发票印刷体数字识别方法的研究

邵 虹,王 佳
(沈阳工业大学 信息科学与工程学院,辽宁 沈阳 110870)

摘 要:在发票图片的采集过程中,由于拍摄不当获取到的图片存在倾斜;受采集环境的影响,采集到的发票图片表面有光照不均匀,部分区域过亮或过暗,不利于数字的定位与识别。针对此类问题,在预处理阶段,首先应用霍夫变换法检测发票图片中的横线并计算其倾斜角,通过旋转对倾斜发票图片进行矫正;其次,对发票图像进行预处理操作,减弱光照以及噪声的影响;接着研究了普通发票版面特征以及数字分布位置,提出了一种基于投影法的定位方法,准确定位出数字区域;最终选用基于数字结构特征的方法判别数字。实验结果表明,该算法识别速度快、精度高。

关键词:预处理;霍夫变换;发票号码;数字定位;号码识别
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2016)07-0173-04
doi:10.3969/j.issn.1673-629X.2016.07.037

Research on Recognition Method on Invoice Printing Number

SHAO Hong, WANG Jia
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: In the process of invoice collection, some pictures may have the tilt problem due to inappropriate photograph method. In addition, non-uniform illumination may happen as a result of the photograph environment, which adds to the difficulty of location and identification of invoice numbers. In view of these problems, Hough transform has been applied firstly in the preprocessing section to detect and calculate the tilt angle of the picture, then rotating to correct invoices. Secondly, preprocessing has been arranged for the invoices to eliminate the effect of noise. Thirdly, a method of location based on projection is proposed by the research on the layout of common invoices, which can identify the number area accurately. Finally, the algorithm based on digital structure feature is adopted to identify. The experiment shows that this algorithm has high identification velocity and precision rate.

Key words: preprocessing; Hough transform; invoice number; digital positioning; digital recognition

0 引 言

随着信息的快速发展,数字世界变得越来越明显。光学字符识别是模式识别领域中的一个重要研究领域。通过前期的努力,这一领域已经取得了丰硕的研究成果。

数字识别是光学字符识别的一个重要研究方向和组成部分,它仅利用计算机就能自动识别阿拉伯数字 0 到 9。是一种有效、可靠、快速的数字识别系统,不仅可以作为单独使用的软件,也可用于识别车牌号码系统以及智能安防系统,具有非常重要的商业价值。因此,数字识别的研究吸引了众多研究者,产生了许多识别算法和研究成果。

文献[1]介绍了银行支票识别系统的基础,针对同时出现的大小写数字,提出多分类器融合算法和人工神经网络算法,分别实现大写和小写数字的鲁棒识别;文献[2]阐述的支票识别系统针对对象中线条特点,给出一种快速线段检测算法和基于特征线检测的单据识别算法。

很长一段时间,单据管理工作由人工完成。在许多企业和政府机构、医院、保险行业,账单处理是沉重和繁琐的手工劳动。如果能够利用计算机自动处理这些发票,从发票印制的数字信息中自动提取,实现数字的精确识别,那么就能减少由于输入数据投入的人力和物力。

收稿日期:2015-11-10 **修回日期**:2016-03-09 **网络出版时间**:2016-06-22
基金项目:辽宁省自然科学基金(201202162)
作者简介:邵 虹(1974-),女,教授,CCF 会员,研究方向为图像处理与模式识别;王 佳(1990-),男,硕士,研究方向为计算机图形学与虚拟现实。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160622.0845.056.html

1 发票的版面结构

普通商业机打发票作为发票类别中最常见也最重要的一类,发票抬头、日期等信息位于发票的上部,不同类别的数据分别列示于下部矩形框内。文中课题有效的信息是发票的数字区域,在发票图片的右上角,其分布形式为上下两行打印。上部是发票代码,12 位印刷体数字,下部是发票号码,8 位印刷体数字。

