一种基于结构特征的票据印刷号码识别方法

谢文彬，李新芳，郑新
（中山火炬职业技术学院，中山 528436）

摘要：目的 为保证含有号码印刷错误的票据不流入社会，研究一种新的用于票据印刷在线检测系统的号码识别方案。**方法** 提出一种基于结构特征的票据号码识别方法。先对票据图像进行采集、灰度化、二值化、去噪、倾斜校正、字符定位、单字符分割及归一化等一系列预处理。建立一种基于结构特征的号码识别分类器，再根据票据中每个号码的结构特征值，对号码进行分类识别。**结果** 实验结果表明，利用文中提出的结构特征方法，票据号码识别率达到 99%以上。**结论** 经过对大量实际发票号码的识别测试实验，该方法有较强的抗干扰性，识别算法速度快、精度高。

关键词：票据；图像处理技术；号码识别系统；结构特征识别
中图分类号：TS801.3；TP309.7 **文献标识码：**A **文章编号：**1001-3563(2018)01-0202-05

Identification Method for Invoice Printing Number Based on Structural Features

XIE Wen-bin, *LI Xin-fang*, *ZHENG Xin*
(Zhongshan Torch Polytechnic, Zhongshan 528436, China)

ABSTRACT: The work aims to research a new number identification method for the invoice printing online inspection system in order to ensure that the invoice containing number printing error does not flow into the society. An invoice number identification method based on structural features was proposed. Firstly, a series of pre-processing for the invoice image was realized, such as collection, graying, binarization, denoising, tilt correction, character positioning, single character segmentation, normalization and so on. A number identification classifier based on structural features was established, and then the number was classified according to the structural characteristic value of each number in the invoice. Based on the experimental results and by means of the proposed structural characteristic method, the invoice number identification rate reached above 99%. After a large number of actual invoice number identification tests, the proposed method has strong anti-interference ability, and the recognition algorithm is fast and the accuracy is high.

KEY WORDS: invoice; image processing technology; number identification system; structural feature recognition

发票号码作为发票中的重要信息，是票据的唯一标识，除了有防伪作用之外，在流通中也具有重要意义，不能出现漏号、重号等质量问题，若出现问题的票据流入社会，将会对国家、票据印刷企业和商家带来不良影响和严重损失[1]，因此，在票据的印制过程中，进行票据号码的自动识别对保证票据的质量变得尤为重要。近年来，随着对发票及各类印刷质量要求的提高[2—4]，传统的人工检测方式受人主观因素影响严重，极易造成错误和漏检，不能满足高效率、高质量的要求，因此，研究开发一个票据号码的在线识别系统，有助于提高生产效率，实现对发票进行客观、自动化、智能化的检测，最重要的是可以保证号码有问题的票据不流入社会，保证质量。

在号码识别领域，许多学者都进行了研究。尤其在票据和人民币方面的研究较多[4—9]。这些号码识别方法中，主要有 4 种方法[10]：模版匹配、号码结构特征、神经网络、支持向量机（SVM）等。宫义山等[11]通过模板匹配的方法实现了票据号码识别，但是其本质上还是依据号码本身结构进行匹配，未提取抽象特征。冯鑫等[12]通过对像素点逐行进行扫描实现了纸币

收稿日期：2016-10-19
基金项目：广东省科技计划（2016A010104002）；中山市科技计划（2015B2333）
作者简介：谢文彬（1983—），男，讲师，主要研究方向为包装印刷和工程技术。

号码识别，其计算量较大。王炎等[13]与苑玮琦等[14]分别采用多类特征对纸币号码进行了检测，都通过按照规则逐步查询以实现识别。ZHAI Xiao-jun 等[15]提出一种基于神经网络的 OCR 方法。FENG Bo-yuan 等[16]提出了提取号码的 Gabor 特征，而后基于 SVM 方法实现人民币字符的识别。总体上模板匹配的号码识别方法过于简单，而神经网络和 SVM 方法需要大量的号码样本，且训练过程时间偏长，而基于结构特征的号码或字符识别方法简单高效。文中的研究主要是针对发票号码的在线识别技术，通过分析票面的特征，在所设计的在线检测平台上，提取发票号码、进行一系列处理，借鉴前面研究者在人民币字符特征识别的算法，为了提高号码的识别率，文中提出一种改进的基于号码结构特征的号码识别方法，含有 14 个特征值。

