OCR 文字识别技术的研究

张婷婷,马明栋,王得玉
(南京邮电大学,江苏 南京 210003)

摘 要:图像中的文字在当下相机高速发展下显得尤为重要,人们开始通过拍摄照片直接进行图像上文字的识别,最常用的就是寄快递收寄地址的识别。其中用到的技术是 OCR(optical character recognition)字符识别技术,其中文名字叫做光学字符识别。它是利用光学技术和计算机技术通过检测字符每个像素的暗、亮模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程。随着日常生活网络化的推进,各种纸质文档的数字化智能化识别进程也在加速。经过二十世纪九十年代的发展,对字符识别技术的研究已经取得了很大的进展,市场上目前正在使用的各种 OCR 识别软件层出不穷。但是以往对证件的识别是一个比较大的难题。文中的研究主要是对普通的文字进行识别。识别系统包括三个模块:图像预处理、图像分割、字符识别。前两个模块又包含图像的二值化分析、灰度化等,对其进行了描述。

关键词:OCR;文字识别;post 方法;图像处理
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)04-0085-04
doi:10.3969/j.issn.1673-629X.2020.04.016

Research on OCR Technology

ZHANG Ting-ting, MA Ming-dong, WANG De-yu
(Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: With the rapid development of cameras, the text in images is particularly important. People begin to recognize the text on images directly by taking photos, and the most commonly used method is to recognize the address of receiving and mailing by express delivery. The technology used is OCR, optical character recognition, which is a process of using optical technology and computer technology to determine the shape of each pixel by detecting the dark and bright mode of the character, and then translating the shape into computer text by character recognition method. With the development of network in daily life, the digital and intelligent recognition process of all kinds of paper documents is also accelerating. After the development of character recognition technology in the 1990s, great progress has been made. Various OCR recognition software are being used in the market. But the identification of documents is a big problem in the past. We mainly focus on the recognition of common characters. The recognition system consists of three modules: image preprocessing, image segmentation and character recognition. The first two modules also include image binarization analysis, grayscale and so on, which are described.

Key words: OCR; character recognition; post method; image processing

0 引 言

OCR 的概念于 1929 年由德国科学家 Tausheck 最先提出。最早对印刷体汉字识别进行研究的是 IBM 公司的 Casey 和 Nagy[1]。在 20 世纪的 60、70 年代,世界各个国家对 OCR 的研究主要集中在对文字的识别方法上,并且仅是对 0 到 9 的数字进行识别。而国内在 OCR 技术的研究相对较晚。在 20 世纪 70 年代,国内学者起初研究的是数字、英文字母及符号的识别,70 年代末开始研究汉字的识别。

OCR[2] 技术在目前互联网及人工智能迅速发展的趋势下,也有了飞速的发展。到目前为止,结合其他方向的技术,特别是人工智能方向,OCR 技术已经发展到可以识别带有地理位置信息的图纸,可以对文字进行高精度的识别,包括生僻字在内的情况[3]。

OCR 技术也普遍应用在日常生活,最为熟悉的是百度网页可以拍照识别图纸上的题目文字,另外还有

收稿日期:2019-04-19 **修回日期**:2019-08-20 **网络出版时间**:2019-12-05
基金项目:江苏省自然科学基金-青年基金项目(BK20140868)
作者简介:张婷婷(1995-),女,硕士研究生,CCF 会员(A9204G),研究方向为无线电通信技术;马明栋,博士,教授,研究方向为地理信息系统平台软件设计与开发等;王得玉,博士,副教授,研究方向为水环境遥感、GIS 软件设计与开发。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20191205.1113.040.html

百度 AI 输入法中的一系列文字识别功能,包括:身份证识别、名片识别、表格识别等等[4]。

1 图像输入预处理

图像输入,对其进行预处理操作。对于不同格式的图像,有着不同的存储格式和压缩方式[5]。文中的图像输入,上传图片方法是实现客户端上传一张本地图片或者使用抓包工具 postman 向百度服务器发送 url 请求[6]时设置参数添加一张本地带汉字的图片。