2 发票图片的预处理

图像的预处理主要是指在图像二值化前对图像所进行的处理工作。由于采像环境的变化和采像设备的影响,会出现不同程度的分布有噪声和亮度不均匀以及图像倾斜的情况,这会严重影响后面的处理效果。图像的预处理可以有效消除噪声、光照反射等情况的不良影响,增强图像中的有效信息。因此,选择适合的预处理方法将对后续二值化、识别工作带来很大的便利[3]。

2.1 图片的倾斜矫正

获取到的发票图片可能存在倾斜,这样的图片会影响后续的数字定位和识别,因此需要对获取到的所有图片进行倾斜检测。霍夫变换是图像处理中的特征提取技术,它是检测一种特定形状的对象投票算法。具体来说,霍夫变换检测图像中的直线,利用双坐标空间的变化,将空间中的相似形状的直线线性映射到另一个坐标空间的点,并将检测直线的问题转化为统计峰值问题[4]。

对于图像中的一条直线而言,在直角坐标系中可以表示为:$y = kx + b$。该直线上任意一点 (x,y) 变更到 $\rho - \theta$ 参数空间将转换为一个“点”,也就是说,将原有空间中所有非零像素变动到 $\rho - \theta$ 参数空间,那么它们将聚焦在一个点上。所以,参数空间中的局部峰值点对应于原始图像空间中的直线。由于这条线的斜率可以无限或无穷小,然后在 $\rho - \theta$ 参数空间是不容易刻画和描述的,因此该线的检测采用极坐标参数空间。在极坐标系中,直线可以表述为以下形式:$\rho = x\cos(\theta) + y\sin(\theta)$。利用霍夫变换检测直线,思路如下:每一个点假设有 n 个方向的直线,通常 $n = 180$,检测角度精度为 1°,分别计算这 n 条直线的 (r,theta) 坐标,得到 n 个坐标点。如果要判断的点共有 N 个,最终得到的 (r,theta) 坐标有 $N \times n$ 个。关于这些坐标,其中 θ 是一个离散的角度,有 180 个值,如果有一行多点,那么必须有一个 θ 等于一定值 theta_i ,和这些点的 R 等于 r_i 。那么,这许多个点都在同一条直线上。

利用霍夫变换方法检测发票图片中的水平最长直线,然后计算出这条最长直线的斜率,最后实现图片的水平矫正,达到预期的目标。具体的算法步骤包括以下几点:

(1)读入彩色图片并进行灰度化。

(2)截取整个发票图片的右上角部分,将检测区域缩小。对发票的研究发现,号码所在位置在整幅图片的上 1/3 到上边界,右 3/4 到右边界部分。缩小直线检测区域将减少处理数据量,提高程序运行速度。

(3)对该区域进行 Roberts 边缘检测,获取边界,滤除竖线,检测所有横线,计算长度并标出最长直线。

(4)计算斜率,矫正图片。设 A、B 两点分别是直线的起点和终点,坐标分别为 (x_1,y_1)、(x_2,y_2)。通过式(1)计算出直线的斜率,然后求出倾斜角 θ,将原始图片逆时针旋转 θ 角度,实现矫正。

$$\tan(\theta) = \frac{y_2 - y_1}{x_2 - x_1} \tag{1}$$

图 1 就是对其中一幅图片进行倾斜矫正的过程图。其中,第一张图片为读入的原始图片,为彩色图片;第二张是局部灰度图,截取的是发票的固定右上角部分,缩小检测范围,提高了检测速度和准确性;第三张是 Roberts 边缘检测的结果,并标注出最长直线;第四张是矫正图,通过最长直线计算出倾斜角度,再旋转原图,得到矫正后的图像。

图 1 图片的倾斜矫正

实验程序总共对 100 张图片进行测试,其中正确的有 100 张。倾斜矫正正确率达到了 100%。

2.2 图片的噪声处理

图像的噪声主要来源于图像采集和传输两阶段。噪声应极大程度地消除,消除噪音可以获得图像的真实数据。消除噪声的方式有很多种,大抵可分为两大类:一类是空间域方法,应用不同的模板算子对原始图像做卷积运算处理,抑制或消除噪声;另一类是频率域方法,把原始图像从空间域转变到频率域,再采用适当的各类滤波器对其进行滤波,经反变换后得到去噪后的图像[5]。

