基于 HALCON 的发票号字符识别研究

张 有 陈晓荣
(上海理工大学 上海 200093)

摘 要：为了解决人工方式识别发票号效率低下、费时费力的问题，以 HALCON 软件为平台，对发票号进行识别，首先，对读取的图像进行图像预处理，包括把原图像转化成单通道图像、图像增强、高斯滤波、阈值分割和连通区域分割，然后对预处理后的图像进行数学形态学处理，接着把粘连的数字拆分开，最后用基于多层感知分类器(MLP)对发票号数字进行识别。通过与人工方式的方法对比，结果表明，用基于 HALCON 的方法，发票号数字能被快速、精确的识别出来，有效的减少了人员的工作量，提高了工作效率。

关键词：HALCON；高斯滤波；粘连；多层感知分类器(MLP)
中图分类号：TN957.52 **文献标识码**：A **国家标准学科分类代码**：510.4050

Character recognition research of invoice numbers based on HALCON

Zhang You Chen Xiaorong
(Shanghai University of Science and Technology, Shanghai 200093, China)

Abstract: In order to solve the problem of low efficiency by manual way of identification, this paper uses HALCON software as a platform to identify the invoice number. Firstly, the image preprocessing is carried out, which includes transforming the original image into single channel image, image enhancement, Gaussian filtering, threshold segmentation and connected region segmentation; Secondly, the mathematical morphology of the pretreated image is processed, And then ,the figures of the split; Finally, the number of the invoice number is identified by MLP (multilayer perceptual classifier). By comparing with the manual method, the results show that the invoice number can be quickly and accurately identified by the method based on HALCON, which can effectively reduce the workload of the staff and improve the efficiency of the work.

Keywords: HALCON; Gaussian filtering; adhesion; multilayer perceptual classifier (MLP)

1 引 言

在实际生活中，越来越多行业银行、学校、公司财务等的业务涉及到发票单据中发票号的识别，比如银行、学校、公司财务等业务，而传统的发票号识别工作一直由人工方式来完成的，发票数据的基数太多，人员的工作面临和存在着任务繁重、效率低下等一系列问题。如果能够利用机器视觉软件，采用数字图像处理的方式代替人工处理操作，实现对发票号数字的、快速、自动、准确地识别，从中提取出版面中的印刷体数字信息，减少手工录入数据和同样的数据重复输入所投入的人力、物力和财力[1]。

基于传统人工方法的缺点，提出了利用机器视觉图像处理软件 HALCON 作为平台，对发票上的发票号进行识别的方法。实验证明：该方法识别准确率高，识别速度快，能有效的减少人员的工作量，提高工作效率。

收稿日期：2016-11

2 HALCON 机器视觉软件介绍

HALCON 是德国著名机器视觉厂商 MVtec 公司开发的一套有完善的数字图像处理的机器视觉软件。HALCON 软件中的 HDEVELOP 是能够与用户交互式的集程序、分析、设计、编程于一体的图像处理界面，其中包含了图像窗口、参变量值的变化观察窗口、程序编辑窗口和算子窗口等数种界面，能够让用户直接能够对平面图像或者影视图像进行编辑和观察变化。它不仅提供功能全面的视觉处理库，而且还提供了几乎所有的最先进和最新的技术算法和算子[2]，其包含了 1 000 多个独立的函数，再加上底层的数据管理核心构成了这个软件。它有数学变换、色彩分析、数学几何变换、校正分类各类滤波、辨识、分类、形状搜索等各种图像处理的功能。目前该软件广泛应用于工业自动化监测检测、遥感探测、遥感监控、医学图像的分析等

方面。

3　发票号识别过程的实现

完整的发票号识别过程应包括以下几个内容:首先对读取的图像进行图像预处理,然后对预处理后的图像进行数学形态学处理,接着把粘连的数字分割开,最后用基于多层感知分类器(MLP)对发票号数字进行识别。由于 HALCON 软件为我们提供了较好的界面窗口,可以通过语句"image (Image, '发票.jpg')",对发票图像进行读取显示,如图 1 所示。

