DOI:10.16644/j.cnki.cn33-1094/tp.2020.10.002

基于图像分析的发票识别与管理系统

杨 蕊，杨 洁

(浙江农林大学暨阳学院，浙江 诸暨 311800)

摘 要：近年来随着图像处理技术的日益发展，越来越多的应用依赖图像处理技术。为了方便财务人员进行发票管理与统计，通过对图像预处理和光学字符识别技术的研究，提出了一种采用OpenCV.js开源图形函数库和Tesseract.js光学字符识别的Web端发票识别与管理系统。图像处理与Web的结合能够在网上方便快捷地把大量纸质文件转化为电子数据，开创了一种经济新模式。

关键词：图像处理；图文识别；OpenCV.js；Tesseract.js；发票

中图分类号：TP391.41 **文献标识码**：A **文章编号**：1006-8228(2020)10-04-04

Invoice recognition and management system with image analysis

Yang Rui, Yang Jie

(*Jiyang College of Zhejiang A&F University, Zhuji, Zhejiang 311800, China*)

Abstract: In recent years, with the development of image processing technology, more and more applications rely on image processing technology. In order to facilitate the financial personnel to manage and count the invoices, by studying the image preprocessing and optical character recognition technologies, this paper proposes a web based invoice recognition and management system using OpenCV.js open source graphic function library and Tesseract.js optical character recognition engine. The combination of image processing and web can easily and quickly convert a large number of paper documents into electronic data on the Internet, which creates a new economic model.

Key words: image processing; image recognition; OpenCV.js; Tesseract.js; invoice

0 引言

随着科技的进步与时代的发展，人类进入了高速发展的数字时代。人们不再满足于传统的纸质办公，而是将需求投放在了电子文稿上。以往，人们在账本上记账，需要留存大量的票据，纸质文件的保存也极为不易；而现在，发票信息的电子录入与财务收支的统计分析变得愈发重要，纸质文件转变为电子数据的过程成为数据存储与运用的必然。

1 现状分析

目前有许多学者对发票识别进行了相关研究，于治楼等人采用BP算法对发票号码进行识别[1]，虞飞对电子发票号码进行研究[2]。但目前图像处理的大多数实践应用主要与硬件实体相结合，将图像处理技术移植到Web服务端的国内外相关实践较少[3]。

所以本文研究Web端的图像处理技术来设计一款实用、操作简单、安全的在线发票识别与管理系统，研究方向包含图像预处理模块、光学字符识别核心模块和电子数据整合模块。并且将尽可能地提高发票识别准确率[4-7]。

2 图片预处理模块

在图像处理中，图片预处理部分的好坏决定了光学字符识别部分的精度和准确度。其目的在于去除

收稿日期：2020-06-02

***基金项目**：浙江省自然科学基金探索项目(Q LQ20F020004)；浙江农林大学暨阳学院人才启动项目(JY2018RC04)

作者简介：杨蕊(1998-)，女，浙江人，本科在读，主要研究方向：计算机网络、图像处理。

通讯作者：杨洁(1989-)，男，重庆人，博士，讲师，主要研究方向：机器学习、图像处理。

图像中影响文字识别的干扰杂质，从而增强发票图像的有效信息。

本文研究的预处理模块主要分为四部分：灰度二值化、图片降噪、图片校正和图片分割。图片预处理流程图如图 1 所示。与此同时，要在 Web 端实现图片的预处理部分，需要引入 OpenCV.js 开源图形函数库进行图像的在线处理。

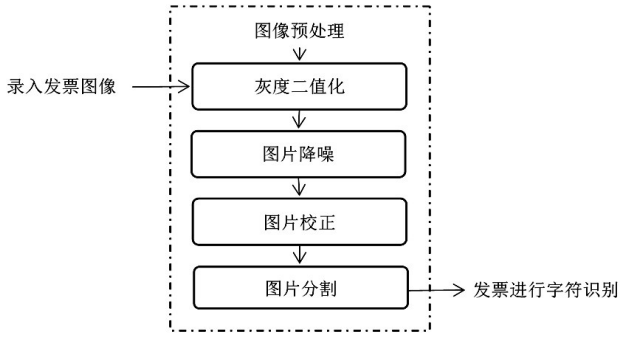


图 1　图像预处理模块

2.1 OpenCV.js 的引入

本发票识别与管理系统引入适用于 JavaScript 运用的 OpenCV.js，实现将 OpenCV 带入 Web 平台，运用 OpenCV.js 主要完成发票图片的预处理部分，包括灰度化、二值化、生成直方图、膨胀、霍夫变换。

