ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 119－125
DOI:10.16337/j.1004-9037.2017.01.014

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于数字结构特征的发票号码识别算法

崔文成　任　磊　刘　阳　邵　虹

(沈阳工业大学信息科学与工程学院,沈阳,110870)

摘　要: 由于印章覆盖、发票折痕等干扰因素的存在,一些发票号码区域会出现噪声粘连区域,这些区域会导致发票号码无法正常分割。针对这一问题,提出了噪声粘连区域修复算法,有效地避免了该情况对数字分割的影响。针对普通发票号码的字体结构和特点,提出了基于数字结构特征的发票号码识别算法。首先定义数字结构特征,包括4种填充区域、2种字符穿越数和4种镂空区域,构成待识别数字的10维特征向量;进而与标准模板库中数字进行模板特征匹配,求得距离最小值所对应的数字作为识别结果。将所提出的方法和基于改进的左右轮廓特征的印刷体数字识别方法进行对比,实验结果表明,本文所提出的识别算法拥有更高的准确率和更快的识别速度,以及对噪声有更强的鲁棒性。

关键词: 发票号码识别;噪声粘连区域;数字结构特征

中图分类号: TP391　　**文献标志码:** A

Invoice Number Recognition Algorithm Based on Numerical Structure Characteristics

Cui Wencheng, Ren Lei, Liu Yang, Shao Hong

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang, 110870, China)

Abstract: Interference factors such as seal cover, invoice crease and so on, cause noise adhesion in number area of some invoice, which would seriously lead to the invoice number segmentation error. Aiming at this problem, a noise adhesion area repairing algorithm is proposed. At the same time, according to the font structure and characteristics of ordinary invoice number, invoice number recognition algorithm based on characteristics of digital structure is proposed. Firstly, define number structure features, including four kinds of fill area, two kinds of number of passing through the character, and four kinds of hollow area, which constitute a 10-dimensional feature vector of the number to be identified. Then, match the feature vector with the template features in the standard template library, by obtaining the Euclidean distance, and regard the corresponding number with the minimum Euclidean distances as the last recognition result. The proposed method and printed number recognition method based on the improved left and right contour features are compared. Experimental results indicate that the proposed identification algorithm has higher accuracy, faster recognition speed and stronger robustness to noise.

Key words: invoice number recognition; noise adhesion area; numerical structure characteristics

---

基金项目: 辽宁省自然科学基金(201202162)资助项目;辽宁省高等学校优秀人才支持计划(LJQ2013013)资助项目。
收稿日期: 2015-05-20;**修订日期:** 2015-06-19

引　言

随着信息时代的高速发展，许多学校和公司都在施行无纸化办公，努力提高办公自动化程度。在财务部门，发票报销往往是比较基本的业务，不可避免要进行发票相关信息的录入，以供存档和后续审批处理。发票号码是税务部门给予发票的编码，是发票的唯一标志，使用统一字体进行印刷。发票号码一般是 8 位，在网上查询发票真伪时，一般要在 8 位发票号码前面输入 12 位发票代码。所以，人们经常提到的发票号码是由 8 位发票代码和 12 位发票号码组成的 20 个数字。工作人员输入大量发票号码极其耗时耗力，人们越来越希望计算机能够代替手工输入，对字符进行自动识别并输入。数字识别属于字符识别的范畴，是计算机对自然数 0～9 这 10 个数字的识别。借力于国内外广大科研人员和学者的潜心钻研，大量识别算法层出不穷，同时应用在不同领域。数字识别大致可以划分为两类：基于全局的统计分析和基于结构的特征分析。基于全局的统计分析大多数应用于模板匹配、特征点和像素点密度等。比如：文献[1]利用局部对比平均法从二值化图像中提取人民币字符；文献[2]将每个数字图像分为非重叠的分区，将每个分区的平均灰度值作为识别的特征矢量，识别钞票的序列号；还有运用反向传播算法训练后的神经网络模型，对煤气表数字进行识别[3]。但基于全局的统计分析的计算量相对较大，对噪声的适应性和对字体形状变化的鲁棒性不好。基于结构的特征分析考虑数字的轮廓和字符形状，包含图像预处理模块、特征提取模块和识别模块[4]。比如：文献[5]基于字符的假想线与相交特征点特征，构建识别判断树完成识别；文献[6]提取各种手写体数字轮廓结构的方向信息，用于检测像素之间的变化，并统计数字图像中的横线数来完成识别。但是基于结构的特征分析在提取特征的过程中，没有关注数字本身的结构特征，导致特征提取过于复杂，所提取特征的紧凑型和区分性还有待增强；同时，该算法完成了不用应用场景下字符的识别，识别对象的字体和图像质量存在着不同程度的差异性，不能直接应用到发票号码识别。针对上述识别算法的不足，本文深入研究了不同数字的结构和形状特征，提出了基于数字结构特征的发票号码识别算法，提取特征数量少，又能很好地区分不同数字，采用更加简单易行的方式完成特征提取过程，从而进行发票号码识别。