图 1　原始发票

3.1　图像的预处理

图像的预处理是排除非字符区域的干扰。预处理包括以下几个步骤:把原图像转化成单通道图像、图像增强、高斯滤波、阈值分割、连通区域分割。通过预处理,发票图像中的发票号数字区域被提取出来,排除了其他干扰,为下一步的数学形态学图像处理提供了基础。其图像预处理流程如图 2 所示。

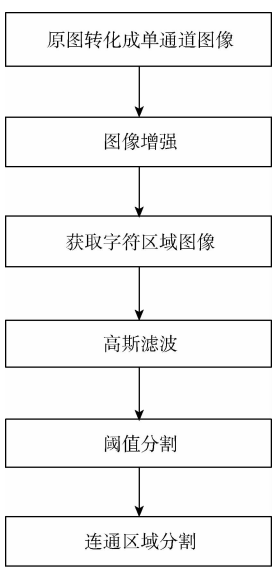

图 2　预处理流程

1)把彩色图像转化成单通道灰度图像

RGB 颜色空间就是说图像上的每一个像素由 R(红色)、G(绿色)、B(蓝色)来表示,这 3 种色也成 3 基色,自然界的任何颜色都可以用它们的组合来表示,并且每一通道的灰度值范围都是 0～255,下图是归一化后的彩色模型,其值都在[0,1]范围内。

图 3 的 3 个坐标轴分别是 R,G,B 通道,空间上每一点的颜色值就取决于这 3 个坐标值的大小,并且由它们各自的大小共同构成像素的颜色值,这就是 RGB 三坐标的含义。在 HALCON 中,用 decompose3(Image, R, G, B)把发票图像分成 R、G、B 三个单通道图像,对比之后选取 R 通道图像。

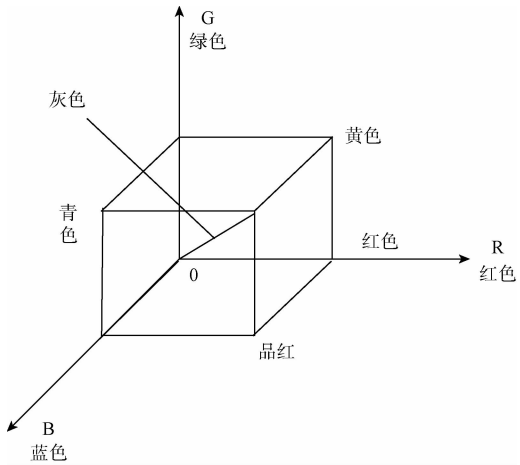

图 3　RGB 彩色模型

2)图像增强

图像增强是为了让感兴趣的字符区域更加突出,直方图均衡化就是其中的一种方法。直方图均衡化:当一幅图像的像素占据了所有的灰度级并且成均匀分布时,则该图像具有比较高的对比度和多变的灰度色调。图像的直方图均衡化就是通过改变图像的全部或局部对比度来进行图像增强的技术。

设为 r_k 为图像的第 k 级灰度值,n_k 为图像中具有灰度值 r_k 的像素的个数,n 是图像中的像素总数,L 为图像灰度级的级数。由于数字图像的灰度级是离散的,所以可用灰度级 r_k 的频数近似替代概率值[7]。这样一幅图像中第 k 个灰度级 r_k 出现的概率为:

$$p_r(r_k) = \frac{n_k}{n}, k = 0,1,2,\cdots,L-1 \tag{1}$$

则直方图均衡化的公式是:

$$s_k = T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}, \quad 0 \leqslant r_j \leqslant 1; k = 0,1,\cdots,L-1 \tag{2}$$