1　研究对象及开发环境

文中以各类票据为研究对象，主要针对发票区域的号码部分进行识别。具体涉及到硬件系统的搭建、算法的研究及实现、软件的开发，文中仅对核心的算法原理及结果作说明。搭建了简易视觉系统，模拟印刷过程并用于动态采集票据图像。实验中选用的部分定额专用发票和普通发票见图 1。

实际票据号码识别实验时，将票据粘贴在运动带上，通过光电传感器触发进行图像拍摄。文中的号码识别算法是基于 VC++进行开发的，能够满足检测的实时性要求。

a 定额票据专用发票图像

b 通用发票

图 1　实验用的发票图像

Fig.1 Invoice image for experiment

2　识别算法总体流程及图像预处理

整个票据号码识别流程主要为：设计了一个模拟印刷实验台，首先基于图像在线采集平台，实现票据号码的实时采集；然后进行一系列的图像预处理过程，得到票据号码的单字符；最后，基于提出得一种基于结构特征的号码识别方法准确快速的识别出号码序列。具体算法流程见图 2。

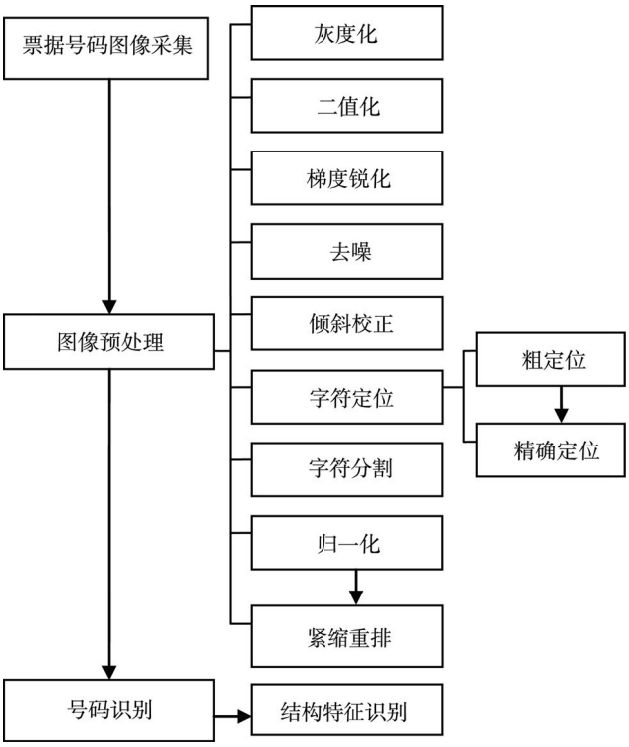

图 2　系统的基本流程

Fig.2 The basic process of the system

2.1　图像的采集及预处理过程

采用工业相机对票据图像进行采集，图片位数是 24 位彩色图片。针对号码识别的需求，将其转换为灰度图象进行处理。经过二值化、倾斜校正、字符定位、分割、归一化及紧密缩排等一系列票据图像预处理工作，最终得到每个待识别的号码的二值图像。图像的二值化原理比较简单，在实际的检测过程中，根据相机拍摄到的感兴趣区域，进行自适应阈值分割方法，结果见图 3。后文仅对核心的号码区域定位及单字符分割和号码的归一化原理和结果作简要说明。

a 灰度化图像　　b 二值化图像

图 3　二值化分割后的效果

Fig.3 The effect after binarization segmentation

2.2 号码区域定位与单字符分割

号码区域定位是识别号码体的关键步骤之一，如果定位不准确，识别过程就无法进行。文中首先根据票据号码位置的先验信息，大致粗定位出号码的区域，而后针对粗定位出的号码区域，采用水平投影和垂直穿越号码体距离的方法[17]，精确定位出号码区域。为了进一步实现号码的识别，利用垂直投影方法实现单字符的分割。具体方法为：针对精确定位出的号码区域，通过竖直的线条从左向右扫描投影图，判断扫描过程中遇见的白色像素决定号码体的起始位置和结束位置。号码区域定位及单字符分割后的效果见图 4。