1　图像预处理

本文的研究重点是发票号码识别算法，因此要锁定发票号码区域。将如图 1 所示的普通发票的号码区域进行定位，得到如图 2 所示的号码区域图片，将图 2 的两行号码图像作为图像预处理对象。图像预处理的主要目的在于有效地避免噪声和光照亮度不均的影响，提高发票号码区域图像的清晰度。首先对图 2 号码区域图像进行灰度化，采用 3×3 模板的自适应中值滤波器对灰度化图像进行去噪处理，因为传统中值滤波器只能有效去除空间密度小的噪声，而自适应中值滤波器可以处理大概率噪声，并平滑非冲激噪声。然后采用形态学处理方法，选用大小为 3×3 的结构元素对图像先后进行开运算和闭运算，对图像进行平滑去噪处理。开运算先对图像进行腐蚀运算，然后对处理结果进行膨胀运算，可以有效去除数字周围的细小噪声点，平滑数字的边界。闭运算则是先膨胀后腐蚀，填充字体本身细小空

图 1　普通发票图片

Fig. 1　Ordinary invoice picture

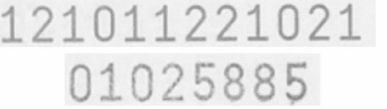

图 2　12 位的发票代码与 8 位的发票号码图片

Fig. 2　Images of 12 bit invoice code and 8 bit invoice number

洞,同时也可以平滑边界。最后利用自适应阈值分割技术[7]求解图像二值化的阈值,对完成去噪处理的图像进行二值化,有效区分目标和背景;由于二值化转化后的图像目标为白色,背景为黑色,所以要在图像二值化之后对图像进行反色处理。门限处理对图像二值化是普遍使用的方法,找到合适的分割阈值是二值化的关键。自适应阈值分割技术通过试探的手段来逐步逼近最终的门限值,具体的算法流程如下。设定随机值 $T'=\mathrm{random}(0.255)$,以 T 为阈值,令 $T=T'$,将图像分割成两部分 G_1,G_2,然后计算两部分的平均灰度值 u_1,u_2。令 $T'=(u_1+u_2)/2$,如果 T 与 T'之间的绝对值之差小于事先的指定值,则图像最终灰度分割阈值定为 T,否则就再令 $T=T'$,重新分割图像[7]。图像二值化效果如图 3 所示。

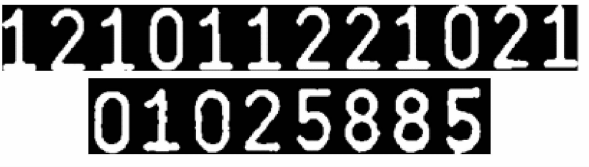

图 3 自适应阈值分割后的二值化图像

Fig. 3 Binary image after adaptive threshold segmentation

2 数字分割

虽然大多数发票的号码区域图像足够清晰,运用传统的投影法[8]就可以顺利分割出单个数字,但是仍然有一些发票,如图 4 所示,其发票号码区域被印章覆盖或者存在折痕,降低了图像清晰度,图像预处理很难滤除这些噪声,导致面积较大的噪声粘连区域被分割出来,在对二值化图像进行数字分割过程中会出现如图 5 所示的现象。