2.2 灰度化+阈值自适应二值化技术

在图像预处理中，由于彩色图像信息量较大，计算机难以区分真实的文字和图案背景，因此需要对图像进行灰度化和二值化以达到去掉图片杂质的目的。由于人眼对颜色的敏感度不同，由式(1)可见对 RGB 进行加权平均能得到更合理的灰度图像，在 OpenCV.js 中采用 COLOR_RGBA2GRAY 的方法进行灰度化可达到该公式效果。

$$Gray = 0.299 \times R + 0.587 \times G + 0.114 \times B \qquad (1)$$

第二步进行二值化将文字与其他背景区分开，在 OpenCV.js 中有许多种二值化的阈值接口，但通常选取三种方法来定义阈值进行图片的二值化：

一是大津法（OTSU）即简单地用一种阈值将图片背景和前景一分为二。大津法又称最大类间方差法，其二值化后的前景与背景的类间方差最大[8]。二值化的过程需要取一个阈值，根据图片整体进行阈值取值的运算，记 u 为图像的总平均灰度。

$$u = w_0 \times u_0 + w_1 \times u_1 \qquad (2)$$

由式(2)可见公式中 w_0 表示前景文字所占图像的比例；$w_1 = 1 - w_0$ 为背景点数占比；u_0 表示 w_0 对应的平均灰度值；u_1 表示 w_1 对应平均灰度值。当方差 G 最大时，前景和后景错分的差别最小，此时的灰度是最佳阈值，方差 G 可由式(3)表示。

$$G = w_0 \times (u_0 - u) + w_1 \times (u_1 - u) = w_0 \times w_1 \times (u_0 - u_1) \qquad (3)$$

二是使用双峰直方图法进行图片前景和背景的分割，两峰之间的谷底所对应的灰度级即为阈值，如图 2 所示，横坐轴表示灰度值，纵坐轴表示幅值，阈值可以很明显得在曲线谷底呈现，大于阈值则视为前景黑色，小于阈值则视为后景白色。

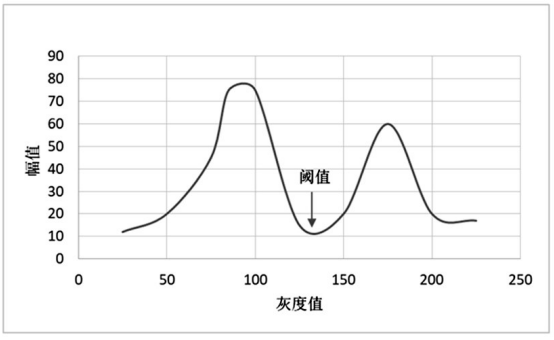


图 2　双峰直方图演示

灰度直方图清晰地概括了一幅图像的灰度级内容[9]，呈现双峰状的灰度图说明物体与背景对比明显，选用双峰直方图法因其适用于层次不同的图片。

(a) 灰度图　　(b) 自适应阈值二值化效果

图 3　自适应阈值二值化效果

三是自适应阈值二值化算法，通过 adaptive-Method 算法[10]将图像的每一个小区块按照局部特征进行阈值计算。如图3所示可以观察到经过阈值二值化处理的发票图像中的数字和文字部分变得十分清晰。

完成的效果如图3所示，自适应阈值二值化算法可以排除图像的光影和纹路干扰，但是会产生较多的颗粒噪点，后续仍需进行多次降噪处理。由于在发票管理系统的发票录入这一步骤中，发票图像的准确度受到人为因素影响较大，所以采用自适应阈值二值化算法较为合适。

2.3 图片降噪

在现实环境中，二值化图片往往图像中会出现随机分布的噪点，对后续光学字符识别（Optical Character Recognition，OCR）中图片的切割识别造成较大的干扰。为纠正颗粒噪点的干扰，通常采用以下两种方法降噪。

一是8邻域降噪法，以每一个像素点为单位，统计其周围8个像素点的灰度值为0的个数，并定义一个阈值，使得统计的个数超过阈值时标记为孤立的噪点颗粒，从而改变该噪点的灰度值，起到降噪的效果。由式(4)可见，阈值 n 为5时效果较好。

$$f(x,y)=\begin{cases}0, & threshold<0\\ 255, & threshold\geqslant 0\end{cases} \qquad threshold\in[0,8] \quad (4)$$