预处理过程,是使用阈值分割法把图片上每个像素二值化[7]。以下是用 Java 语言实现的预处理函数,函数是根据图片高度和宽度遍历图片上的每个像素点,通过 ISWHITE 来判断当前像素值。

```
Int width=img. getWidth( );
Int height=img. getHeight( );
For( int x=0;x<width;++x) {
For( int y=0;y<height;++y) {
If( ISWHITE( img. getRGB( x, y))= =1) {
Img. setRGB( x,y,color. WHITE. getRGB( ));
//像素红绿蓝在一定范围置成白色
} else {
Img. setRGB( x,y,color. WHITE. getRGB( ));
//反之置成黑色
}
}
}
```

灰度化处理,在 RGB 模型中,如果 $R = G = B$ 时,则彩色表示一种灰度颜色,其中 $R = G = B$ 的值叫灰度值。灰度图每个像素只需一个字节存放灰度值[8],其范围是 0 ~ 255。将彩色图像中的三分量的亮度作为三个灰度图像的灰度值[9]。其中 $f_k(i,j)(k = 1,2,3)$ 为转换后的灰度图像在 (i,j) 处的灰度值。采用分量方法[10]:

$$f_1(i,j) = R(i,j) \tag{1}$$

$$f_2(i,j) = G(i,j) \tag{2}$$

$$f_3(i,j) = B(i,j) \tag{3}$$

二值化过程[11],根据灰度图像中像素的灰度级值的取值范围为 c_0[12],希望能够更加显现出图像中的文字部分,一般图像中的文字为黑色,在灰度图像中灰度值较小。二值化过程使用 Otsu 算法实现,Otsu 算法又称为最大类间方差法[13]。

设一幅图像大小为 $(M * N)$,$f(x,y)$ 是该图像中点 (x,y) 处像素的灰度值,灰度级为 L,则 $f(x,y) \in [0,L-1]$。若灰度级 i 的所有像素个数为 f_i,则第 i 级灰度出现的概率为:

$$P(i) = \frac{f_i}{M * N} \tag{4}$$

$$\sum_{i=0}^{L-1} P(i) = 1 \tag{5}$$

将图像中的像素按灰度级用阈值 t 划分为两类,即背景 c_0 和目标 c_1[14]。背景 c_0 的灰度级为 0 ~ $t-1$,目标的灰度级为 t ~ $L-1$。背景 c_0 和目标 c_1 对应的像素分别为 $f(i,j)$。背景 c_0 部分出现的概率和目标 c_1 部分出现的概率分别为:

$$w_0 = \sum_{i=0}^{t-1} P(i) \tag{6}$$

$$w_1 = \sum_{i=t}^{L-1} P(i) \tag{7}$$

其中,$w_0 + w_1 = 1$。

背景 c_0 部分和目标 c_1 部分的平均灰度值分别为:

$$u_0(t) = \sum_{i=0}^{t-1} i * \frac{P(i)}{w_0} \tag{8}$$

$$u_1(t) = \sum_{i=t}^{L-1} i * \frac{P(i)}{w_1} \tag{9}$$

图像的总平均灰度值为:

$$u = \sum_{i=0}^{L-1} i * P(i) \tag{10}$$

图像中背景和目标的类间方差为:

$$\delta^2(k) = w_0 (u - u_0)^2 + w_1 (u - u_1)^2 \tag{11}$$

令 k 的取值从 0 ~ $L-1$ 变化,计算不同 k 值下的类间方差 $\delta^2(k)$,使得类间方差 $\delta^2(k)$ 最大时的那个值就是所要的最佳阈值。

2 图像分割

图像分割就是把图像分成若干个特定的、具有独特性质的区域并提出感兴趣目标的技术和过程。图像分割方法近年来又有新型神经网络图像分割法,文中使用阈值分割法进行图像分割。

阈值分割方法实际上是输入图像 f 到输出图像 g 的变换[15],公式如下:

$$\begin{cases} g(i,j) = 1, f(i,j) \geqslant T \\ g(i,j) = 0, f(i,j) \leqslant T \end{cases} \tag{12}$$

其中,T 为阈值,对于物体的图像元素 $g(i,j) = 1$,对于背景的图像元素 $g(i,j) = 0$。

如果能确定一个合适的阈值就可准确地将图像分割开来。阈值确定后,将阈值与像素点的灰度值逐个进行比较,而且像素分割可对各像素并行地进行,分割的结果直接给出图像区域[16]。

3 汉字识别

前边的图像处理都是为此步骤做铺垫。二值化和图像分割已经将字符提取出来,但是以单个汉字为基础的识别要将每个字从图像中提取出来[17]。文中的汉字识别,主要借助百度的 OCR 识别技术,识别出通