在 HALCON 中,用语句 equ _ histo _ image (R, ImageEquHisto)来实现直方图均衡化,扩大 R 通道的图像

的灰度级，然后用 emphasize（ImageEquHisto，ImageEmphasize，7，7，1）来增加对比度，突出字符区域，效果图如图 5 所示。

图 5　图像增强之后的图

3）获取字符区域图像

由于只是对发票图像中的票号数字部分进行识别读取，为了减少图像处理的运行时间，提高运行效率，只把发票图像中含有发票号的区域提取出来。用 gen_rectangle1（Rectangle，1307.12，1087.43，1415.94，1856.03）和 reduce_domain（ImageEmphasize，Rectangle，ImageROI）来提取发票号字符区域，效果图如图 6 所示。

图 6　字符区域图像

4）高斯滤波

数字图像常会受一些随机误差而退化，或者在图像获取的过程中因环境条件、成像设备和传感器原件自身质量的影响，在图像的传输过程中因所用传输新到的干扰污染等，会产生噪声。高斯噪声是一种含有强度服从正态分布的噪声。高斯滤波属于线性平滑的滤波方法，广泛应用于对图像中高斯噪声的消除，常用于平滑图像。在图像处理中，高斯滤波的实现方式主要有两种，一种是通过傅里叶变换，另一种是用离散化窗口卷积。其高斯离散逼近函数式是：

$$\exp(x)=\frac{1}{2\pi\sqrt{\sigma}}\exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{3}$$

高斯滤波就是对整副图像进行加权平均的过程，每一个像素的值都由其本身和领域内的其他像素经过加权平均后得到。高斯滤波的平滑过程：用一个模板扫描图像中的每一个像素，用模板确定邻域内像素的加权平均灰度值去替代模板中心像素点的值。在 HALCON 软件中，可以用以下算子 gauss_filter（ImageROI，ImageGauss，5）来进行高斯滤波来减少噪声，效果图如图 7 所示。

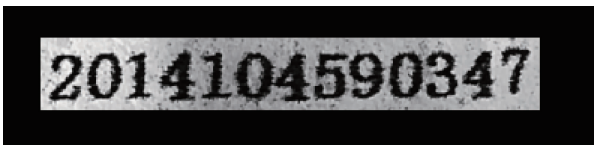

图 7　高斯滤波后的字符区域

5）阈值分割

在数字图像中，目标物体和无用背景物体常常混合在一起，怎样把目标物体从图像中分割出来是个难点。阈值分割，又称图像二值化，是一种基于灰度的分割技术，它对目标物体与背景物体有较强的对比度的图像的分割特别有用。二值化的方法非常多，但是又没有对任何对象都普遍适用的方法，必须根据具体的处理对象。灰度阈值的方法是把数字图像的灰度分成不同的等级，然后用设置灰度阈值的方法确定要分割的区域。假设输入图像为 $f(x,y)$，输出图像为 $g(x,y)$，则输出图像表达式为：

$$g(x,y)=\begin{cases}1, f(x,y)\geqslant t\\ 0, f(x,y)<t\end{cases} \tag{4}$$

或

$$g(x,y)=\begin{cases}1, f(x,y)\leqslant t\\ 0, f(x,y)>t\end{cases} \tag{5}$$

这就是图像的阈值分割，它的目的就是求一个阈值 t，并用 t 将图像 $f(x,y)$ 分成目标区域和背景区域。由于实际得到的图像其目标和背景不可能只分布在两个灰度范围内，此时就需要两个或者多个阈值来分割提取目标区域[3]。、通过设定相关阈值，将小于该阈值的轮廓去除，大于该阈值的轮廓保留[4]。方法如下：选一个区间 (t_1,t_2) 作为阈值，用下面的公式对图像进行阈值处理：

$$g(x,y)=\begin{cases}1, \text{若 } t_1\leqslant f(x,y)\leqslant t_2\\ 0, \text{其他}\end{cases} \tag{6}$$

或

$$g(x,y)=\begin{cases}1, \text{其他}\\ 0, \text{若 } t_1\leqslant f(x,y)\leqslant t_2\end{cases} \tag{7}$$

选取合适的阈值对阈值的结果会有很大的影响，其中依据灰度直方图分布判断阈值的方法效果最好[5]，其字符区域灰度值直方图如图 8 所示。

图 8　字符区域灰度直方图

从灰度直方图中可知阈值选取 0～69,最合适。用语句 threshold (ImageROI, Regions, 0, 69)将字符形状区域大致分割出来,如图 9 中浅色部分。

