DOI: 10.19322/j.cnki.issn.1006-4710.2016.04.004

自适应字符切分及提取算法研究

金海燕，夏　婷，王　彬

（西安理工大学 计算机科学与工程学院，陕西 西安 710048）

摘要：在字符识别技术日趋成熟的现状下，单个字符的正确切分及提取已经成为制约字符识别精确度的关键因素。本文针对二手车发票上印刷体的日期数字（阿拉伯数字），对图像二值化处理后，采用垂直方向投影和轮廓特征两种策略进行自适应字符切分及提取。实验结果表明，该方法提高了从图像中定位出来的字符串的切分率，并保证了字符切分和提取的正确率平均达到99%。

关键词：字符切分及提取；字符识别；图像二值化

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1006-4710(2016)04-0399-04

Study of adaptive character segmentation and extraction algorithm

JIN Haiyan, XIA Ting, WANG Bin

(School of Computer Science and Engineering, Xi' an University of Technology, Xi' an 710048, China)

Abstract: In the current the situation of character recognition technology is becoming more and more mature, and the correct segmentation and extraction of single character has become a key factor to control the accuracy of character recognition. This paper focuses on second-hand car invoice printing of digital date (Arabic numerals), with both vertical projection and contour feature combination strategies for adaptive character segmentation and extraction being conducted after image binarization processing. Experimental results show that using the proposed method can improve the segmentation rate of the string from the image, ensuring that the average accuracy rate of the extracted image can reach 99%.

Key words: character segmentation and extraction; character recognition; image binarization

单一字符识别技术发展迅速，目前，多数字符识别系统对于印刷字符数字的识别已经达到较高识别率，此时字符的正确切分就成为进一步提高识别率的关键因素。

目前，一些学者已经提出了一些字符切分方法。投影法[1]主要是对字符图像二值化后向水平方向投影，投影曲线两个相邻波谷的位置就是切割线的位置，该方法简单、快速，但是对噪声等干扰因素敏感。上下轮廓(Upper/Lower Contour)特征法[2]通过逐列搜索扫描来确定字符的上下轮廓曲线，并根据其轮廓曲线来近似确定单个字符。该方法对字符宽度、间隔不固定的图像切分效果不好。滴水算法(Drop-Falling)[3]根据水滴运动所经过的路径轨迹来确定字符的分割路径，该方法对粘连字符有很好的切割效果，但对于有倾斜或本身是凹形结构的字符会造成字符断裂，对某些左中右结构的汉字会造成过度切分。连通域法[4]是基于字符本身的像素连续性，每一个字符或者字符的一部分构成一个连通域，该方法抗噪性较好，但是对一些由互不相连的几部分组成的汉字，会将其分割成几部分，从而导致切分不正确。

综上所述，字符切分中的最小单元是单个字符，描述字符单元最精确的知识就是字符形状[5]。字符切分过程就是对单个字符的外轮廓边界准确定位的过程，根据字符的外轮廓确定字符的外边界，进而确定切割的位置。而单个字符数字内部的间断、字符间的粘连以及图像背景噪声不同程度的污染等，都是造成不能精确切分的主要因素及难点[6]。本文就二手车发票日期字符的识别提出自适应字符切分及提取算法。该算法采用两种策略多次自适应切分及提取，首先采用上下轮廓追踪法切分，如果切分结果不理想，再进行垂直投影法切分。切分过程中可以根据字

收稿日期：2016-03-04

基金项目：国家自然科学基金资助项目(61472204, 61272283)

作者简介：金海燕，女，教授，博士，研究方向为图像处理，模式识别，智能信息处理。E-mail:jinhaiyan@xaut.edu.cn

符的宽高来判断切分结果是否正确,最后根据日期的有效位数去除年、月、日后,提取 8 位数字图片。

1 自适应字符切分及提取算法

1.1 字符的切分及提取算法

本文算法主要采用直线检测法定位出日期位置,因此对日期位置没有明确要求。假设日期位置位于发票的左上方,首先定位日期位置,切取发票的日期数字部分,再对该部分进行初步的预处理[7-9]。本文提出的字符切分及提取方法分为两步。

1.1.1 上下轮廓追踪

首先,用简单统计法[10]将灰度位图二值化,并对二值位图进行边缘噪声清理和消除孤立点,然后,追踪搜索上下轮廓特征,分析空白间隔,切除不需要的部分,对保留下来的部分搜索极值点,进而于特征点处切分。