二是8连通泛水填充法 FloodFill，通过计算相连通的像素点面积来保留相连通的字符，从而过滤掉非字符的孤立噪点。

2.4 图片矫正

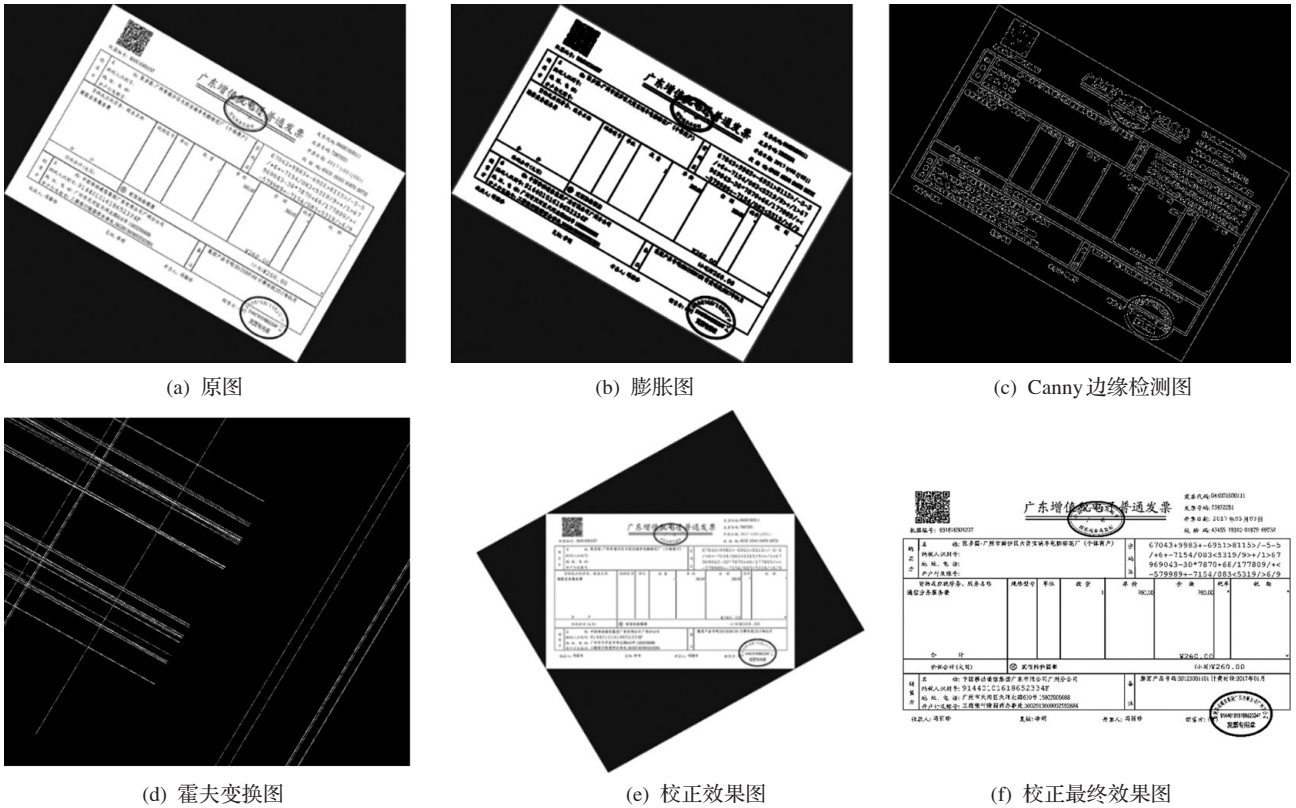
通常所拍摄的图像偶有图片倾斜的情况，此时需要计算倾斜角度将图像进行校正。

第一步进行膨胀操作将图片黑色部分的文字进行扩大，膨胀效果如图4(b)所示，黑色的文字前景区域相应扩大。

第二步进行 Canny 边缘检测，以减小图像噪点并检测出文字的边缘。Canny 边缘检测算法其效果如图4(c)所示，可以看出该算法对于文字边缘的检测效果较好。

第三步运用霍夫变换将二值化的图片进行行列检测，其阈值越大，往往精确度越高，但阈值设定过高则会导致检测效果较差，阈值过低则程序运行速度较慢。经过大量试验后，可以确定一个适合的阈值，其霍夫变换效果如图4(d)所示。