图 9　阈值分割之后的字符区域

6)连通区域分割

利用阈值分割得到的数字区域是一个整体,其中含有其他一些干扰区域,要把这些干扰区域排除,必须计算出阈值分割后所得到的区域内包含的所有连通区域[6],在 HALCON 中,用算子 connection (Regions, ConnectedRegions) 把阈值后分割出的区域,连通成一个个小区域,放在了 ConnectedRegions 中,然后利用面积特征,排除其他干扰用算子 select_shape(ConnectedRegions, SelectedRegions, 'area', 'and', 150, 99999)把字符区域选出来,这样就得到了只含数字的 SelectedRegions 区域。效果图如图 10 所示。

图 10　选出的只含字符的区域

3.2　数字形态学处理

数学形态学是以几何学为基础的用于图像处理的新方法,主要以图像的形态特征为研究对象[7]。从图 10 中可以看到连通后的数字区域中,数字周围有一些突出的毛刺,数字内部也有一些较小的孔洞,对字符区域进行数学形态学开运算能有效地消除物体周围的毛刺,闭运算能有效的填充物体内部的孔洞。数学形态学的基本运算包括两种:膨胀和腐蚀。闭运算和开运算就是由这两种基本运算进行推导和[8],设 A 为目标图形,B 为结构元素, 表示 A 的补集,则目标图像 A 被结构元素 B 腐蚀运算定义为:

$$A \ominus B = \{c \mid B + c \subset A\} \tag{8}$$

目标图像 A 被结构元素 B 膨胀运算定义为:

$$A \oplus B = [A^c \ominus (-B)]^c \tag{9}$$

闭运算是腐蚀和膨胀的结合,先对图像进行膨胀,再进行腐蚀运算,运算定义为:

$$A \bullet B = [A \oplus (-B)] \ominus (-B) \tag{10}$$

开运算是腐蚀和膨胀的结合,先对图像进行腐蚀,再进行膨胀运算,运算定义为:

$$A \circ B = (A \ominus B) \oplus B \tag{11}$$

形态学闭运算可以填充区域中间的孔洞,开运算可以去除区域边界附近的细小毛刺[9]。

在 HALCONN 软件中,本文先用开运算的算子 opening_circle(SelectedRegions,RegionOpening, 2)消除数字周围的毛刺部分,然后再把开运算之后的区域 RegionOpening 用数学形态学进行闭运算 closing_circle (RegionOpening, RegionClosing1, 3)来填充字符内部的小孔洞,为了便于查看,用算子 shape_trans(RegionClosing1, RegionTrans0, 'rectangle1')把字符区域转化成小矩形显示,效果图如图 11 所示。

图 11　形态学处理后的字符

3.3　分割粘连字符

从上图可以看出形态学处理后的区域中,"4"和"5"两个字符粘连在一起了,在 HALCON 中, partition_dynamic (RegionClosing1,Partitioned,70, 70)算子可以把粘连的字符分割开,然后再用 shape_trans (Partitioned, RegionTrans,'rectangle1'),把分割后的单个字符转化成矩形便于显示,效果图如图 12 所示。

图 12　粘连字符被分割后的字符

4　多层感知分类器(MLP)

要识别字符,就必须将拆分的字符分类,也就是把每个分割出的区域赋予一个符号标记,HALCON 提供了一些可用于 OCR 的训练过的字体,从文件中直接读出训练过的 OCR 分类器。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、模板匹配法、基于数字结构特征的识别算法、基于组合特征的识别算法等[10]。

本文使用的分类器是基于一种特殊形式的多层神经网络感知器分类器。多层感知器(MLP)也叫神经网络,多层感知器神经网络分类法是基于多层前馈神经网络[11],是一种前向结构的人工神经网络,映射一组输入向量到一组输出向量。MLP 就是模仿生物神经元传输的机制,神经网络中每个神经元如同多层感知器的每个感知器,它由多个节点层组成,除了输入层,它中间可以有很多层,每层都全连接到下一层且每一层都互相连接的,也就是每一个神经元节点与下一层都有连接,最底层是输入层,中间是隐藏层,最上面是输出层。