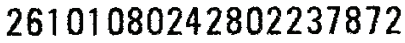

图 4 定位与分割后的号码图像

Fig.4 The number image after location and segmentation

2.3 号码的归一化

通过相机拍摄的图像中的号码在大小尺寸方面存在着一定的差异。所谓的归一化是指将单个号码字符的大小进行位置和大小的归一化，以便识别的标准性更强，准确率更高。其中，位置的归一化是将号码放在图像的中心；大小归一化是将号码规范成统一大小的图像，根据水平和垂直 2 个方向字符像素的分布缩放到统一大小，文中统一将号码归一化为 32×16 像素大小。归一化后的效果见图 5，这样有利于单个字符的结构特征值的准确，从而保证最后的识别效果。

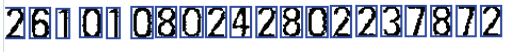

图 5 归一化后的号码图像

Fig.5 The normalized number image

3 基于结构特征的号码识别算法

对于经过一系列预处理得到的待识别号码，先对待识别号码抽取稳定的结构特征值，再用提取的特征值根据最近邻分类器进行号码的识别，因此，提取号码的结构特征是票据号码识别的关键。特征值提取的核心是选取稳定且有效的特征，其计算量相对较小，识别速度快。结构特征识别是基于号码的结构特征构造编码器，完成号码识别。根据印刷体数字的结构特征，提取数字的 4 种稳定特征：横线特征、竖线特征、水平方向过线数和垂直方向过线数。对于横线特征，在水平方向上，定义式为：

$$R_{\text{hori}} = H_{\text{blacknum}} / L_{\text{width}} \tag{1}$$

式中：L_{width} 为归一化后单个字符图像的宽度，用像素数来度量；H_{blacknum} 为该数字在水平方向上连续出现的黑色像素点数。当比值 $0.6 \le R_{\text{hori}} \le 1$，则认为数字中这些连续出现的黑像素点构成了一条横线。

对于竖线特征，在垂直方向上定义：

$$R_{\text{vert}} = V_{\text{blacknum}} / L_{\text{height}} \tag{2}$$

式中：L_{height} 为单个数字图像的高度，用像素数来度量；V_{blacknum} 为该数字在垂直方向上连续出现的黑像素点数。当比值 $0.4 \le R_{\text{vert}} \le 1$，则数字中这些连续出现的黑像素点构成了一条竖线。

号码的过线数：把数字分别水平和垂直均分成几部分，在每个部分中扫描线穿过数字，统计每条扫描线穿越号码时不相邻的交点数，在各部分得到的最大交点数定义为数字该部分的过线数。水平扫描得到的是水平过线数，垂直方向扫描数字得到各部分的垂直过线数。

依据印刷体号码这 4 种稳定的基本特征，构造分类器，对各号码正确的分类识别。显而易见：数字 0 的上下横线数为 0，左右竖线数分别为 1，水平 1/2 高度处的过线数为 2，各部分过线数均为 2。根据以上的结构特征，对 0~9 的数字进行特征抽取，可以以各号码的 4 种稳定特征构造编码器，完成对号码的识别。

考虑到仅采用这 4 种特征鲁棒性不够强，此处的鲁棒性也叫强壮性，是指针对不同的号码提取的号码特征和识别效果的稳定性。为了提高号码的识别率，通过对印刷体票据号码进行分析，在以上特征基础上，对号码的特征进行了进一步细分，除了上面提及的上横线、下横线、左竖线、右竖线这 4 种特征外，还统计了号码的其他更细的特征，具体有水平方向上的 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8 处的过线数以及垂直方向上 1/4, 1/2, 3/4 处的过线数。因此，文中设计的号码结构特征提取方法，针对每个单独号码，都包含 3 类以上特征共同组成的 14 个特征值。以数字 0 为例，主要的号码结构特征见图 6。

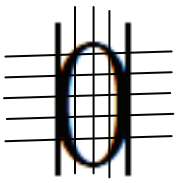

图 6 号码的结构特征

Fig.6 The structural feature of the number

根据文中所设计的号码特征统计方法，可以对票据中 0—9 共 10 个号码进行特征值的编码，具体各号码的特征编码表见表 1 竖列。后面识别的过程，对经过预处理的票据号码统计每个单字符的这 14 个结构特征值，而后基于最简单的 K 最近邻方法（KNN）和表 1 中竖列所示的 0—9 的结构特征编码值作比较，最近邻的字符值即为识别的票据号码结果，即文中基于表 1 的号码特征，使用的最近邻分类器实现票据号码识别。