霍夫变换检测出大量直线后，计算出直线与图片边框的角度并进行旋转校正。处理过的发票图片校正效果如图4(e)所示，最终的图片可达到如图4(f)的效果。



(a) 原图　(b) 膨胀图　(c) Canny边缘检测图

(d) 霍夫变换图　(e) 校正效果图　(f) 校正最终效果图

图4　校正效果演示图

2.5 图片分割

图文识别分割的准确度决定了图文识别的准确度，通常采用投影分割法来进行分割。生成二值化的像素分布直方图如图5所示，可以看出当黑色像素点个数为0时，即为图像文字间隔的白色区域。由此可分割成不同的文字区块。

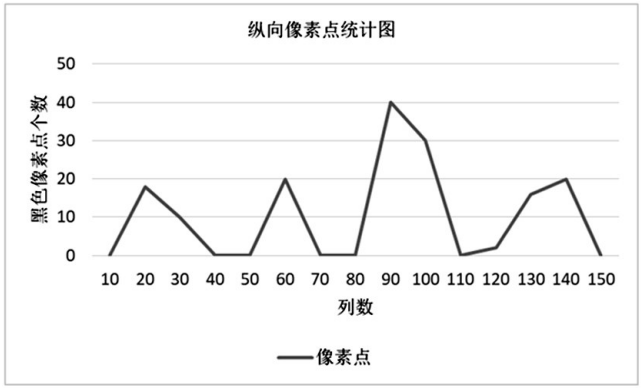


图5 纵向直方折线图模拟演示

3 字符识别核心模块

预处理后的图像主要使用Tesseract引擎来进行简单的打印字符识别，但Tesseract引擎无法识别手写字体，所以本文引入jTessBoxEditor进行Tesseract字符库训练，通过扩展字符库的方法提高识别的准确率，训练测试结果如表1所示。同时，在Web端进行字符识别时需要引入Tesseract.js进行网页端的图像处理。

表1 测试结果对比

测试方法	测试字符数/个	识别率/%
原字符库	135	74.8
训练字符库	135	90.4

其训练方法如图6所示，对手写的样本文字图片（tif或tiff格式）进行操作，生成BOX文件，然后定义字符配置文件，达到对字符矫正的目的。同时Web端进行OCR操作还需要引入Tesseract.js，可以用于在浏览器进行相关图像处理操作。

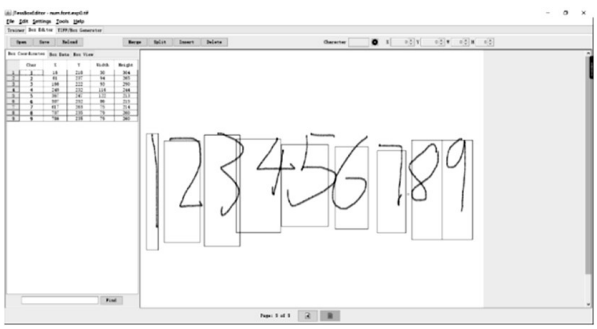


图6 训练校正

4 基于图像分析的发票识别与管理系统

4.1 办公应用功能

第一部分是注册登录功能，用户可进行登录或注册操作，并进行数据加密和验证。

第二部分是发票相册功能，用户录入的发票图像都会在这一栏目排列显示，鼠标划过任意一张发票图像即可放大查看详细的图像细节。

第三部分是图片预处理功能，用户可以自由地在图片预处理页面进行图片的灰度化、二值化、膨胀、霍夫变换等操作。

第四部分是发票支出统计分析功能，可以将所支出的发票金额自动形成柱状统计图，以直观的看出当前公司的财务支出用度情况，计算机可以自动根据数据库中的相关金额数据进行图表演示。

4.2 发票记录模块

该模块的功能有发票的录入，查看发票的录入时间、发票图像和经手人签名信息。同时在操作区域，用户可以做修改、删除和识别操作，发票记录模块效果如图7所示。

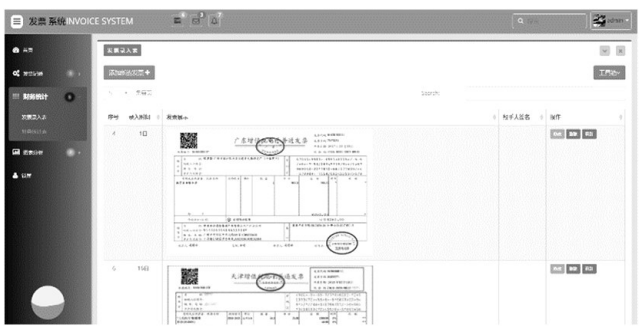


图7 发票记录模块

4.3 财务统计模块

图8 财务记录表

用户可以在左侧菜单栏“财务统计”中的“财务记录表”查看所识别的发票具体信息，包括开户名、金额、收款人、税号和发票种类五类信息，同时也可做修改和删除操作。财务统计模块效果如图8所示，该模块