用字。

此测试环境需要的配置如下:

(1)百度官方的 Java SDK 压缩包;

(2)Jdk 需要 1.7 以上;

(3)IDE 使用 Eclipse 新建工程,导入下载的工具包;

(4)配置通用文字识别的客服端,以及服务器代理设置,代码如下:

```
Public class test {
Public static void main(string[ ] args) {
AipOcr Client=new AipOcr(app_id,api_key,secret_key);
//网络设置
Client. setConnectionTimeoutInMillis(2000);
Client. setSocketTimeoutInMillis(60000);
//代理设置,http 代理
Client. setHttpProxy("proxy_host",proxy_port);
//接口调用
String path="test. jpg";
JSONObject res = client. basicGeneral( path, new HashMap<
string,string>( ));
System. out. println(res. toString(2));
}
}
```

上述代码中接口说明如表 1 所示。

表 1　测试代码接口说明

接口	说明
setConnectionTimeoutInMillis	建立连接的超时时间(单位:毫秒)
setSocketTimeoutInMillis	通过打开的连接传输数据的超时时间(单位:毫秒)
setHttpProxy	设置 HTTP 代理服务器
Width	定位位置的长方形的宽度

向服务器请求识别某张图片中的所有文字。Java 语言的配置代码如下:

```
public void sample(AipOcr client) {
    //传入可选参数调用接口
     HashMap < String, String > options = new HashMap <
String,String>( );
    options. put("language_type","CHN_ENG");
options. put("detect_direction","true");
    options. put("detect_language","true");
    options. put("probability","true");

    //参数为本地路径
    String image="test. jpg";
    JSONObject res=client. basicGeneral(image,options);
    System. out. println(res. toString(2));

    //参数为二进制数组
    byte[ ] file=readFile("test. jpg");
    res=client. basicGeneral(file,options);
    System. out. println(res. toString(2));

    //通用文字识别, 图片参数为远程 url 图片
    JSONObject res = client. basicGeneralUrl(url, options);
    System. out. println(res. toString(2));

}
```

Postman 使用 post 方法请求识别图片文字,请求参数设置如表 2 所示。

表 2　url 请求参数设置

参数名字	默认值	说明
Image		本地图片路径或者二进制数据
URL	https://aip. baidubce. com/rest/2. 0/ocr/v1/general	图片完整 URL,URL 长度不超过 1 024 字节,URL 对应的图片 base64 编码后大小不超过 4 M,最短边至少 15 px,最长边最大 4 096 px,支持 jpg/png/bmp 格式,当 image 字段存在时 url 字段失效
language_type	CHN_ENG	CHN_ENG,中英混合

请求返回结果参数说明如表 3 所示。

表 3　请求结果参数说明

参数	说明
words_result_num	图片上文字的个数
Words	汉字
Height	定位位置的长方形的高度
Width	定位位置的长方形的宽度

4 结束语

介绍了 OCR 识别的过程和相应模块的代码实现，理论公式推导。国内在 OCR 方向上的发展是很迅速的，尤其是国内公司在其上的应用，这种技术已经渗透到日常手机打字的软件百度输入法中。文中的二值化方法只是平常方法中的一个，还有其他很多方法未涉及，图像分割也是如此。

OCR 大体可以分为两类：手写体识别和印刷体识别。文中使用 Java 语言基于百度 OCR 的 API 实现 OCR 扫描识别印刷体图片上文字的一个客户端（普通文字识别），操作可以借助百度公司 OCR 的 API 利用抓包工具 postman 向其服务器发送 post 请求，在请求参数中带上一张带有文字的图片或者使用实现的客户端进行图片上传识别。还介绍了图像文字识别刚开始的图像处理，其中图像分割步骤至关重要，是识别率高的关键点。文中默认识别通用文字中文，高精度的识别或者带位置信息的识别，生僻字的识别，此处不做研究。Post 请求提交时，请求头 Header 设置为 x-wwww-form-urlencoded 形式。

参考文献：

[1] 靳天飞. 脱机手写汉字识别中笔段提取算法研究[J]. 山东大学学报：理学版，2008，43(5)：39-44.

[2] RADWAN M A，KHALIL M I，ABBAS H M. Neural networks pipeline for offline machine printed arabic OCR[J]. Neural Processing Letters，2018，48：769-787.