表 1　各号码的结构特征值
Tab.1 Values of structural feature for every number

数字特征	0	1	2	3	4	5	6	7	8	9
上横线	0	0	0	0	0	1	0	1	0	0
下横线	0	0,1	1	0	0,1	0	0	0	0	0
左竖线	1	0,1	0	0	0	1,0	1,0	0	0	0
右竖线	1	0,1	0	0,1	0,1	0	0	1,0	0	0,1
水平1/8	2	1	1	2	1	2	2	1	2	2
水平1/4	2	1	1	1,2	1	1,2	2	1	2	1
水平3/8	2	1	1,2	1,2	1,2	1,2	2	1	2	1,2
水平1/2	2	1	1	1	1,2	1	2,1	1	1,2	1,2
水平5/8	2	1	1	1,2	2	1,2	1,2	1	2	2
水平3/4	2	1	1,2	1,2	2	1	1	1	2	2
水平7/8	2	2	2	2	2	1	2	2	2	2
垂直1/4	2	1,2	2,3	2	2	3	3	1	3	3
垂直1/2	2	1	3	2,3	2	3	3	1,2	3	3
垂直3/4	2	1	2,3	3	1	2,3	3	1,2	3	3

4　实验结果及分析

在所设计的票据号码在线采集平台上，主要针对收集到的多张陕西省西安市定额专用发票和普通发票为对象，进行了算法效果测试。其中陕西省定额专用发票 100 张，普通发票 50 张。每张发票上的票据号码都含有 20 个单号码。实验结果表明，基于文中提出的结构特征识别方法，能够很好地完成号码的识别。对于每个票据上的 20 个号码中所包含的 0—9 数字，绝大部分识别率达到 100%，个别数字的识别较差。分析识别错误的号码，主要由于获得的票据图像存在采集环境、号码印刷质量等外界客观条件的影响，对于印刷质量差的票据中的号码识别效果也较差。总体上，正常印刷没有质量问题的号码识别率在 99%以上。过去收集到的定额专用发票的号码识别结果见图 7。

a 票据图像　　b 识别结果

图 7　票据号码识别结果
Fig.7 The identification result of invoice number

5　结语

针对票据中号码的识别问题，提出了一种改进的基于号码结构特征的识别方法。首先通过对在线采集的图像进行一系列的预处理，然后针对每个号码统计其结构特征值，最后基于最近邻方法完成了整个票据号码区域的号码识别过程。文中方法速度快，识别精度高，具有一定的抗干扰性，总体上达到了预期要求。针对印刷体号码数字的识别，若想进一步提高号码的识别率和鲁棒性，结构识别上进一步提高识别的措施有：增加水平和垂直方向的扫描线数；不同的过线数给以不同的优先权。当然，为了提高识别的鲁棒性，可以将几种结构特征或不同的识别方法进行最后的决策层融合，从而可保证号码的识别率为 100%。

参考文献：

[1] 贾彦金. 票据印刷号码自动识别技术研究[D]. 西安: 西安理工大学, 2008.
JIA Yan-jin. Study on Note Number Automatic Recognition Technique[D]. Xi'an: Xi'an University of Technology, 2008.

[2] 王文举, 赵萍, 陈伟, 等. 彩色印刷品缺陷快速精确检测方法研究[J]. 包装工程, 2015, 36(17): 112—130.
WANG Wen-ju, ZHAO Ping, CHEN Wei, et al. A Fast and Accurate Method of Defect Detection of Colour Printing Image[J]. Packing Engineering, 2015, 36(17): 112—130.

[3] 谢剑斌, 刘通, 陈章永, 等. 一种基于先验信息和多模板匹配的票据水印检测算法[J]. 光子学报, 2010, 39(9): 1708-1711.
XIE Jian-bin, LIU Tong, CHEN Zhang-yong, et al. An Algorithm for Detecting Bills Watermarks Based on Prior Information andMulti-template Matching[J]. Acta Photonica Sinica, 2010, 39(9): 1708—1711.

[4] ZHAO Ya, GU Xiao-dong. Vehicle License Plate Localization and License Number Recognition Using Unit-Linking Pulse Coupled Neural Network[J]. Neural Information Processing, 2012, 7667: 100—108.