预处理原始图像,裁剪出左上角位图,从而大大减少后续操作数据量,并按固定比例进行缩放,记录此时高度 h_0、宽度 w_0。

用最大类间方差(OTSU)[11]算法二值化该灰度图,利用霍夫变换检测表格线和倾斜角度,定位出表格左上角及角度的矫正。以位图左上角为坐标原点,将二值图像像素值为 0 的直角坐标系下的坐标点(x,y)转换为极角坐标下的坐标点(r,θ),并将累加个数存储于 ***hough*** 矩阵数组中,如公式(1)所示。

$$r = x \times \cos\theta + y \times \sin\theta \tag{1}$$

式中,$r\in(0,(h_0^2+w_0^2)^{1/2})$,$\theta\in(0°,180°)$,$(r,\theta)\in$ ***hough***[ ]。

从 $r=h_0-a$ 处开始降序搜索 ***hough*** 矩阵中同一(r,θ)的数量值,当首次出现大于预设的门限值 b 时设:

$up\_dis=r$;

$angle=\theta$;

式中,up_dis 为表格最上边直线距原点的距离,$angle$为表格倾斜角度,a、b 为常量。

同理,从 $r=a$ 处升序搜索 ***hough*** 矩阵中同一(r,θ)的数量值,当首次出现大于预设的门限值 c 时设:

$left\_dis=r$;

式中,$left\_dis$ 为表格最左边直线距原点的距离,c 为常量。

若 $\theta\neq0$,判断表格倾斜角度 $angle$ 及表格左上角坐标(x_0,y_0),编写如下程序:

$x_0=left\_dis$;

if($angle>90°$)

$angle=180°-angle$;

$y_0=h_0\times\cos(angle\times3.14\div180)-up\_dis$;

else

$y_0=h_0\times\cos(angle\times3.14\div180)-up\_dis+w_0\times\sin(angle\times3.14\div180)$;

$angle=-angle$;

采用邻近插值算法[12]以 $angle$ 角度对位图进行倾斜矫正。

以(x_0,y_0)为矩形左下角,按固定比例从原灰度图中裁剪出目标所在区域部分,此时位图矩阵记为 $\boldsymbol{image}_0$。

采用简单统计法二值化 $\boldsymbol{image}_0$,搜索 8 邻域连通域消除孤立噪声点,并进行边缘清理,记录此时二值位图为 I_1。

在 I_1中定位字符的上下轮廓特征,可采用先自上而下再自下而上沿纵轴方向扫描像素点的方法,用一维数组 $U(1,I_{1\_x})$、$D(1,I_{1_x})$分别记录上下轮廓的纵坐标点,得到上下轮廓之间距离的离散值 $span(1,I_{1\_x})$,如公式(2)所示。

$$span(1,i)=U(1,i)-D(1,i) \tag{2}$$

式中,I_{1_x}为 I_1 的宽,i 初始化为 0,上限为 $I_{1\_x}-1$,移动步长为 1。

分析 $span(1,i)$空白间隔处,并自适应选取缩小目标区域的切割点(x_1,y_1),基于 $\boldsymbol{image}_0$图裁剪出日期矩阵 $\boldsymbol{image}_1$。

依据先递减后递增的规律从左向右逐一查找 $span(1,i)$中的极值点,设小于 d(注:d 为常量经验值)的极值点对应的坐标点为切分点,基于 $\boldsymbol{image}_1$ 灰度位图裁剪,最终切分成字符碎片,用最大类间方差算法生成二值化效果图。

对字符碎片做轮廓临界点裁剪,二次切分自适应提取有效字符片位图。首先,统计字符碎片高度、宽度,排序查找出现次数最多的高度、宽度分别记为 h、w;其次,鉴于此时二值化位图细节清晰,只需取上轮廓特征极值点,参考 h、w 及非有效字符的提取情况,自适应的进行二次切分,直至取到所有有效字符片存于数组矩阵 ***digital_matrix*** 中。

1.1.2 垂直投影

对经过 1.1.1 中切分处理后仍然切分错误的字符图像,采用垂直投影[13]进一步分析。对 1.1.1 中定位裁剪出的日期灰度图 $\boldsymbol{image}_0$,采用最大类间方差算法二值化记为 I_0,再以 8 领域连通域搜索消除孤立噪声点,并进行边缘清理。

