垃圾邮件图像中的文字角点检测

万明成,耿 技,程红蓉,曾志华
WAN Ming-cheng,GENG Ji,CHENG Hong-rong,ZENG Zhi-hua

电子科技大学 计算机科学与工程学院,成都 610054
School of Computer Science and Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China
E-mail:mcwan@foxmail.com

WAN Ming-cheng,GENG Ji,CHENG Hong-rong,et al.Detecting corners of text in spam images.Computer Engineering and Applications,2009,45(14):170-172.

Abstract: Combining edge extraction methods and a circular template,this paper proposes a new method for detecting corners of text in spam images.Firstly,the edges of text in spam images are extracted by color-roberts algorithm and threshold segmentation. Then,the corners is detected by using a circular template.Most noise in spam image is eliminated by edge extraction and threshold segmentation.Besides,the novel circular template allows the corner detection is insensitive to the orientation of text.A comparative experiment engages SUSAN method and the proposed method for the corners detection of text in images from a real spam image archive.The results indicate that the method has a better performance and can obtain the angle magnitude of corners contemporarily.
Key words: corner detection;edge extraction;circular template;spam image

摘 要:为提取垃圾邮件图像中文字的角点信息,提出一种新的基于图像边缘和圆形模板的角点检测算法。算法首先利用彩色边缘检测算子和阈值分割方法获取文字图像的边缘,然后采用圆形模板提取文字的角点信息。边缘检测和阈值分割降低了干扰背景和噪声对角点检测的影响,圆形模板使得角点检测对文字方向变化不敏感。实验表明,在真实的垃圾邮件图像中文字角点定位精度略高于 SUSAN 算法,并能同时获取角点角度的大小。
关键词:角点检测;边缘检测;圆形模板;垃圾邮件图像
DOI:10.3778/j.issn.1002-8331.2009.14.052 **文章编号**:1002-8331(2009)14-0170-03 **文献标识码**:A **中图分类号**:TP391.4

1 前言

文字包含有丰富的角点信息,通过对角点信息的提取可以有效定位图像中的文本区域。目前,角点信息已被广泛应用于各种图像的文本区域定位中,包括灰度图像[1]、自然场景图像[2]、视频帧图像[3]等。典型的角点检测算法可以分成两类:基于灰度的角点检测算法[4]和基于边缘的角点检测算法[5-6]。但是对于垃圾邮件图像而言,文字角点的提取相对困难。一方面垃圾邮件图像多为彩色图像,彩色图像到灰度图像的转换可能丢失颜色信息;另一方面垃圾邮件图像含有大量干扰,Wang[7]等人已分析出 17 种随机化干扰方式,包括光栅、短线、噪声点等。并且,文字角点分布密集,一些算法可能在检测一个字符的角点时受到邻近字符的影响。另外,仅有较少的角点检测算法能够在定位角点的同时计算出角点处张角的大小[8-9]。

提出一种基于文字边缘,利用圆形模板检测文字角点的算法。实验表明,将该算法应用于提取实际的垃圾邮件图像中的文字角点时,不仅能获得较高的文字角点定位精度,同时还能获取角点的近似角度值。

2 彩色边缘检测与阈值分割

边缘检测算法的选取至关重要,文字边缘的定位精确度直接影响到角点定位的准确度。基于灰度图像的边缘检测算子需要对彩色图像做灰度转换,可能导致颜色信息丢失,不利于边缘检测。为此,选用张引[10]等人提出的彩色边缘检测算子——ColorRoberts 进行边缘检测。该算子首先利用扩展的 Roberts 算子将彩色图像转换至灰度边缘图像 I_1,然后利用 log 算子作二次边缘提取以获得最终的单像素边缘图像 I_2。

垃圾邮件图像通常含有大量的干扰背景,利用 log 算子进行二次边缘提取可能会强化干扰背景的边缘,从而影响角点定位的精度。通过对大量垃圾邮件图像的深入分析后发现,经扩展 Roberts 算子变换后获得的灰度边缘图像中, 干扰背景点的灰度值通常较低, 而文字边缘点的灰度值较高, 一般不低于 80。另外,Roberts 算子是一种利用局部差分算子寻找边缘的算子。利用扩展的 Roberts 算子获得彩色图像的灰度边缘图像以后,在边缘的拐角处,灰度值通常较低。其它边缘点的灰度值大约是拐角处边缘点灰度值的$\sqrt{2}$倍。如果不对这些拐角处的