实验中选取空间域处理方法,选用非线性滤波器

中的中值滤波器,由于在实际运算过程中它并不需要图像的统计特征,所以具有简单与便利的优点,能够消除线性滤波器引起的图像细节模糊难题。实验中对发票图片进行中值滤波,采用 3×3 大小的模板。

2.3　发票数字的定位

在滤除噪声之后,发票号码数字需要被精确定位并截取为数字串存储下来,以便于后续的分割和识别操作。仔细观察发票版面后发现,感兴趣的发票号码数字位于发票图像中最大矩形框的上方,以这个矩形框右上角的顶点为基准点,分别向左、向上截取合适的宽度和高度,构成一个特定的矩形把两行数字包含起来,实现对数字的初次定位,最后将两行数字进行水平切割,将发票代码和号码分别储存起来。实验过程如下:首先对整幅发票图片作预处理操作,对二值图像作水平投影,选取水平投影图中左侧波峰中最大的位置,记录其行位置 x_0;其次对二值图像作垂直投影,选取垂直投影图中右侧波峰中最大的位置,记录其列号位置 y_0;最后得到基准点的坐标(x_0,y_0)。由于发票上的号码数字是规范化印刷体,其每个数字的宽度和高度大小都是固定值,故可以统计出 12 个数字的宽度和高度,统计所有发票图片,最后宽度采用 Width/5,高度采用 Height/5。Width、Height 分别是原始图像的宽度和高度。

3　数字分割

数字图像分割是把每个图数字串的数字分开,使它成为一幅单一的数字图像。这里,如果数字分割的准确率很高,那么对后续的单个数字提取特征将非常有利。数字分割算法有很多,实验使用投影法的数字分割法,其过程有以下几点:先竖直投影,找出每个数字的左右边界,分割出单个数字;其次再对每一个数字进行水平投影,找出其上下边界,至此,每一个数字都被一个最小外接矩形包围,也就是数字分割工作完成;最后,将每一个数字做归一化,使其所有单个数字大小一致。这里进行归一化所采用的方法为双线性插值法,将数字归一化到 40×80 大小[6]。图 2 为采用投影法分割数字的效果图。

图 2　数字分割结果

4　数字识别

在过去的数十年中,研究者们提出了各种各样的识别方法,如神经网络法[7]、模板匹配法[8]、基于数字结构特征的识别算法[9-10]、基于组合特征的识别算法[11]等。

4.1　基于穿越号码次数的结构识别算法

该算法[12]使用的特征是:航程(包括上、下、左、右航程)、穿越号码体次数(水平和垂直)、第一次穿越号码体空体航程、长横和长竖。识别方法根据结构特征采用逐级判断的方法:

(1)字符宽度小于最大字符宽度 1/3 的为“1”;上航程面积大于右航程面积设定值的为“4”;下航程面积大于上航程面积设定值的为“7”。

(2)左、右、上、下航程面积几乎多为零的可能是“0,6,8,9”,水平穿越上半部分一次的为“6”;水平穿越下半部分一次的为“9”;垂直穿越中部两次的为“0”;垂直穿越中部大于等于三次的为“8”。

(3)左航程面积大于右航程面积设定值的为“3”;左航程面积等于 1/2 下半部分左航程面积且右航程面积等于 1/2 上半部分右航程面积的为“5”;另一个为“2”。

4.2　基于结构特征的号码识别算法

该算法[13]使用的特征是:水平、垂直方向穿线数。把数字从上到下平均分成 8 部分,在每部分中分别以水平方向扫描线从左到右穿过数字,计算每条扫描线穿越黑像素区域互不相邻的交点数,统计每部分的最大交点数。在上 $i/8$(i =1,2,3,4)部分的最多交点数定义为该数字上 $i/8$ 高度处的过线数;在下 $i/8$(i =1,2,3,4)部分的最多交点数定义为该数字的下 $i/8$ 高度处的过线数。同理可得该数字的左 $i/8$(i =1,2,3,4)宽度处的过线数和右 $i/8$(i =1,2,3,4)宽度处的过线数。从 10 个数字中寻找稳定而有效的特征来构造编码器,如表 1 所示,根据编码器识别印刷体数字。