将 I_0进行垂直投影,投影累加值记录在一维矩阵 $\boldsymbol{S}(1,I_{0_x})$中,如公式(3)所示。

$$\boldsymbol{S}(1,i)=\sum_{j=0}^{I_{0\_y}-1}(1-I_0(i,j)) \tag{3}$$

式中，I_{0_x}为 I_0的宽，I_{0_y}为 I_0的高，$I_0(i,j)$为坐标(i,j)点的像素值（注：0 为黑色像素点，1 为白色像素点），$\boldsymbol{S}(1,i)$为在横坐标为 i 下所有像素为 0 的点的个数。

基于空白间隔处进行字符切分，得到字符碎片 $\boldsymbol{piece\_matrix}[sum]$（注：$sum$ 为碎片总数）。对字符碎片做轮廓临界点裁剪，二次切分自适应提取有效字符片位图。

1.2 算法流程图

首先，将读入的图像进行倾斜矫正；然后，从中定位出表格左上角的坐标，从而定位出日期；最后，结合两种策略对日期进行单个字符的切分及提取。算法流程图如图 1 所示。

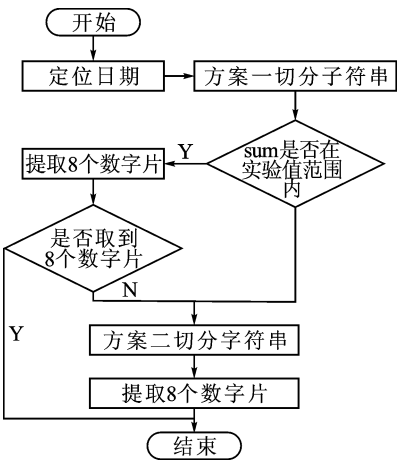

图 1 算法流程图

Fig. 1 Algorithm flowchart

2 实验分析

以奥迪 4S 店实际的二手车售车发票为实验样本进行分析。图 2 为发票源图的灰度图。

图 2 发票源图的灰度图

Fig. 2 The original invoice grayscale

步骤 1：选取经过加权平均法灰度化的灰度图图 2 作为源图。初步裁剪出左上角位图，并按固定比例进行缩放，记录此时高度 $h_0=290$（像素）、宽度 $w_0=600$（像素），左上角位图如图 3 所示。

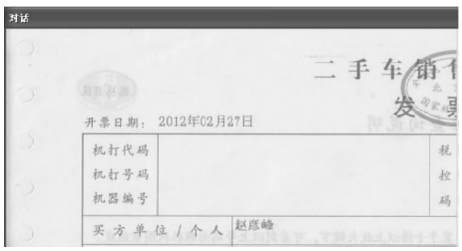

图 3 发票左上角图

Fig. 3 The image of invoice upper left corner

步骤 2：采用最大类间方差算法对图 3 进行二值化，利用霍夫变换检测表格线和倾斜角度，定位出表格左上角及角度的矫正。以左上角为坐标原点，检测到表格左上角坐标为（99，93）、角度为 0°。

步骤 3：位图无倾斜，不需要矫正。

步骤 4：以（99，93）为矩形左下角，按固定比例从原灰度图中裁剪出目标所在区域部分，记为 $\boldsymbol{image}_0$，如图 4 所示。

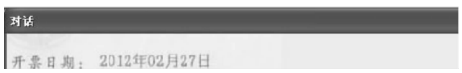

图 4 定位开票日期头位图

Fig. 4 The billing date bitmap head positioning

步骤 5：采用简单统计法二值化 $\boldsymbol{image}_0$，搜索 8 邻域连通域消除孤立噪声点，并进行边缘清理，记录此时二值位图为 I_1。

步骤 6：在 I_1中用扫描像素的方法定位字符的上下轮廓特征，得到上下轮廓之间距离的离散值 $span(1,I_{1\_x})$，$U(1,I_{1\_x})$、$D(1,I_{1_x})$分别记录上下轮廓的纵坐标点。自左向右分析 $span(1,i)$空白间隔处，缩小目标所在区域范围，切除“开票日期：”部分后的左下角坐标为（80，0），基于灰度位图 $\boldsymbol{image}_0$再次切割，记为 $\boldsymbol{image}_1$，如图 5 所示。

