1)首先选取图片中的基准，既相对容易识别的部分。例如，发票选取货物或应税劳务、服务名称这几个字符为基准，其文本检测框的 4 个坐标分别为 (x_1,y_1)，(x_2,y_2)，(x_3,y_3) 和 (x_4,y_4)。（以最左上角为第一个顶点，顺时针方式进行坐标的标注）。

(2)当已知检测文本框的 4 个顶点坐标时，通过公式得到发票中每个汉字的高度和宽度，如式(1)和式(2)所示。

$$H = \frac{(x_3 + x_4)}{2} - \frac{(x_1 + x_2)}{2} \tag{1}$$

$$W = \frac{\dfrac{(y_2 + y_4)}{2} - \dfrac{(y_1 + y_3)}{2}}{L} \tag{2}$$

其中，H 代表发票中每个字符的高度；W 代表发票中每个字符的宽度；L 代表检测框所包含的字符数。

(3)选择预测字段，并确定其所在范围后进行文字处理。以购买方开户行及账号为例，文本框左上点所在范围一定位于基准的 (x_1,y_1) 的上方，并且 x_{1n} 范围在 $0.5*H \sim 2*H$ 之间，y_{1n} 范围在 $0.5*W \sim 1.5*W$ 之间，则此范围的文字就可认为是买方开户行及账号。

(4)选择其它的字段，重复第(3)步骤。

其中：x_{1n} 代表预测字段检测框的左上坐标的 x 轴；y_{1n} 代表预测字段检测框的左上坐标的 y 轴。

2　实验与分析

为了评估本文方法的有效性，收集 53 张火车票和 31 张发票进行了实验与分析。

本文 OCR 部分采用的是由百度开源 PaddleOCR。PaddleOCR 由文本检测、文本框矫正和文本识别 3 部分组成。文本检测的目的是定位图像中的文本区域；文本框矫正则是将文本检测出的文本进行旋转和翻转到正面方向上，便于后续的文本识别；文本识别是将文本框中的文字识别出来。在内容匹配中，对于火车票，本文选取了”开”和“12306”字符串为基准。当检测到的文本有这些字符时，就认为其是火车票；同样对于发票，本文选取了“纳税人识别号”字符串为基准，当检测到的文本

有这些字符串时,则认为其是发票,否则为其他。当识别出票的种类后,采用 N 版本程序设计策略。由于算法二的精度相对较高,故表决器以算法二为基准,当算法二未检测到相关字段时,若算法一检测到相应的结果,就可以补充算法二的缺陷,从而提高最终的识别精度。

3 结果与分析

3.1 火车票识别结果

本文收集的 53 张火车票来自于网络上的爬取,涉及到了不同年份,不同光照,不同场景。下面从定性和定量两方面进行火车票识别结果的评价。

3.1.1 实验结果定性分析

本文选取了 4 张火车票进行定性分析。如图 5 所示。

从图 5 的实验结果可以看到,在自然场景下,本文的算法对不同年份的火车票都取得了一个很好的识别效果,充分证明了本文算法具有很好的鲁棒性。

3.1.2 实验结果定量分析

为了证明本文算法的有效性,对火车票中关键字段的准确率识别做了统计,结果见表 1。

从表 1 可以看出,本文的算法在涉及到不同年份、不同光照、不同场景的火车票时,各个关键字段都有一个较好的识别结果。

表 1 火车票各字段准确率

Tab. 1 The accuracy rate of each field of the train ticket

关键字段	准确率/%
始发站	89
终点站	91
车次	98
开车时间	93
金额	87
座位类型	89
姓名	91

(a) 火车票原始图片

(b) 带有文本检测框的原始图片

(c) 检测出来的非结构化数据

(d) 识别的结构化数据

图 5 火车票定性实验结果

Fig. 5 Train ticket qualitative experiment results

3.2 发票识别结果

本文收集的 31 张发票来自自然场景下的手工拍摄。下面从定性和定量两个方面进行发票识别结果的评价。

3.2.1 实验结果定性分析

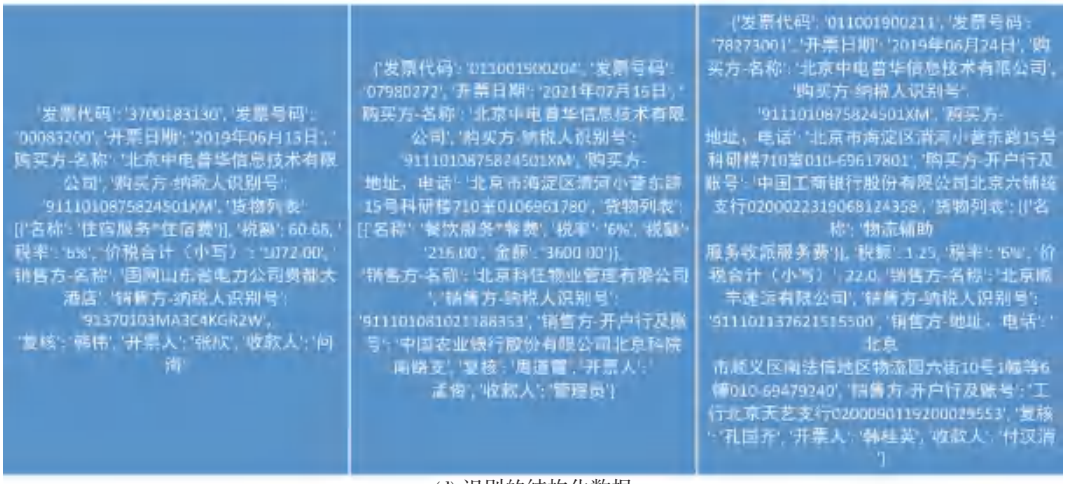
本文选取了 3 张发票进行定性评价,其分别是

增值税专用发票，增值税普通发票和增值税普通电子发票各一张，实验结果如图 6 所示。图中每一行代表一个测试样本，其中第一列代表原始的发票图片，第二列为带有文本检测框的原始图片；第三列为检测出来的非结构化数据打印在固定的位置上，第四行为识别的结构化数据。