表 1　编码器

数字特征	0	2	3	4	5	6	7	8	9
上 1/8	1	1	1	1	1	1	1	1	1
上 2/8	1,2	1	2,1	1	1	1,2	1	1,2	2,1
上 3/8	2	2	2	1	1	2	1	2	2
上 4/8	2	2,1	1	1	1	1,2	1	2,1	1,2
下 4/8	2	1	2,1	2,1	1,2	1,2	1	1,2	1
下 3/8	2	1	2,1	1,2	1,2	2	1	1,2	2,1
下 2/8	2,1	1	1,2	1	2	2	1	2,1	1,2
下 1/8	1,0	1	0,1	1	1	1	1	1,0	0,1
左 1/8	1	2,1	2,1	1	1,2	1	1	2,1	1,2
左 2/8	1	2	2	1	2	1	1,2	2,1	2
左 3/8	1	2,3	2	1	2	1	2,1	1	2
左 4/8	2	3,2	3	2,1	3,2	2,3	2	3,2,1	3
右 4/8	2	3,2	3	2,1	3	3	1,2	3	3
右 3/8	2,1	2,3	1,2	1	3,2	3,2	1	1,2,3	1,2
右 2/8	1	2,1	1,2	1,2	2	2	1	2,1	1
右 1/8	1	2,0,1	2	1,0	2,1,0	2	1	2,1	1

注:上 1/8 代表上 1/8 穿线数,以此类推。

4.3 基于数字结构特征的数字识别算法

特征提取的关键是选取稳定且有效的结构特征，提取不同的特征，识别率不同。实验中提取的结构特征有：上横线、下横线、水平交线个数以及垂直交线个数。

横线，指水平扫描号码体，如果存在某行连续为号码体像素的数目超过号码体宽度的三分之二，则定义为横线。根据横线所处的不同位置可分为：上横线，即横线位于数字的顶部；下横线，即横线位于数字的底部。

交线个数，指水平或垂直扫描号码体，以像素为单位不同位置的穿线个数可能不同，一个像素位置有一个穿线次数的结果，但一般为一次、二次或三次，选取指定位置出现最多的次数，定义为交线个数[14]。

识别过程如下：首先，数字 5 和 7 在顶部有上横线特征，通过上横线特征分类出数字 5 和 7；数字 1 和 2 在底部有下横线特征，在剩余的所有数字中分类出数字 1 和 2；对其余的 6 个数字 0、3、4、6、8、9，通过水平交线个数和垂直交线个数两种特征进行分类。对于数字 0 和 8，在水平 1/3 处和 2/3 处的交线个数都是 2，但是数字 0 的垂直交线个数为 2，数字 8 的垂直交线个数是 3，通过垂直交线个数分类出数字 0 和 8；对于数字 4 和 9，在水平 1/3 处和 2/3 处的交线个数分别均是 2 和 1，但数字 4 的垂直交线个数为 2，数字 9 的垂直交线个数为 3，通过垂直交线个数分类出数字 4 和 9；对于数字 3 和 6，数字 3 在水平 1/3 处和 2/3 处的交线个数分别是 1 和 1，数字 6 在水平 1/3 处和 2/3 处的交线个数分别是 1 和 2，2/3 处水平交线个数为 2 的是数字 6，否则是数字 3。

数字识别流程图如图 3 所示。

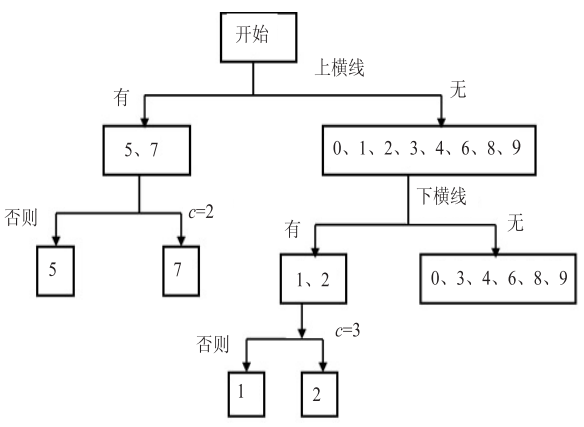

图 3 数字识别流程图

对于数字 0，3，4，6，8，9 应采用两种交线特征来进行识别。表 2 为特征编码表。其中，a 为 1/3 处水平交线个数；b 为 2/3 处水平交线个数；c 为垂直交线个数。