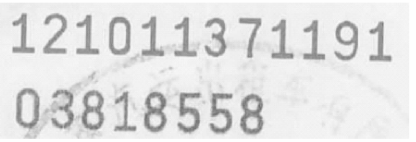

图 4 被印章覆盖的发票号码图像

Fig. 4 Invoice number image covered with seals

噪声粘连区域是指存在噪声点并导致数字粘连的区域。如何判断是否存在噪声粘连区域,以及如何处理这样的区域,成为数字分割环节的难点。针对这一问题,提出了噪声粘连区域修复算法,在使用投影法分割数字之前,先对噪声粘连区域进行修复,避免印章痕迹对数字分割带来的影响,具体算法如下:

(1)首先计算单个数字最小连通面积,进行统计对比分析,确定最小连通面积 $S_{\min}$。

(2)将 $S_{\min}$作为判断图片非数字的噪声区域的门限值,图像中连通面积低于 $S_{\min}$的区域判定为噪声粘连区域。实验表明,$S_{\min}$取值 160 时,噪声粘连区域判定最为准确。

(3)对噪声粘连区域进行反色处理,从而消除分割过程中的噪声粘连区域和去噪环节无法去除的杂点,从而进行正常分割。

针对如图 4 所示的发票图片,先对其进行噪声粘连区域修复,再运用投影法进行分割,最后利用双线性插值法将分割好的单个数字图像按照 20×40 尺寸标准进行大小归一化,最终结果如图 6 所示。

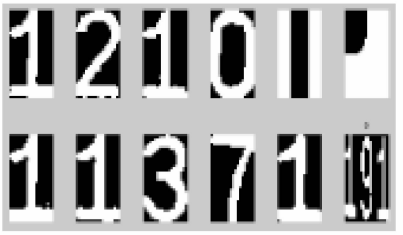

图 5 噪声粘连区域

Fig. 5 Noise adhesion area

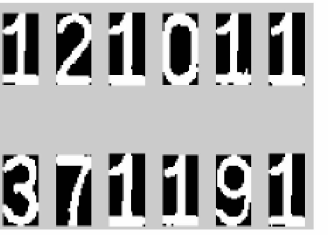

图 6 大小归一化的数字分割结果

Fig. 6 Segmentation results of size normalized number

3 数字识别

首先通过对识别数字进行结构特征的提取，包括顶部、底部、左侧和右侧填充区域，水平和垂直字符穿越数，左上、右上、左下和右下镂空区域，构成 10 维的特征向量；接着运用特征匹配方法，把待识别数字通过特征提取构成的特征向量和标准模板库中 0～9 的 10 个特征向量分别求取两者的欧式距离，取得 10 个欧氏距离的最小值，识别结果就是最小值对应的标准模板库中的数字。

3.1 填充区域

在单个数字图像中，定义在某一固定区域内，如果字符像素占据该区域绝大部分，则称该区域为填充区域。按照方向对填充区域进行分类，在水平方向上，划定顶部填充区域和底部填充区域；在垂直方向上，划定左侧填充区域和右侧填充区域。

(1)水平方向

$$H_L = h_w \div w \tag{1}$$

式中：w 为切割完成后数字图像的宽，也就是每一行的像素数之和；h_w 为每一行白色像素点不间断出现的次数；H_L 表示每一行白色像素点所占的比例。当 $H_L \in [0.75, 1]$ 时，可以确定在该行存在一条由数字像素点组成的横线。

在大小为 20×40 的图片中，定义水平方向前 5 行像素为顶部区域，水平方向最后五行像素为底部区域。在顶部区域内，通过逐行判断像素点颜色，计算每一行白色像素点所占比例 H_L。若有连续三行 $H_L \in [0.75, 1]$，则判定该数字具有顶部填充区域。同理，在底部区域，逐行扫描像素点，若有连续三行 $H_L \in [0.75, 1]$，则判定该数字具有顶部填充区域。以数字“1”“5”为例，由图 7 可知，根据上述的描述，“1”有底部填充区域，“5”有顶部填充区域。