图 5 定位日期位图

Fig. 5 Location date bitmap

步骤 7：依据先递减后递增的规律从左向右逐一查找 $span(1,i)$中的极值点（注：i 从 80 开始），设小于 $d(d=7)$的极值点对应的坐标点为切分点，其横坐标为：29、39、47、57、64、74、83、94、110。基于灰度位图 $\boldsymbol{image}_1$裁剪，最终切分成字符碎片，并采用

OTSU 算法二值化。

步骤 8:对字符碎片做轮廓临界点裁剪如图 6 所示,放大字符碎片效果如图 7 所示。二次切分自适应提取有效字符片位图。首先,统计字符碎片高度、宽度,排序查找出现次数最多的高度、宽度分别记为 $h=13$、$w=8$,其次,鉴于此时二值化位图细节清晰,只需取上轮廓特征极值点,参考均值 h、w 以及非有效字符的情况,自适应的进行二次切分,从中提取 8 个数字片矩阵,分别考虑年、月、日、数的情况。依据均宽 w、均高 h,分别对日期数、"日、月、年"的宽度、高度做一个大概定位,不在这个范围内的,自适应的进行二次切分或者向前选取新的矩阵进行提取,最终只提取到 8 个数值矩阵存储于 ***digital_matrix*** 矩阵数组中。提取出的 8 个数字片如图 8 所示。

图 6 字符碎片位图
Fig. 6 Character fragment bitmap

图 7 字符碎片放大位图
Fig. 7 Character fragmentation bitmap

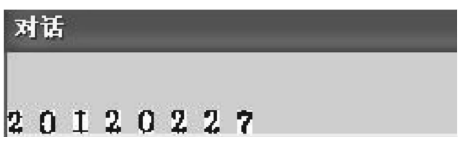

图 8 单个数字片位图
Fig. 8 Single digital bitmap

以上实验中提取的有效字符数量及切分结果说明字符切分正确,无需再进行垂直投影处理,而对有些发票,当采用上下轮廓追踪处理后有效字符数不为 8 或出现无法识别的字符时,需要采用垂直投影重新切分处理,以达到切分的最佳效果。运用本文算法对 100 张不同时间日期的二手车发票进行了日期字符的切分及提取,实现了对日期字符的精确定位、切分和提取,并且字符的边界划分较为清晰准确,取得了比较满意的结果,提取出的待识别数字正确率平均达到 99%。

3 结 语

自适应字符切分及提取方法,是将读入的图像进行倾斜矫正,从中定位出发票表格左上角的坐标 (x_0, y_0),并以此设置矩形框裁剪出图像,最后进行单个字符的切分及提取。本文结合多种二值化图像的效果,采用垂直方向投影和轮廓特征两种策略进行自适应字符切分及提取,提高了从图像中定位出来的字符串的切分率,并保证了提取出的待识别数字的正确切分率。

参考文献:

[1] 迟小君,孟庆春. 基于投影特征值的车牌字符分割算法[J]. 计算机应用研究,2006,32(7):256-257.
CHI Xiaojun, MENG Qingchun. Character segmentation of license plate based on projection and eigenvalue[J]. Application Research of Computers, 2006, 32(7): 256-257.

[2] STRATHY N W, SUEN C Y, KRZYZAK A. Segmentation of handwritten digits using contour features[C]// Document Analysis and Recognition, 1993, Proceedings of the Second International Conference on, Tsukuba, 1993:577-580.

[3] KHAN S A. Character segmentation heuristics for check amount verification[D]. Cambridge: Massachusetts Institute of Technology, 1998.

[4] 陈艳,孙羽菲,张玉志. 基于连通域的汉字切分技术研究[J]. 计算机应用研究,2005,22(6):246-248.
CHEN Yan, SUN Yufei, ZHANG Yuzhi. Research on Chinese character segmentation based on connected domain[J]. Computer Application Research, 2005, 22(6): 246-248.

[5] 刘刚,丁晓青,彭良瑞,等. 多知识综合判决的字符切分算法[J]. 计算机工程与应用,2002,38(17):59-62.
LIU Gang, DING Xiaoqing, PENG Liangrui, et al. A character segmentation algorithm based on synthetic decision[J]. Computer Engineering and Applications, 2002, 38(17): 59-62.

