基于图像型垃圾邮件过滤系统的研究

福州大学物理与信息工程学院　代立华　黄立勤

【摘要】在互联网技术迅猛发展的背景下，电子邮件凭借着成本低、方便快捷的特点在人们日常交流和沟通中的应用越来越广泛。但需要注意的是，大量垃圾邮件的出现严重影响了人们的使用体验，尤其近年来图像型垃圾邮件的出现，给众多用户和企业带来了严重的困扰，甚至导致各种损失的出现，这就给垃圾邮件过滤软件提出了更高的要求。基于以上，本文提出了一种基于OCR过滤方法的图像型垃圾邮件过滤系统，分析了图像向垃圾邮件的过滤策略、文本检测和文本识别，研究了图像型垃圾邮件过滤系统的设计结构。

【关键词】图像型垃圾邮件；文本检测；文本识别；过滤系统

1 前言

近年来，研究人员对图像垃圾邮件的识别和过滤技术的研究较为关注，但当前研究出的过滤系统都不能够很好的实现垃圾邮件图像的识别和分类，难以满足图像型垃圾邮件过滤的准确性、实时性及高效性要求。基于以上，本文提出了一种以OCR技术为基础的图像向垃圾邮件过滤系统，旨在为相关研究和实践提供参考。

2 图像型垃圾邮件过滤策略分析

就目前来看，图像型垃圾邮件的过滤方法主要包括贝叶斯过滤算法、支持向量机分类算法、黑白名单过滤算法及决策树过滤算法等。本文以传统垃圾邮件检测过滤技术为基础，融合OCR技术（光学字符识别技术），具体的过滤步骤如下：首先采用黑白名单过滤算法对图像型垃圾邮件进行过滤，之后利用OCR技术对图像型邮件中的文本进行提取，最后以支持向量机分类算法为基础，对邮件进行明确的分类，以此来实现对图像型垃圾邮件类型的判断。

OCR技术主要以模式识别方法为基础，能够将带有文字的图像文件转换为可以进行编辑的文本文件，利用OCR软件能够有效提取二值化文档图像中的文字。具体来说，首先需要处理图像，检测出图像中的文本区域，之后进行文本区域的二值化处理，最后提取文字信息[1]。

3 图像文本检测

3.1 提取图像边缘集

3.1.1 求图像边缘

当前有着众多图像边缘检测算法，其中John F·Canny提出的Canny算子检测算法以最优化算法为基础，是最为有效也是应用最为广泛的一种图像边缘检测算法。因此，本文以此方法为基础来对邮件图像的垂直边缘和水平边缘进行检测。具体步骤如下：①采用高斯滤波平滑图像来减少或去除图像噪声；②以一阶微分偏导数有限差分方法为基础，对图像中各个像素点的梯度值和方向进行计算和分析[2]；③采用非极大值方法来实现图像梯度幅值的抑制，④利用双阈值算法，对图像边缘进行检测和连接，尽可能消除图像边缘中的伪边缘段。

3.1.2 图像边缘细化

SPTA细化算法是一种有效的图像边缘细化方法，在处理图像后能够保证图像的圆润性，且能够有效避免出现图像纹理断裂的问题，时间复杂度较低，鉴于SPTA算法的众多优势，本文选此方法来对得到的目标区域边缘图像进行边缘细化处理。具体来说，使用窗口模式来扫描目标区域边缘图像的所有像素，按照一定的规则计算像素点邻域，之后在横向和纵向上对像素点进行检测，判断出可能要删除的像素点和安全像素点，以此来实现对目标区域边缘图像的细化处理。

3.2 候选文本区域融合

经过上述步骤得到的图像边缘连通性优良，边缘清晰、圆润，但需要注意的是，在提取图像边缘集的过程中，一些与图像文本相似的、有着一定规则性的背景物体被保留了下来，为了保证邮件图像关键文字的提取效果，需要对这些背景图像即非文本区域进行去除，去除干扰。具体步骤如下：①以颜色视觉特征为依据，对图像区域进行聚类处理；②使用小波变换方法，分解分布特征近似的区域图像，以此来实现后续处理工作的简化；③构造区域能量图像，利用文字方向投影断层检测算法，对文字块进行构建，从而实现对非文本区域进行去除。

3.3 验证候选文本区域融合

融合图像候选文本区域之后，能够对文字方向进行明确，并取出文字重叠部分，之后的工作需要分离候选文本区域中的文本区域和非文本区域。本文选用支持向量机SVM分类方法，实现相应特征的分类，以此来识别并获取图像中的文本区域。

4 图像文本识别

在采用检测算法处理图像之后，能够识别和筛选出图像中的文本区域，但需要注意的是，要想通过OCR软件对图像中的文字进行处理，需要对图像进行二值化处理，而二值化处理的过程中会受到背景图像的影响，容易引入噪声，影响文本的识别率，因此，在二值化处理图像完成后进行图像去噪是十分必要的。以小波变换为基础的去噪方法十分有效，能够保证原始图像纹理细节的完整性，不会对边缘轮廓造成破坏。在识别的过程中，将文字壁画特征图算法和图像文本颜色层算法结合，通过组合过滤的方式来保证获取二值化图像的高质量。