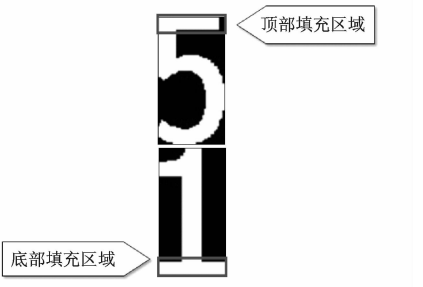

图 7 顶部填充区域和底部填充区域

Fig. 7 Top and bottom filling areas

(2)垂直方向

$$V_L = v_w \div h \tag{2}$$

式中：h 为切割完成后数字图像的高，也就是每一列的像素数之和；v_w 为每一列白色像素点不间断出现的次数；V_L 表示每一列白色的像素点的所占比例。当 $V_L \in [0.6, 1]$ 时，可以确定在该列存在一条由数字像素点组成的竖线。

在大小为 20×40 的图片中，定义垂直方向左边 5 列像素为左侧区域，右边 5 列像素为右侧区域。

在左侧区域内，通过逐列扫描像素点，计算每列中白色像素点所占的比例 V_L。若存在连续 3 列 $V_L \in [0.6, 1]$，则判定该数字具有左侧填充区域。同理，在右侧区域，逐列扫描像素点，若有连续 3 列 $V_L \in [0.6, 1]$，则判定该数字具有右侧填充区域。如图 8 所示，“0”同时具有左侧填充区域和右侧填充区域。

3.2 字符穿越数

当每条扫描线穿越白像素区域边界时，有黑白像素的跳变或者起始位置为白色像素点，这样的情况判定为扫描线与字符相交。将每一条扫描线与白色区域相交的次数定义为字符穿越数。按照扫描方向划分，字符穿越数包括水平字符穿越数和垂直字符穿越数两种。结合发票号码字体的特殊性，本文对穿越数的计算进行调整，定义数字图片下半部分的字符穿越数为水平字符穿越数，数字图片右半部分的字符穿越数为垂直字符穿越数。水平字符穿越数的计算方法是水平扫描数字图像下半部分的所有行，若发生黑白跳变或者首个像素点为白，均视为与扫描线相交，并记为一个相交次数。对比所有扫描行的相交次数，确定相交次数的最大值，该值则为水平字符穿越数，如图 9 所示。

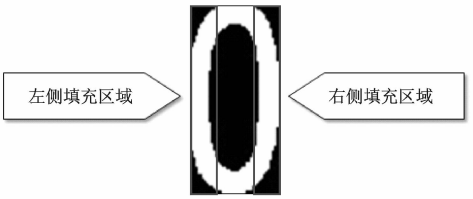

图 8 左侧填充区域和右侧填充区域

Fig. 8 Right and left filling areas

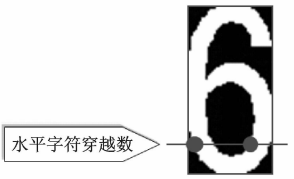

图 9 水平字符穿越数

Fig. 9 Number of horizontal passing through characters

垂直字符穿越数的计算方法是垂直扫描右半部分所有列,若发生黑白跳变或者首个像素点为白,均视为与扫描线相交,确定所有的相交次数最大值,该值则为垂直字符穿越数,如图 10 所示。

3.3 镂空区域

镂空区域是指数字图片某固定区域内至少有两行像素是全黑色像素。按照左上、右上、左下和右下四个方向,把单个数字图像平均划分成四个区域,判断这四个固定区域是否可以称作镂空区域。镂空区域的判定方法是在数字图像等分的 1/4 区域内,扫描每行像素点,若超过两行全部为黑色像素则判定为镂空区域。因此可以分为 4 个镂空区域:左上镂空区域、右上镂空区域、左下镂空区域和右下镂空区域。图 11 展示了几个典型数字的 4 类镂空区域。