多层感知器的结构组成如下:

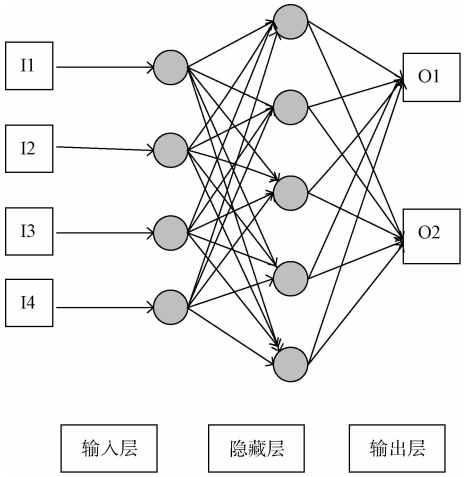

图 13　多层感知器结构组成

除了输入节点，每个节点都是一个带有非线性激活器函数的神经元或者是处理单元，每一节点的激活输出值由结点输入，激活函数及偏置量所决定，可以解决线性不可分问题。

除输出层外的激活函数可以是逻辑激活函数：

$$f(x)=\frac{1}{1+e^{x}} \tag{12}$$

也可以是双曲线正切激活函数：

$$f(x)=\tanh\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{13}$$

输出层中使用 softmax 激活函数：

$$f(x)=\frac{e^{x_i}}{\sum_{j=1}^{n}e^{x_j}} \tag{14}$$

前一层的输出为下一层的神经元的输入，每层神经节点只接受前一层的神经元节点的输出信号。对于输入信号，要先向前传播到隐含层节点，经作用函数后，再把隐藏节点的输出信号传播到输出节点，最后给出输出结果[12]。首先用 sort_region (Partitioned, SortedRegions′ first_point′, ′true′, ′column′)，把分割后的单个数字区域进行排序，为下面识别准备，然后用算子 read_ocr_class_mlp(′DotPrint_0-9.omc′, OCRHandle) 和 do_ocr_multi_class_mlp (SortedRegions, ImageGauss, OCRHandle, Class, Confidence) 来识别出字符，在原图像中显示识别结果如图 14 所示。

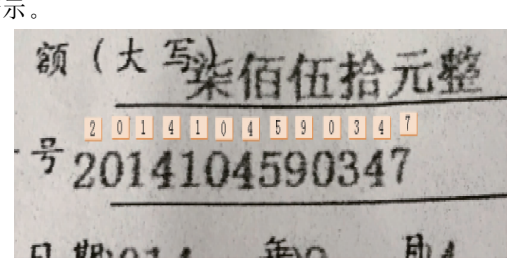

图 14　识别结果

5　实验结果

以下是采用人工方式和基于 HALCON 的图像处理方式分别对学校的财务处发票进行对比实验得出的数据统，然后做成的统计分析表格，如表 1、2 所示。

表 1　人工方式数据

发票个数/张	总花费识别时间/s	准确率/%
300	9 000	94.67

表 2　基于 HALCON 方式数据

发票个数/张	总花费识别时间/s	准确率/%
300	1 200	99.33

上面的数据是对学校的财务处发票进行对比是别的的结果。从上面两个表格可以看出。在相同发票个数的前提下，人工方式识别的速度较慢，准确率也较低，而基于 HALCON 的方法，虽然有一些人为干扰和环境因素等的影响，但是其准确率达到了 99%，识别时间，大为减少，总体达到了试验设计目的。

6　结　　论

通过试验数据对比，提出的基于 HALCON 软件的方法不但成功的识别出了发票号数字，在准确率方面，尤其是时间上，明显缩短了识别的时间，节省了成本，提高了识别效率。但是由于本次实验只是对一种发票进行识别，没有对其他类型的发票进行测试，且实验的样本数据还不够多，所以识别的算法还有待进一步测试和优化。

参考文献

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

[2]　金贝. 基于 HALCON 的机器视觉教学实验系统设计[D]. 北京：北京交通大学，2012：1-29.