表 2 特征编码表

数字	a	b	c
0	2	2	2
8	2	2	3
3	1	1	3
6	1	2	3
4	2	1	2
9	2	1	3

5 实 验

5.1 实验结果

实验使用的 PC 机基本信息如下：

操作系统：Windows 7 旗舰版；

处理器：AMD 速龙 X2 ql-64 2.10 GHz；

内存：2 GB；

系统类型：32 位操作系统。

通过测试 20 张 12 位数字的发票代码图片和 20 张 8 位数字的发票号码图片，总计 400 个数字，其中正确识别的数字有 394 个，识别率达到 98.5%。

5.2 实验分析

实验对比分析见表 3。

表 3 实验对比分析

指标	基于穿越号码次数的结构识别算法	基于结构特征的号码识别算法	文中实验方法
识别个数	400	400	400
正确个数	380	392	394
错误个数	20	8	6
识别率/%	95	98	98.5
运行时间/s	1.651 8	1.431 2	1.131 8

由表 3 可以看出，实验所用的识别算法识别率高于另外两种数字识别算法，并且运行时间短，识别速度快，具有明显的优势。由于提取的特征减少，计算量相应减少，识别效率有所提高，识别率也较高。

6 结束语

实验对图像的获取、图像的预处理、数字分割以及数字识别四个步骤进行了研究和分析。在图像的获取阶段，是用普通 300 W 像素摄像头拍摄，图片中会存在不同程度的噪声及倾斜。在预处理阶段，应用霍夫变换检测水平直线，计算直线斜率，计算图像的倾斜角度，再逆时针旋转图片，实现图像的倾斜校正。使用中值滤波法对图像进行噪声滤除，极大减弱了噪声影响，图片更加清晰，有利于后续操作。对感兴趣数字区域进行定位，定位出识别所需的数字，为后续数字的分割和识别做技术准备。基于投影法的数字分割结果非常理想，将数字串分割为单一的数字；基于数字结构特征

（下转第 182 页）

的基于局部搜索的离散人工蜂群算法来解决云制造服务组合问题。实验仿真证实了 LSDABC 的有效性及可行性。然而,基于 QoS 的评估模型仅仅考虑了服务的技术指标,忽略了服务使用者的感受,因此仍有待改进。所以下一步的研究工作将集中于兼顾用户体验质量综合考虑,优化云制造服务组合路径。

参考文献:

[1] 李伯虎,张 霖,柴旭东. 云制造概论[J]. 中兴通讯技术,2010(4):5-8.

[2] 陶 飞,张 霖,郭 华,等. 云制造特征及云服务组合关键问题研究[J]. 计算机集成制造系统,2011,17(3):477-486.

[3] Xiang F,Hu Y,Yu Y,et al. QoS and energy consumption aware service composition and optimal-selection based on Pareto group leader algorithm in cloud manufacturing system [J]. Central European Journal of Operations Research,2014,22(4):663-685.

[4] 向 峰. 云制造系统中基于能耗的服务组合关键技术研究[D]. 武汉:武汉理工大学,2013.

[5] Huo Y,Zhuang Y,Gu J,et al. Discrete gbest-guided artificial bee colony algorithm for cloud service composition[J]. Applied Intelligence,2015,42(4):661-678.

[6] 刘卫宁,李一鸣,刘 波. 基于自适应粒子群算法的制造云服务组合研究[J]. 计算机应用,2012,32(10):2869-2874.