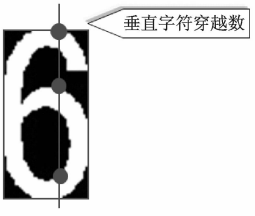

图 10 垂直字符穿越数

Fig. 10 Number of vertical passing through characters

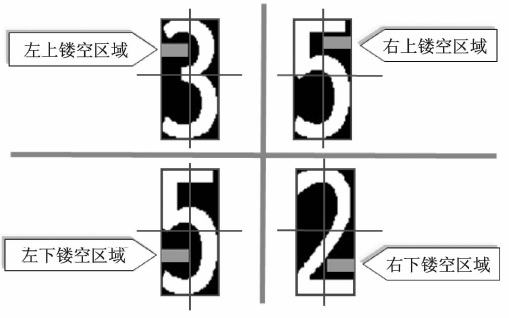

图 11 四类镂空区域

Fig. 11 Four types of hollow areas

3.4 特征匹配与识别

通过对发票号码的标准数字模板库的数字 0～9 进行结构特征提取,汇总标准数字 0～9 模板的特征,如表 1 所示,存在填充区域或镂空区域用数值 0 表示,否则用数值 1 表示。对待识别号码的多结构

表 1 数字 0～9 的特征

Tab. 1 Characteristics of number 0～9

数字	0	1	2	3	4	5	6	7	8	9
顶部填充区域	0	0	0	0	0	1	0	1	0	0
底部填充区域	1	1	1	0	0	0	0	0	0	0
左侧填充区域	1	0	0	0	0	0	1	0	0	0
右侧填充区域	1	0	0	0	0	0	0	0	0	1
水平字符穿越数	2	1	1	2	1	2	2	1	2	1
垂直字符穿越数	2	1	3	3	2	3	3	3	3	3
左上镂空区域	0	0	1	1	1	0	0	1	0	0
右上镂空区域	0	0	0	0	0	1	1	0	0	0
左下镂空区域	0	0	0	1	1	1	0	0	0	1
右下镂空区域	0	0	1	0	0	0	0	0	0	0

特征提取之后，为其建立 1 个 10 维特征向量，得出它与表 1 中标准模板库中 0～9 这 10 个数字特征向量之间的 10 个欧式距离。取 10 个欧氏距离中的最小值，识别结果就是该最小值对应的标准库中的数字。

4 实验结果与分析

将发票号码正确识别率和运行时间，作为算法的评价标准。与基于统计特征的数字识别方法[9]进行对比，该方法首先假设每个字符存在于一个矩形框里，在框里设定出 3 条特征线，然后统计通过线上像素点的变化次数来提取每个字符的特征值。在单个数字的二值化图像中，分别从数字横向 2/5 和 2/3 处以及字符纵向 1/2 处作扫描线，分别命名为 X_1，X_2 和 Y，统计 3 条扫描线上数字变化的次数，得到 3 个穿越次数特征，初步将数字分为 8 类；因为总计 10 个数字，被归为同一类的两个数字可以再根据第 1 个发生变化的像素所在列与 Y 的位置关系来判断，最终识别出 10 个数字。

在程序运行设备、运行软件版本、识别对象、图像预处理和数字分割环节完全一致的情况下，对比两种识别算法的识别率和运行时间，从而对算法做出客观公正评价。实验使用计算机的基本信息如表 2 所示，软件运行环境是 Matlab R2013a 版本。

表 2 实验过程所使用计算机的基本信息

Tab. 2 Basic computer information used in experiment

类别	参数指标
操作系统	Window7 旗舰版
处理器	Inter(R)Core(TM) i5-3230M CPU@2.60 GHz
安装内存(RAM)	8.00 GB(7.86 可用)
系统类型	64 位操作系统

识别对象是经过定位后的发票号码区域的 JPG 格式图片，如图 2 所示。具体包括 200 张 12 位数字的发票代码图片和 200 张 8 位数字的发票号码图片。实验数据对比如表 3 所示。根据表 3 的实验对比数据，可以看出本文提出的基于数字结构特征的发票号码识别算法识别率更高，同时运行时间也更短。在特征提取环节，本文方法更能体现发票号码的结构特点，区分度更大；算法的整体流程更加简单易行，执行效率也更高。