基金项目:国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z41)。
作者简介:万明成(1985-),男,硕士研究生,主要研究方向:图像处理、信息安全;耿技(1963-),男,教授,主要研究方向:计算机网络、信息安全;程红蓉(1975-),女,博士生,讲师,主要研究方向:信息安全、模式识别及其应用研究;曾志华(1983-),男,硕士研究生,主要研究方向:模式识别、自然语言处理。
收稿日期:2008-03-17 **修回日期**:2008-06-10

边缘点做特殊处理，则这些拐角点可能丢失，从而导致文字边缘出现断裂现象。

为此，首先对灰度边缘图像 I_1 做阈值分割来去除大量干扰背景的边缘点和一些非文字边缘点，然后利用非极大值抑制来获取文字边缘。垃圾邮件图像中的文字边缘像素点数量通常占整幅垃圾邮件图像中像素点数量的比例较大，但基本上不会超过 12%。因此，本文的分割阈值选取算法如下：

```
int Th0=N*0.12;
int count=0;
int Th=80;//存放分割阈值
for(int i=255;i>80;i++)
{
  count+=h(i);
  if(count > Th0)
  {
    Th=i;//获得分割阈值
    break;
  }
}
```

算法中 N 为灰度图像 I_1 中的像素点总数，$h(i)$中存放的是 I_1 中灰度值为 i 的像素点数量，Th 为最后获得的分割阈值。阈值分割公式如式(1)所示：

$$I_2(i,j)=\begin{cases} I_1(i,j) & I_1(i,j)\geqslant Th \\ 0 & I_1(i,j)<Th \end{cases} \tag{1}$$

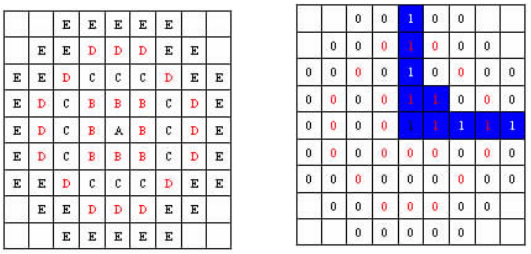
利用式(1)去除灰度图像 I_1 中的大部分干扰背景点以后，采用局部区域非极大值抑制以进一步去除非边缘像素点。一个点是局部区域极大值时，对该点予以保留。为防止出现边缘断裂，当一个点不是局部区域极大值，但是局部区域极大值与该点的灰度值形成近似$\sqrt{2}$倍关系，即该值的 T_1 倍小于局部区域的极大值，该值的 T_2 倍大于局部区域极大值时，对这些点予以保留。在实现时，取 3×3 的窗口作为局部区域，T_1 取为 1.35，T_2 取为 1.5。分割后获得的边缘图像 I_2 可能不是单像素宽。提出的角点检测算法并不要求边缘必须是单像素宽，所以不需要边缘细化处理。

3　基于圆形模板的角点检测算法

关于角点检测目前已做了大量的研究工作，提出了很多算法。但是对角点一直没有形成统一的定义，不同的检测算法对角点可能有不同的定义。Shen[5]将两条或多条直线相交的点定义为角点；Zhang[11]将梯度幅度和梯度方向变化都很大的点定义为角点；侯北平[6]倾向于认为边界上的大曲率点为物体的角点；陈良[12]按照人的视觉响应特点，将角点定义为边界上具有局部特征和全局特征的点的集合。本文将局部区域角度值变化最大的点视为文字的角点，并使用一个圆形模板来检测边缘图像 I_2 中边缘曲线上各点处的近似角度值。

采用圆形模板检测文字角点，使得角点检测算法对文字方向变化不敏感。设计的圆形模板如图 1(a)所示。它是一个半径约为 5 的圆，点 A 为圆心，然后依次为 B,C,D,E 四个圆环。