[5] 冯博远, 任明武, 张煦尧, 等. 人民币冠字号码识别预处理算法研究[J]. 计算机工程与科学, 2015, 37(6): 1148—1153.
FENG Bo-yuan, REN Ming-wu, ZHANG Xu-yao, et al. Image Preprocessing for RMB Serial Number Recognition[J]. Computer Engineering & Science, 2015, 37(6): 1148—1153.

[6] 段敬红, 栾丹. 人民币号码自动识别方法研究[J]. 计算机工程与科学, 2008, 30(1): 66—68.
DUAN Jing-hong, LUAN Dan. Research on an Automatic Number Recognition Methods for RMB Banknotes[J]. Computer Engineering & Science, 2008, 30 (1): 66—68.

[7] 刘洪刚. 纸币号码识别系统的设计与实现[D]. 长沙: 中南大学, 2007.
LIU Hong-gang. Design and Realization for Bill Serial Number Recognition System[D]. Changsha: Central South University, 2007.

[8] 刘红刚, 贺建飚. 基于模板匹配的纸币号码识别系统[J]. 计算机测量与控制, 2007, 15(8): 1077—1079.
LIU Hong-gang, HE Jian-biao. Recognition System of

Paper Currency Numbers Based on Template Matching [J]. Computer Measurement& Control, 2007, 15(8): 1077—1079.

[9] 潘虎, 陈斌, 李全文. 基于二叉树和 Adaboost 算法的纸币号码识别[J]. 计算机应用, 2011, 31(2): 396—398.
PAN Hu, CHEN Bin, LI Quan-wen. Paper Currency Number Recognition Based on Binary Tree and Adaboost Algorithm[J]. Journal of Computer Applications, 2011, 31(2): 396—398.

[10] TRIE O D, JAIN A K, TAXT T. Feature Extraction Methods for Character Recognition-A Survey[J]. Pattern Recognition, 1996, 29(4): 641—662.

[11] 宫义山, 王鹏. 基于模板匹配的发票号码识别算法[J]. 沈阳工业大学学报, 2015, 37(6): 673—678.
GONG Yi-shan, WANG Peng. Identification Algorithm for Invoice Number Based on Template Matching[J]. Journal of Shenyang University of Technology, 2015, 37(6): 673—678.

[12] 冯鑫, 吴庆洪. 一种基于结构特征的纸币号码识别方法[J]. 辽宁科技大学学报, 2013, 36(4): 385—388.
FENG Xin, WU Qing-hong. One Paper Currency Character Recognition Method Based on Structural Feature [J]. Journal of University of Science and Technology Liaoning, 2013, 36(4): 385—388.

[13] 王炎, 刘洋, 宋百春. 人民币纸币号码识别算法研究[J]. 计算机工程与科学, 2013, 35(8): 103—108.
WANG Yan, LIU Yang, SONG Bai-chun. Research on RMB Paper Currency Number Recognition Algorithm [J]. Computer Engineering & Science, 2013, 35(8): 103—108.

[14] 苑玮琦, 金灿. 基于结构特征的纸币号码识别方法[J]. 计算机工程与应用, 2014, 50(8): 118—121.
YUAN Wei-qi, JIN Can. Paper Currency Number Recognition Method Based on Structural Features[J]. Computer Engineering and Applications, 2014, 50(8): 118—121.

[15] ZHAI Xiao-jun, BENSAALI F, SOTUDEH R. Real-time Optical Character Recognition on Field Programmable Gate Array for Automatic Number Plate Recognition System[J]. IET Circuits Devices Systems, 2013, 7(6): 337—344.

[16] FENG Bo-yuan, REN Ming-wu, ZHANG Xu-yao, et al. Automatic Recognition of Serial Numbers in Bank Notes[J]. Pattern Recognition, 2014, 47(8): 2621—2634.

[17] 郭艳平, 丁万山. 基于投影法定位和分割的美元号码识别系统[J]. 航空计算技术, 2007, 37(5): 45—48.
GUO Yan-ping, DING Wan-shan. An US Dollar Number Recognition System for Location and Division Based on a Projection Method[J]. Aeronautical Computing Technique, 2007, 37(5): 45—48.