5 过滤系统

根据上述步骤来提取图像中的文本信息之后，将文本信息与事

（下转第178页）

气势上占据绝对优势。然后按照六丁六甲排列放置就成了六丁六甲阵。整个矩阵如果一半拉成线（可随意变化），一半如同四门兜底阵一般，即电磁蜘蛛网组成的北斗七星阵。

6.郯庐断裂带

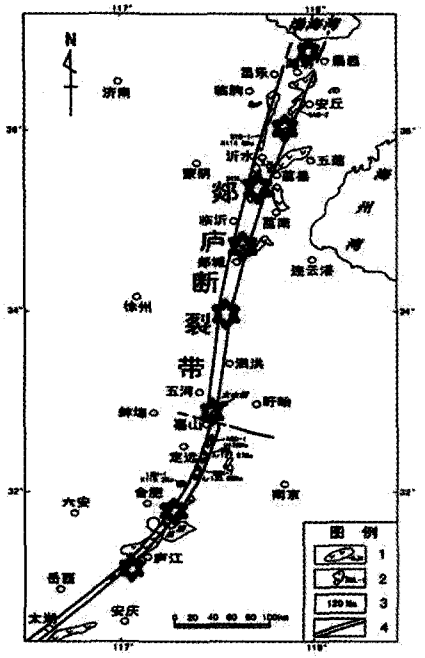

图1 已经在郯庐断裂带安放的电磁蜘蛛网

已经在郯庐断裂带安放的电磁蜘蛛网如图1所示。郯庐断裂带在中国境内长约二千多公里，从经安徽、江苏、山东，进入渤海，之后又在辽宁登陆，进入东北以后，分为了两支（另一说是分三支），然后继续北上，进入俄罗斯。它形成于2亿年前的三叠纪，那个时候恐龙才开始出现。它的形成应该与华北地块、华南地块的碰撞紧密相关。从谷歌地图可以看到因华北、华南两个地块碰撞在它们中间形成了一条高耸的山脉，就是秦岭_大别山山脉。由于亿万年风化剥蚀风韵犹在。自这条断裂形成以后，经历了复杂的演化。这次首先在合肥市安放然后从沈阳往下布局安放，前期必须经过了精心选择，谷歌地图上可见有一条非常明显的线郯庐断裂带控制着我国东部的大地构造格架。在郯庐断裂带安放电磁蜘蛛网，具有方便检测到异常电磁场的实际研究价值。郯庐断裂带规模宏伟，结构复杂。是地壳断块差异运动的接合带，是地球物理场平常带和深源岩浆活动带。

参考文献

[1]Jiang Min. Multi point timing control circuit[J].Electric world fourth,1992(4):42-42.

[2]Jiang Min. Sound and light indicator of electronic compass[J]. Electric world,second 1996(2):33.

[3]Jiang Min.Earthquake prediction micro integrated measurement recorder [J] Technology Innovation Herald 2010,(29):22-23.

[4]Min Jiang,2013.11 How to weave the electromagnetic spider web to predict earthquakes,2013 3rd International Conference on Education and Education Management(EEM 2013),p546-551.

（上接第176页）

先构造词库中的敏感词进行比对，确定图像型垃圾邮件的类别。选用来自于Spam Archive数据集中的训练样本与测试样本，采用基于ORC的图像型垃圾邮件过滤系统进行实验。以谷歌OCR开源代码为基础，在相关软件环境下调试来生成可执行文件，获取文本信息后将提取结果在一个文件中保存。

其中广告类图像型垃圾邮件共有200幅，涉嫌违法类图像型垃圾邮件共有200幅，分别为票证类邮件图像100幅，色情类邮件图像50幅，反动类邮件图像50幅，具体过滤实验结果如表1所示：

表1 图像型垃圾邮件过滤系统实验结果

分类	系类别	图像总数量	过滤正确率
广告类		200	90.25%
违法类	票证类	100	88.67%
	色情类	50	86.36%
	反动类	50	81.57%

由表1可知，本文设计的基于ORC的图像型垃圾邮件过滤系统过滤效果良好，对广告类、票证类、色情类及反动类图像型垃圾邮件的过滤正确率都达到了80%以上，值得进一步推广和应用。

6 结论

综上所述，电子邮件在人们日常生活和工作中的应用越来越广泛，但垃圾邮件的出现则影响了用户体验，尤其对于图像型电子邮件来说，其检测、识别和过滤困难。本文以ORC技术为基础提出了一种图像型垃圾邮件过滤系统，探讨了具体系统原理和整个流程，并通过实验验证了系统过滤效果，实验表明，本文提出的图像型垃圾邮件过滤系统过滤效果良好，值得推广和应用。在未来的发展中，应当积极融合各种技术来提升垃圾邮件过滤效果。

参考文献

[1]李鹏,崔刚.图像型垃圾邮件过滤技术研究进展[J].智能计算机与应用,2013,03:28-32+36.

[2]万明成,耿技,程红蓉,陈佳.图像型垃圾邮件过滤技术综述[J].计算机应用研究,2008,09:2579-2582.

作者简介：

代立华（1987—），男，大学本科，助理工程师，研究方向：图像处理。

黄立勤（1973—），男，博士，教授，研究方向：图像处理与计算机视觉、机器学习算法在视频数据分析、医学图像处理及辅助诊断中的应用。