(a) 发票票原始图片

(b) 带有评议本检测框的原始图片

(c) 检测出来的非结构化数据



(d) 识别的结构化数据

图 6　发票定性实验结果

Fig. 6　Qualitative experimental results of invoices

从图 6 中可以看出，本文的算法在自然场景的情况下，对发票的关键字段识别具有很好的识别效果。

3.2.2　实验结果定量分析

为了证明本文算法对发票的有效性，本文对发票中关键字段的准确率识别做了统计，统计结果见表 2。

表 2 发票各字段准确率

Tab. 2 Invoice field accuracy rate

关键字段	准确率/%
发票号码	100
开票日期	97
购买方-名称	97
购买方-纳税人识别号	100
购买方-地址、电话	61
购买方-开户行及账号	21
销售方-名称	97
销售方-纳税人识别号	100
销售方-地址、电话	52
销售方-开户行及账号	68
货物名称	100
货物税率	94
金额	97
开票人	97
复核人	100
收款人	94

从表 2 可以看出,本文的算法在发票号码、购买方-纳税人识别号、销售方-纳税人识别号、货物名称和复核人在 31 张发票测试中可达到 100%的识别效果。对于购买方-地址、电话、购买方-开户行及账号、销售方-地址、电话和销售方-开户行及账号的识别效果比较低,主要因为字符串较长,并且字符相对较小。从整体来看,本文的算法在发票识别上取得了不错的效果。

4 结束语

本文提出了一种基于 OCR 技术的票据识别算法。通过大量实验得到如下结论:

(1)本文将深度学习方法与传统的方法相结合,实现了自然场景下票据的识别,无需特定设备进行扫描。

(2)采用 OCR 技术+N 版本策略的方法,解决了票据识别需要大量标签的问题,无需标签就可以实现对票据的识别。

(3)利用 N 版本程序设计策略,设计了两种算法。实验结果表明,本文算法对火车票和发票都有很好的识别效果。

广泛的实验结果都证明了本文所提出的方法的有效性。本文提出的 N 版本程序设计策略中目前只有两种算法,在接下来的工作中,我们将研发更多的算法加入到 N 版本程序设计策略当中,使本文的识别精度进一步提高。

参考文献

[1] 梁林森. 基于 OCR 技术的医疗收费票据自动录入系统研究[J]. 电力设备管理,2021(4):198-199.

[2] 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.

[3] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]//European conference on computer vision. Springer, Cham, 2016: 56-72.

[4] 孟丹丹,李如玮. 基于改进 CRNN 网络的手写体票据字符识别研究[J]. 电脑编程技巧与维护,2021(4):105-108.

[5] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[6] 晏文仲,李光. 基于字符分割与新型 LENET 网络的票据识别算法[J]. 包装工程,2020,41(21):244-250.

[7] LECUN Y. LeNet-5, convolutional neural networks[J]. URL: http://yann. lecun. com/exdb/lenet, 2015, 20(5): 14.

[8] 闫茹,孙永奇,朱卫国, 等. 基于 CNN 与有限状态自动机的手写体大写金额识别[J]. 计算机工程,2021,47(9):304-312.

[9] 张振宇,姜贺云,樊明宇. 一种面向银行票据文字自动化识别的高效人工智能方法[J]. 温州大学学报(自然科学版),2020,41(3):47-56.

[10] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[11] 吴琼, 李锵, 关欣. 基于多尺度残差式卷积神经网络与双向简单循环单元的光学乐谱识别方法[J]. 激光与光电子学进展, 2020, 57(8): 081006.

[12] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 369-376.

[13] 韩炜, 臧红伟. N 版本编程技术的软件可靠性分析[J]. 微电子学与计算机, 2003, 20(5): 62-63.

(上接第 100 页)

[4] 朱洪玲,刘畅,张博,等. 激光超声可视化图像处理研究[J]. 中国激光,2018,45(1):174-181.

[5] 罗娜,李学国. 图像去雾 DCP 算法的透射率容差参数修正[J]. 科技通报,2018,34(9):218-221.

[6] 李武周,余锋,王冰,等. 基于形态学滤波的红外图像背景补偿[J]. 红外技术,2016,38(4):333-336.

[7] 张然,赵凤群. 基于双向扩散与冲击滤波的雾天图像增强算法[J]. 计算机工程, 2018, 44(10): 221-227.

[8] 贾银亮,冀凯伦,张驰宇,等. 基于暗通道先验的视频去雾算法[J]. 电子测量技术,2018,41(20):98-101.

[9] 李博. 基于视觉传达的多帧图像高分辨率重建仿真[J]. 计算机仿真,2021,38(3):113-116,121.

[10] 张迅,李建胜,王安成,等. 无人平台视觉导航算法验证仿真系统的设计与实现[J]. 测绘科学技术学报,2021,38(1):9-14,20.

[11] 李松卿,丁刚毅. 基于成像逼真度的视觉仿真方法[J]. 中国电子科学研究院学报,2021,16(1):14-20.

[12] 邓志鹏, 孙浩, 雷琳, 等. 基于多尺度形变特征卷积网络的高分辨率遥感影像目标检测[J]. 测绘学报, 2018, 47(9):1216-1227.