表 3 实验数据对比

Tab. 3 Comparison of experimental data

识别算法	基于统计特征的识别算法	本文算法
识别数字/个	4 000	4 000
正确识别数/个	3 685	3 873
误识个数/个	315	127
正确识别率/%	92.13	96.83
运行时间/s	17.50	13.71

5 结束语

发票号码识别过程中，数字分割和识别算法是两个较为关键的环节，其效果的好坏直接影响最终的识别效果。在进行数字分割的过程中，本文针对发票号码区域被印章覆盖影响分割的问题，提出了噪声粘连区域修复方法，有效地去除了印章痕迹的影响，保证了发票号码区域图片可以顺利分割成单个数字

图片。对于发票号码识别算法,特征提取是最为重要的环节,成功的特征提取方法要求所提取特征能体现数字之间的差异性,同时又能保证算法的可行性。本文提出了基于数字结构特征的发票识别算法,提取填充区域、字符穿越数和镂空区域这些具有较大区分性的结构特征,不仅避免了特征冗余,而且所提取特征简单又具有代表性,更能体现 10 个数字之间的差异性。但所提出的发票识别算法利用计算欧氏距离来进行模板特征匹配,是相对传统的匹配方法。如何在不提高算法复杂性的同时,使用更加适合的特征匹配方法,进一步提高该算法的正确识别率,将是后续研究的工作重点。

参考文献:

[1] Feng Boyuan, Ren Mingwu, Zhang Xuyao, et al. Extraction of serial numbers on bank notes [C]// 12th International Conference on Document Analysis and Recognition. Washington, DC: IEEE, 2013: 698-702.

[2] Gai Shan, Yang Guowei, Zhang Sheng, et al. New banknote number recognition algorithm based on support vector machine [C] // 2nd IAPR Asian Conference on Pattern Recognition. Naha: IEEE, 2013: 176-180.

[3] Li Pei, Li Chaofeng, Ju Yiwen, et al. A new method for recognizing digital numbers on coal gas meters [C] // 6th International Congress on Image and Signal Processing. Hangzhou, China: IEEE, 2013:469-473.

[4] 王静娇,孙晶,周玉冰,等. 基于 TMS320DM642 的人民币图像特征识别系统[J]. 数据采集与处理,2012,27(S2):206-211.
Wang Jingjiao, Sun Jing, Zhou Yubing, et al. RMB image feature identification system based on TMS320DM642[J]. Journal of Data Acquisition and Processing,2012,27(S2): 206-211.

[5] Cao Xinyan, Ma Lin. A recongnition system of real time paper currency[C]// 2nd International Conference on Computer Science and Network Technology. Changchun: IEEE, 2012:198-201.

[6] Lee S W, Wu H C. Effective multiple-features extraction for off-line SVM-based handwritten numeral recognition [C] // 3rd International Conference on Information Security and Intelligent Control. Yunlin, Taiwan, China: IEEE, 2012:194-197.

[7] 陈明华. 印刷体数字识别算法研究[D]. 武汉:华中科技大学,2012:12-14.
Chen Minghua. Study on printed numeral recognition [D]. Wuhan: Huazhong University of Science and Technology, 2012: 12-14.

[8] Li Yueqin, Li Jinping, Han Lei, et al. A bank note number automatic identification method[C]// International Conference on Environment Science. Melbourne: IEEE, 2012:185-192.

[9] 高振斌,赵盼,王霞,等. 印刷体数字识别系统的 FPGA 实现[J]. 重庆邮电大学学报:自然科学版,2015,27(2):213-218.
Gao Zhenbin, Zhao Pan, Wang Xia, et al. Printed digit recognition system based on field programmable gate array [J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2015, 27(2): 213-218.

作者简介:

崔文成(1973-),男,讲师,研究方向:智能信息处理,E-mail:576022085@qq.com。

任磊(1990-),男,硕士研究生,研究方向:智能信息处理。

刘阳(1965-),男,副教授,研究方向:视频及图像处理、虚拟现实技术。

邵虹(1974-),女,教授,研究方向:图像处理与模式识别、智能信息处理。