[6] 王叶. 车牌识别系统中字符切分和识别技术的研究[D]. 北京:北京邮电大学,2009.
WANG Ye. Study of character recognition in vehicle license plate system[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[7] 童立靖,张艳,舒巍,等. 几种文本图像二值化方法的对比分析[J]. 北方大学学报,2011,23(1):25-33.
TONG Lijing, ZHANG Yan, SHU Wei, et al. Comparison and analysis of several document image binarization algorithm[J]. Journal of North College, 2011, 23(1): 25-33.

(下转第 415 页)

[4] 宋少民，杨柳，徐国强. 石灰石粉与低品质粉煤灰复掺对混凝土耐久性能的影响[J]. 土木工程学报，2010，43(S2)：368-372.
SONG Shaomin，YANG Liu，XU Guoqiang. Influence of limestone powder and low quality fly ash on the durability of concrete[J]. China Civil Engineering Journal，2010，43(S2)：368-372.

[5] 寇世聪，潘志生. 不同强度混凝土制造的再生骨料对高性能混凝土力学性能的影响[J]. 硅酸盐学报，2010，40(1)：7-11.
KOU Shicong，POON Chisun. Effect of quality of parent concrete on the mechanical behavior of high performance recycled aggregate concrete[J]. Journal of the Chinese Ceramic society，2010，40(1)：7-11.

[6] 中华人民共和国住房和城乡建设部. 普通混凝土配合比设计规程：JGJ55—2011[S]. 北京：中国建筑工业出版社，2011.

[7] 中华人民共和国建设部. 普通混凝土用砂、石质量及检验方法标准：JGJ52—2006[S]. 北京：中国建筑工业出版社，2006.

[8] 中华人民共和国建设部，中华人民共和国国家质量监督检验检疫总局. 普通混凝土拌合物性能试验方法标准：GB/T 50080—2002[S]. 北京：中国建筑工业出版社，2002.

[9] 中华人民共和国建设部. 普通混凝土力学性能试验方法：GB/T 50081—2002[S]. 北京：中国建筑工业出版社，2003.

[10] 中华人民共和国住房和城乡建设部，中华人民共和国国家质量监督检验检疫总局. 普通混凝土长期性能和耐久性能试验方法标准：GB/T 50082—2009[S]. 北京：中国建筑工业出版社，2009.

（责任编辑　王绪迪，王卫勋）

（上接第 402 页）

[8] 李钊，李鸿. 几种去噪方法的比较与改进[J]. 工业控制计算机，2012，25(8)：9-10.
LI Zhao，LI Hong. Comparison and several de-noising methods[J]. Industrial Control Computer，2012，25(8)：9-10.

[9] 张雅兰. 扫描图象二值化处理的研究[J]. 广西光学院学报，2002，13(3)：26-28.
ZHANG Yalan. A study of binarization processing of scanning image[J]. Journal of Guangxi Institute of Light，2002，13(3)：26-28.

[10] 陈杰，王振华，窦丽华. 一种尺度自适应 Canny 边缘检测方法[J]. 光电工程，2008，35(2)：79-84.
CHEN Jie，WANG Zhenhua，Dou Lihua. Scale adaptive canny edge detection method[J]. Opto-Electronic Engineering，2008，35(2)：79-84.

[11] OHTSU N. A threshold selection method from gray-Level histograms[J]. Systems Man & Cybernetics IEEE Transactions on，1979，9(1)：62-66.

[12] 于亚龙，穆远彪. 插值算法的研究[J]. 现代计算机，2014，007(05)：32-35.
YU Yalong，MU Yuanbiao. Research on interpolation algorithms[J]. Modern Computer，2014，007(05)：32-35.

[13] 毛永明，祁宁，张东伟，等. 智能交通系统车牌字符分割研究[C]//第十二届沈阳科学学术年会，第十二届沈阳科学学术年会论文集，沈阳，2015.
MAO Yongming，QI Ning，ZHANG Dongwei，et al. Intelligent transportation system character segmentation study[C]//The Twelfth Annual Conference of Science and Technology in Shenyang，Shenyang Science Twelfth Annual Conference Proceedings，Shenyang，2015.

（责任编辑　周　蓓）