可以让用户随时查看所识别过的发票信息数据。

5 结束语

为了解决发票数据人工录入的问题，本文提出了基于Web端的发票自动识别与管理系统，经过大量的图像分析测试之后，该发票识别与管理系统能较好地批量录入发票、进行发票的识别和财务数据分析，为图像处理开创了一种新模式。

在后续的研究与改进中，将会研究机器学习和卷积神经网络对发票信息进行特征识别，同时将加强该系统的安全性，保障使用者的财务数据不会被泄露。未来图像处理的市场广阔，数字化金融时代也更加离不开票据的智能化识别分析技术。

参考文献(References):

[1] 于治楼,信晓敏,黄正茂.BP算法在发票号码识别中的应用研究[J].信息技术与信息化,2014.3:113-115

[2] 虞飞.通用机打普通商业发票识别系统研究与实现[D].华南理工大学,2013.6:44-48

[3] 李光磊.Web OCR系统设计、实现及多平台对比与选型[D].天津大学,2004.

[4] PAI N, KOLKURE V S. Optical Character Recognition: An Encompassing Review[J]. International Journal of Research in Engineering and Technology,2015.4:407-409

[5] LI X G, CHEN J H, et al. English Sentence Recognition Based on HMM and Clustering[J]. American Journalof Computational Mathematics, 2013.3:37-42

[6] SHAFFIE A M, Elkobrosy G A.A Fast Recognition System for Isolated Printed Characters Using Center of Gravity and Principal Axis[J]. Applied Mathematics, 2013.4:1313-1319

[7] NAZ S, HAYAT, et al. The Optical Character Recognition of Urdu-Like Cursive Scripts[J]. Pattern Recognition, 2014.47:1229-1249

[8] 刻亚文,余孝源,周飞舟等.基于金字塔权重的QR二维码重构[J].控制工程,2020.27(4):641-647

[9] 梁华为.直接从双峰直方图确定二值化阈值[J].模式识别与人工智能,2002.15(2):253-256

[10] Ivan Ozhiganov. Deep Dive Into OCR for Receipt Recognition[EB/OL]. 2017-6-21. https://dzone. com/articles/using-ocr-for-receipt-recognition.

(上接第3页)

法，以无讼案例网中3000余例食品安全裁判文书为数据样本，进行自然语言分析，实现了非结构化文本数据中关键命名实体(人名、地名、组织机构名、食品、毒害物、危害后果)的提取，取得了比较好的效果，对于食品安全相关案件的犯罪预测预警，具有重要的理论和实践意义。下一步，在命名实体识别的基础上，继续研究食品安全实体关系抽取，进而构建食品安全知识图谱。

参考文献(References):

[1] 徐飞,宋英华.海量食品安全事件下的命名实体识别研究[J].科研管理,2018.39(7):131-138

[2] 唐剑.条件随机场模型在中文人名识别中的研究与实现[J].现代计算机(专业版),2012.21:3-7

[3] 张剑,吴青,羊昕莳等.基于条件随机场的农业命名实体识别[J].计算机与现代化,2018.1:123-126

[4] 张华平,刘群.基于角色标注的中国人名自动识别研究[J].计算机学报,2004.1:85-91

[5] 俞鸿魁,张华平,刘群等.基于层叠隐马尔可夫模型的中文命名实体识别[J].通信学报,2006.2:87-94

[6] 郭剑毅,薛征山,余正涛等.基于层叠条件随机场的旅游领域命名实体识别[J].中文信息学报,2009.23(5):47-52

[7] 叶枫,陈莺莺,周根贵等.电子病历中命名实体的智能识别[J].中国生物医学工程学报,2011.30(2):256-262

[8] 杨锦锋,关毅,何彬等.中文电子病历命名实体和实体关系语料库构建[J].软件学报,2016.27(11):2725-2746

[9] 鞠久朋,张伟伟,宁建军,等.CRF与规则相结合的地理空间命名实体识别[J].计算机工程,2011.37(7):210-212,215

[10] 李航.统计学习方法[M].清华大学出版社,2012.

[11] https://github.com/hankcs/HanLP.