[3] 肖 坚. 基于学习的 OCR 字符识别[J]. 计算机时代，2018(7)：48-51.

[4] 麦尔旦·吐拉江. 基于光学字符识别维汉翻译软件的研究与实现[D]. 乌鲁木齐：新疆大学，2018.

[5] 邓子平. 基于图像处理的叶轮给煤机控制系统[J]. 电子技术与软件工程，2019(6)：85-86.

[6] 胡冀川. 基于请求内容的 Web 应用 QoS 方法研究[D]. 北京：北京邮电大学，2018.

[7] 郭宪军，赵海旭，姚 新，等. 声呐图像分割中的改进 Otsu 算法[J]. 声学与电子工程，2018(2)：1-4.

[8] 李育冰，安小旭，蒙 杰. 基于激光图像处理的接触网测量方法[J]. 电气化铁道，2019(2)：60-63.

[9] 张宝华，刘 鹤. 采用子带分量阈值估计的红外图像去噪方法[J]. 中国激光，2014，41(8)：224-231.

[10] GILL H S，KHEHRA B S，SINGH A，et al. Teaching-learning-based optimization algorithm to minimize cross entropy for selecting multilevel threshold values[J]. Egyptian Informatics Journal，2019，20(1)：11-25.

[11] 韩 萍，刘则徐. 基于灰度级分组的 X 光行李图像增强改进方法[J]. 中国民航大学学报，2011，29(4)：23-26.

[12] 燕红文，邓雪峰. OTSU 算法在图像分割中的应用研究[J]. 农业开发与装备，2018(11)：103.

[13] 颜 微. 改进的二维阈值图像分割方法[D]. 湘潭：湘潭大学，2016.

[14] 任红萍，陈敏捷，王子豪，等. 二值网络的分阶段残差二值化算法[J]. 计算机系统应用，2019，28(1)：38-46.

[15] 张 然，陈 权，牛青松，等. 一种基于种子优化算法的图像分割方法[J]. 电脑知识与技术，2019，15(6)：193-197.

[16] 吴 睿. 基于遗传算法的多级自动阈值分割方法研究[J]. 电视技术，2018，42(9)：11-14.

[17] RYAN M，HANAFIAH N. An examination of character recognition on ID card using template matching approach[J]. Procedia Computer Science，2015，59：520-529.

（上接第 25 页）

[12] 陈颖频，彭真明，李美惠，等. 基于交叠组稀疏广义全变分的地震信号随机噪声衰减[J]. 石油地球物理勘探，2019，54(1)：24-35.

[13] LI S，HE Y M，CHEN Y P，et al. Fast multi-trace impedance inversionusing anisotropic total p-variation regularization in the frequency domain[J]. Journal of Geophysics and Engineering，2018，15(5)：2171-2182.

[14] LIU X，CHEN Y P，PENG Z M，et al. Infrared image super-resolution reconstruction based on quaternion fractional order total variation with Lp quasinorm[J]. Applied Sciences，2018，8(10)：1864.

[15] WANG L Z，CHEN Y P，LIN F，et al. Impulse noise denoising using total variation with overlapping group sparsity and Lp-pseudo-norm shrinkage[J]. Applied Sciences，2018，8(11)：2317.

[16] CHARTRAND R. Shrinkage mappings and their induced penalty functions[C]//IEEE international conference on acoustics，speech and signal processing. Florence：IEEE，2014：1026-1029.

[17] WU L，CHEN Y P，JIN J，et al. Four-directional fractional-order total variation regularization for image denoising[J]. Journal of Electronic Imaging，2017，26(5)：053003.

[18] WOODWORTH J，CHARTRAND R. Compressed sensing recovery via nonconvex shrinkage penalties[J]. Inverse Problems，2016，32(7)：1-25.

[19] DUAN F B，CHAPEAU B F，ABBOTT D，et al. Non-Gaussian noise benefits for coherent detection of narrowband weak signal[J]. Physics Letters A，2014，378(26-27)：1820-1824.

[20] REN Z，HE C，ZHANG Q. Fractional order total variation regularization for image super-resolution[J]. Signal Processing，2013，93(9)：2408-2421.

[21] WANG Z，BOVIK A C，SHEIKH H R，et al. Image quality assessment：from error visibility to structural similarity[J]. IEEE Transaction on Image Processing，2004，13(4)：600-612.