检测角点时，将待检测的点(i,j)与圆心 A 对齐。为避免邻近文字的边缘影响到角点的判断，本算法只将与待检测点相连通的边缘点在圆形模板内标出，如图 1(b)所示。将由全“0”组成的连通区域称为“0”域，将与圆心连通的“0”域称为有效“0”域。在图 1(b)中存在两个“0”域，但只有左下角那个“0”域才是有效“0”域。对有效“0”域内各圆环上标记为“0”的像素点数量进行统计就可以求出点(i,j)处的近似角度值。角度值的具体计算方式如下：



(a)圆形模板　　(b)边缘像素点分布图

图 1　圆形模板及边缘像素点分布图

$$B_d(n)=45\times[N_b(n)+1] \tag{2}$$

$$C_d(n)=30\times[N_c(n)+1] \tag{3}$$

$$D_d(n)=22.5\times[N_d(n)+1] \tag{4}$$

$$E_d(n)=11.25\times[N_e(n)+1] \tag{5}$$

$$ZA_d(n)=\frac{B_d(n)+C_d(n)+D_d(n)+E_d(n)}{4} \tag{6}$$

$$D_0(i,j)=\max(ZA_d(n)),n=1,2,3\cdots \tag{7}$$

式(2)~(5)中 $N_b(n),N_c(n),N_d(n),N_e(n)$分别代表第 n 个有效“0”域在 B,C,D,E 四环上灰度值为 0 的像素点数量，$B_d(n),C_d(n),D_d(n),E_d(n)$分别代表第 n 个有效“0”域在 B,C,D,E 四环上计算所得到的夹角度数。式(6)中，$ZA_d(n)$为第 n 个有效“0”域上 4 环所得角度值的平均值，并以此作为该有效“0”域的近似角度值。利用公式(7)取点(i,j)周围所有有效“0”域所获得的最大角度值作为该点的近似角度值。垃圾邮件图像中，文字排列方向变化较大，因此角点方向对文本区域定位意义不大。在不考虑角点方向时，X 度和 360–X 度所代表的弯曲程度相同。所以算法通过式(8)将所有角点角度值 $D(i,j)$都限定在(0,180)以内。

$$D(i,j)=\begin{cases} 360-D_0(i,j) & D_0(i,j)\geqslant 180 \\ D_0(i,j) & D_0(i,j)<180 \end{cases} \tag{8}$$

最后对局部区域内角度值变化非极大值的点进行抑制来筛选出最终的角点。所选择的局部区域为与图 1(a)所示的圆形模板同样大小的一个圆形区域。

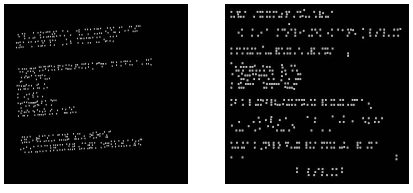
4　实验结果与性能分析

垃圾邮件图像中含有大量的信息，如虚假股票信息，商品广告信息等。获取垃圾邮件图像中的文本信息可以为垃圾邮件的判定提供非常有力的证据。而角点信息提取是获取文本信息极为重要的第一步，为验证本文所提出的算法在提取垃圾邮件图像中文字角点时的性能，从真实的垃圾邮件图像集[13]中选取两幅典型的含有干扰的垃圾邮件图像作为测试图像。图 2(a)为一幅文字倾斜的垃圾邮件图像，并且含有一些光栅式的干扰背景。图 2(b)则为含有短线、噪声点等干扰的垃圾邮件图像，其文字方向为水平方向。分别用 SUSAN 算法和本文所提出的算法来提取图像中的角点信息。

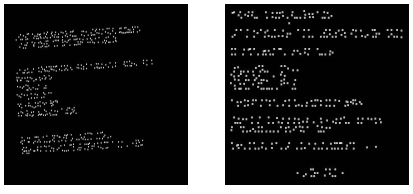
(a)文字倾斜、背景干扰 (b)短线、噪声干扰

图 2 垃圾邮件图像

角点检测结果如图 3 所示,图 3(a)为 SUSAN 算法对两幅图像的角点检测结果,图 3(b)为本文的角点检测算法所获得的角点检测结果。对于文字倾斜的图像(图 2(a)),SUSAN 算法共提取出角点 501 个,本文算法共获得角点 535 个;对于含有短线、噪声点等干扰的图像(图 2(b)),SUSAN 算法共提取出角点 438 个,本文算法共检测出角点数为 418 个。角点检测结果表明,本文算法所识别出的角点数量和角点定位精度与 SUSAN 算法相近。经过分析发现,SUSAN 算法漏检了较多的“X”型角点,并将部分短线和噪声点视为角点。因此本文的算法性能比 SUSAN 算法效果稍好,更加稳定。