[3]　谢凤英. 数字图像处理及应用[M]. 北京：电子工业出版社，2014.

[4]　石晓伟. 图像处理在安全壳缺陷检测中的应用[J]. 电子测量技术，2016，39(6)：89-93.

[5]　陈银. 基于扫描笔的发票识别系统设计[D]. 成都：电子科技大学，2014.

[6]　王佳. 发票印刷体数字识别方法的研究[D]. 沈阳：沈阳工业大学，2016.

[7]　黄明鑫，杨龙兴，庄立东，梁栋. 基于 HALCON 图像处理的粘连零件颗粒计数方法研究[J]. 机械设计与制造工程，2016(2)：85-89.

[8]　李旭辉. 数字图像处理李俊山. 北京：清华大学出版社，2007.

[9] 贺瑞芳. 面向视觉假体的复杂图像处理技术[J]. 电子测量技术,2015,38(11):55-59.
[10] 迟国炜. 商业发票手写体数字识别系统的设计与实现[D]. 沈阳:沈阳工业大学,2006.
[11] 吴雪芬,李昊昱,陈功,等. HALCON 软件在车牌图像处理中的应用[J]. 电子质量,2014(12):49-54.
[12] 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.

作者简介

张有,1990 年出生,硕士研究生,主要研究方向测试计量技术及仪器、数字图像处理识别与缺陷检测。
E-mail:582565501@qq.com

陈晓荣,女,1974 年出生,副教授,研究方向为图像处理、在线检测、信号与信息处理。
E-mail: 1906182543@qq.com

(上接第 116 页)

[9] 王秀平,白瑞林,刘子腾,等. 由任意平行四边形确定摄像机内参数的方法[J]. 上海交通大学学报,2015,49(3):366-370.
[10] 赵振庆,叶东,吴斌,等. 消隐点共线约束逐点畸变校正算法[J]. 光学精密工程,2015,23(4):1196-1204.

作者简介

马志学,学士,高级工程师,主要研究方向为电力高压测试和电力图像处理识别等。

袁爱仙,学士,工程师,主要研究方向为电力工程测量等。

黄晓波,学士,工程师,主要研究方向为电力高压试验等。

蔡建峰,学士,助理工程师,主要研究方向为电力高压试验等。

(上接第 121 页)

作者简介

徐志军,1992 年出生,现为西南交通大学牵引动力国家重点实验室硕士研究生,主要研究方向为动态检测技术及数据处理。
E-mail:xzj20104393@163.com

陈建政,1968 年出生,现为西南交通大学牵引动力国家重点实验室研究员,硕士研究生导师,主要研究方向为仪器与检测 、机车车辆动态检测技术及数据处理、机车车辆故障诊断及安全检测。
E-mail:13980050615@126.com

(上接第 125 页)

[7] 刘一凡,蔡振江,索雪松. 基于基础矩阵与 HEIV 模型的双目相机标定[J]. 电子测量与仪器学报,2016,30(9):1425-1431.
[8] 王隆. 基于数字近景摄影测量的隧道变形监测研究[D]. 重庆:重庆交通大学,2012.
[9] 梁灵飞,基于交比不变和间距比不变的线阵相机畸变标定方法[J]. 科学技术与工程,2014,14(14):248-251.
[10] 付阳. 基于 2D 单应矩阵约束的图像匹配方法[J]. 首都师范大学学报:自然科学版,2015,36(3):73-77.
[11] 姚庆源. 多线阵 CCD 摄像机标定及应用研究[D]. 洛阳:河南科技大学,2013.

作者简介

黄建斌,1995 年出生,现就读于江苏大学机械工程学院,主要研究方向为仪器与测试技术、图像处理和自动化技术等。
E-mail:1074576125@qq.com