[7] 敬石开,姜 浩,许文婷,等. 考虑执行可靠性的云制造服务组合算法[J]. 计算机辅助设计与图形学学报,2014,26(3):392-400.

[8] Tao F,LaiLi Y,Xu L,et al. FC-PACO-RM:a parallel method for service composition optimal-selection in cloud manufacturing system[J]. IEEE Transactions on Industrial Informatics,2013,9(4):2023-2033.

[9] Xia Y M,Cheng B,Chen J L,et al. Optimizing services composition based on improved ant colony algorithm[J]. Jisuanji Xuebao (Chinese Journal of Computers),2012,35(2):270-281.

[10] Fan Y,Zhao D,Zhang L,et al. Manufacturing grid:needs,concept,and architecture[M]//Grid and cooperative computing. Brilin:Springer,2004:653-656.

[11] Tao F,Hu Y F,Zhou Z. Study on manufacturing grid & its resource service optimal-selection system[J]. The International Journal of Advanced Manufacturing Technology,2008,37(9-10):1022-1041.

[12] Tao F,Zhao D,Hu Y,et al. Resource service composition and its optimal-selection based on particle swarm optimization in manufacturing grid system[J]. IEEE Transactions on Industrial Informatics,2008,4(4):315-327.

[13] 秦全德,程 适,李 丽,等. 人工蜂群算法研究综述[J]. 智能系统学报,2014,9(2):127-135.

[14] 张 平,刘三阳,朱明敏. 基于人工蜂群算法的贝叶斯网络结构学习[J]. 智能系统学报,2014,9(3):325-329.

(上接第 176 页)

的识别算法识别速度快,识别率高。

实验中还存在一些问题:第一,通过拍摄或扫描的图片清晰度各异,经过图像预处理后仍有不清晰现象;第二,由于特征提取算法本身的问题,提取的特征可能出现误差。

参考文献:

[1] 林 强. 基于 OCR 的支票识别系统的研究与实现[D]. 北京:北京邮电大学,2010.

[2] 李 琥,卜佳俊,陈 纯. 一种新的基于特征线检取的票据识别算法[J]. 浙江大学学报:工学版,2003,37(2):173-177.

[3] 严国莉,黄 山,李岱璋,等. 印刷体数字快速识别算法在身份证编号数字识别中的应用[J]. 计算机工程,2003,29(1):178-179.

[4] Zhang Zongjian,Chen Guanghua,Li Jianwei. The research on digit recognition algorithm for automatic meter reading system [C]//Proceedings of the 8th world congress on intelligent control and automation. Jinan,China:[s. n.],2010:5399-5403.

[5] Li Yueqin,Li Jinping,Han Lei,et al. A bank note number automatic identification method[C]//Proc of international conference on environment science. Melbourne:IEEE,2012:185-192.

[6] 徐 哲,楼文高. 基于模版对比的手写体数字识别神经网络模型[J]. 计算机工程与应用,2008,44(9):226-228.

[7] 戴 静,胡钊政,白建川. 一种基于交点特征的印刷体数字识别方法[J]. 电视技术,2014,38(13):28-30.

[8] 高 菊,叶 桦. 一种有效的水表数字图像二次识别算法[J]. 东南大学学报:自然科学版,2013,43(S):153-157.

[9] 滕书华,孙即祥,邵晓芳. 一种鲁棒性的印刷体数字识别算法[J]. 光学与光电技术,2005,3(6):12-15.

[10] 倪桂博,梁晓尊. 基于结构形状的印刷体数字识别方法[J]. 软件导刊,2010,9(5):67-68.

[11] 张翼成,陈 欣,杨红军,等. 基于组合特征的 BP 神经网络数字识别方法[J]. 计算机系统应用,2013,22(3):113-116.

[12] 李春宇. 金融发票印刷体数字及面值识别方法的研究[D]. 沈阳:沈阳工业大学,2006.

[13] 邬建瓴. 数字识别及其应用[D]. 武汉:华中科技大学,2006.

[14] 徐 敬,刘 炜. 基于特征矩阵的高效数字识别算法[J]. 软件导论,2014,13(1):59-61.