(a)SUSAN 算法角点检测结果



(b)本文算法角点检测结果

图 3 角点检测结果

所提出的文字角点检测算法不仅能够有效定位文字角点,还能估计出角点处张角的近似角度值。角度大小往往被一些角点检测算法所忽视,但这些角度对文字区域定位有着极为重要的作用。文字区域角点密集,角度大小的分布也有一定的规律。图 4 为本文算法所提取的两幅测试图像的角点角度值分布图。

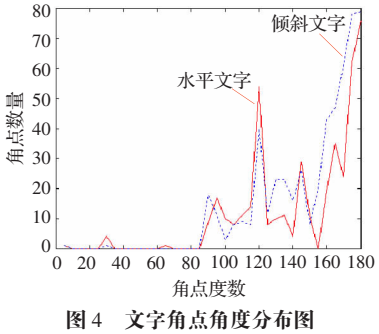


图 4 文字角点角度分布图

可以看出,尽管两幅图像的内容大相径庭,文字方向也不一样,但两条曲线非常相似。这些文字角点角度值的分布规律有助于区分文字区域和非文字区域。因此,本文的角点检测算法所提取的文字角点信息对文本区域定位有极为重要的意义。

5 结论

角点检测的应用领域较多,不同的角点检测算法侧重于特定的应用领域。本文提出的基于边缘和圆形模板相结合的角点检测算法适用于检测有人为干扰的垃圾邮件图像中的文字角点。该算法思想简单,并能处理非单像素宽的文字边缘。由于无复杂操作,其运行速度快。实验表明,本文的算法和 SUSAN 算法相比,对含有人为干扰的垃圾邮件图像进行文字角点检测时有更好的稳定性。在确保提取出图像中大部分角点的同时能估计出角点处张角的大小,为文本区域定位提供更多的信息。因此,本文的算法具有较高的实用性。

参考文献:

[1] Nourbakhsh F,Pati P B,Ramakrishnan A G.Text localization and extraction from complex gray images[C]//Lecture Notes in Computer Science.Berlin:Springer-Verlag,2006:776-785.

[2] Shen Hui-ying,Coughlan J.Finding text in natural scenes by figure-ground segmentation[C]//Proceedings of the 18th International Conference on Pattern Recognition.Hong Kong:IEEE,2006:113-118.

[3] Guo Ge,Jin Jin,Ping Xi-jian,et al.Automatic video text localization and recognition[C]//Proceedings of the Fourth International Conference on Image and Graphics.Chengdu:IEEE,2007:484-489.

[4] Smith S M,Brady J M.SUSAN-a new approach to low level image processing[J].International Journal of Computer Vision,1977,23(1):45-78.

[5] Shen Fei,Wang Han.Corner detection based on modified hough transform[J].Pattern Recognition Letters,2002,23:1039-1049.

[6] 侯北平,李平,宋执环.基于滑动窗口的自适应角点检测研究[J].电路与系统学报,2006,11(6):133-137.

[7] Wang Z,Josephson W,Lv Q,et al.Filtering image spam with near-duplicate detection[C]//CEAS 2007-Fourth Conference on Email and AntiSpam,California,2007.

[8] 王克勇,郑链,潘乐义,等.一种实用的目标图像角点特征提取方法[J].探测与控制学报,2003,25:10-12.

[9] 殷润民,柴旭东,李伯虎.基于灰度差统计的角点检测方法[J].计算机工程,2006,32(22):184-186.

[10] 张引,潘云鹤.复杂背景下文本提取的彩色边缘检测算子设计[J].软件学报,2001,12(8):1229-1235.

[11] Zhang Kun-hua,Wang Jing-ru,Zhang Qi-heng.Corner detection algorithm based on multi-feature[C]//Proceedings of International Society for Optical Engineering,INIST,2001:85-90.

[12] 陈良,高成敏.基于边界的角点无阈值识别算法[J].计算机工程,2006,32(11):200-205.

[13] Princeton spam image benchmark[EB/OL].http://www.cs.princeton.edu/cass/spam/.