第34卷 第15期 计 算 机 工 程 2008年8月
Vol.34 No.15 Computer Engineering August 2008
·工程应用技术与实现· 文章编号：1000—3428(2008)15—0231—03 文献标识码：A 中图分类号：TP391.1

彩色文档图像的版面分析

黄海凌，刘列根，张 宇
(华南理工大学计算机应用工程研究所，广州 510641)

摘 要：文档图像处理技术是实现对网络上以“图片化”形式发送的垃圾邮件进行检测和过滤的有效手段。该文对彩色文档图像的版面进行分析，目的是分割出图像中的特定目标，便于分析并检测出文档图像中是否含有特别字符信息，从而使得网络垃圾邮件过滤系统可以根据这些信息判断是否过滤该邮件。实验结果表明，上述方法可以在不同颜色深度和不同几何结构的彩色文档图像中进行有效的检测，具有较好的实用性和应用价值。

关键词：文档图像；版面分析；连通元；归一

Color Document Image Layout Analysis

HUANG Hai-ling, LIU Lie-gen, ZHANG Yu
(Research Institution of Computer Application Engineering, South China University of Technology, Guangzhou 510641)

【Abstract】 Document image analysis technology provides an effective tool for filtering junk mails in a graphic form. The aim of analyzing the color document image layout is to segment particular objects in the document image, so that the downstream steps can analyze and inspect whether there are special words in the document image. The network junk-mail-filter system can use this information to identify whether to filter the mail or not. Experiments on this system show that the method is efficient in inspecting different color and gray document images with different geometric structure. The proposed method has potential applications in document image information extraction and filtering.

【Keywords】 document image; layout analysis; connected components; normalization

1 概述

彩色文档图像的应用背景十分广泛，随着网络的大面积普及，彩色文档图像在网络中的应用越来越多，形式也越来越多样化，例如电子广告，电子报刊。但是，随之而来的负面影响也不小，例如在垃圾邮件中时常会夹杂着电子广告，或是某些非用户所订阅的电子报刊。

目前，防堵垃圾邮件以内容过滤、语意分析技术居多。随着垃圾邮件发送者手段的越发高明，发送垃圾邮件的手法多变且语意内容不断变化，以这种彩色文档图像的形式发送垃圾邮件就能完全躲避传统的关键字过滤。

本文就是在此基础之上，利用文本图像分析技术，实现彩色文档图像的版面分析，为进一步识别出“图片化”形式的垃圾邮件提供技术手段。

目前版面分析主要有两类方法：自顶向下(top to down method)和自底向上(bottom to up method)[1]。自顶向下的方法重视全局图像信息，从整个图像入手，根据对文档版面的形式语言描述，将文本递归分割成足够小的区域，是一个分解的过程。自底向上的方法重视局部图像信息，从图像细节入手，将图像小区域逐步合并成较大区域，是一个合并的过程。本文所采用的方案是以自底向上分析为主。

而近年来，从图像中提取文字在国内外也都有大量的研究。Ohya 使用灰度门限法对西文字符进行分割[2]；Lopresti 和 Zhou 使用图像分析法对 lnternet 上的静态图像进行了文字分割[3]；文献[4]提出了一个分 4 步从图像中检测和抽取文字的系统；文献[5]介绍了一种适合报纸、网页和普通的图像的文字定位方法。本文是根据印刷字体本身的特点以及彩色文档图像的特点来提取文字的。

2 系统框架

为了更好地理解和把握本方法和系统，先简要地介绍系统的框架结构，如图 1 所示。

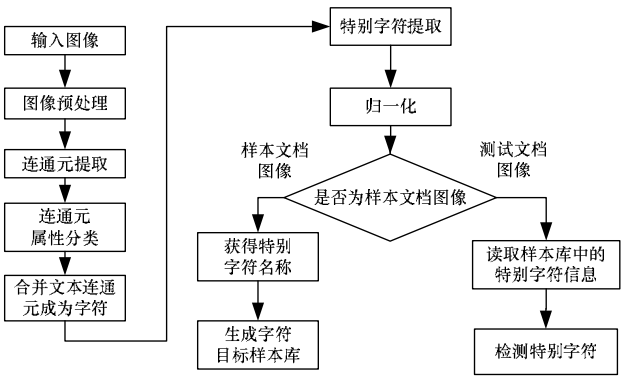

图 1 系统框架

为了能够对彩色文档图像实现与颜色特征无关的版面分析处理，必须先对文档图像进行预处理(二值化等)。接着，提取文档图像中的连通元，并对这些连通元进行分析、分类，然后合并文本属性的连通元，提取出特别字符的特征信息。为了检测方便，还需要对所提取的特别字符进行归一化。

如果用户输入的是一个样本文档图像，系统会先提示用户输入所提取出的特别字符的名称，并连同特征信息一起生成字符目标样本库，以供检测之用。如果用户输入的是一个

作者简介：黄海凌(1985－)，女，硕士研究生，主研方向：电信网络，图像处理；刘列根，硕士；张 宇，高级工程师、博士
收稿日期：2007-09-10 **E-mail：**helly505@vip.sina.com

测试文档图像，系统则会先读取字符目标样本库中关于特别字符的特征信息，再将这些信息与当前在测试文档图像中提取的特别字符的特征信息进行一一对比，得出最终检测结果。

3 基于连通元的分析方法

本文定义连通元 R_i 是八元组 $R_i=(x_l,y_l,x_r,y_r,x_c,y_c,z_f,c_n)$，其中，$(x_l,y_l)$ 为连通元的矩形区域 R_i 左上角的坐标；(x_r,y_r) 为右下角的坐标；(x_c,y_c) 为 R_i 的中心点的坐标，且它们之间满足以下关系：

$$x_c=(x_l+x_r)/2 \tag{1}$$

$$y_c=(y_l+y_r)/2 \tag{2}$$

z_f 为该矩形区域的标识，用来标识该连通元的属性(文本、图片等)，所有关于标志的处理都会修改到该项；c_n 表示该连通元的彩色信息。本文采用基于连通元成分标记算法提取连通元。

3.1 连通元属性分类

连通元属性分类的主要目的是提取文本信息，同时去除文档图像中非文本属性的连通元，以便后面的分析工作。

可以通过面积和矩形区域的长宽以及它们各自的彩色信息去除这些非文本属性的连通元。但是有时图片信息元中会存在一些小的连通元或“噪音”，这些小的连通元也应该与图片信息元一同去掉，所以应该先进行包含结构的合并，这样就可以把图片及其“噪音”全部归并为一个大的连通区域，方便之后的其他处理工作。

主要的分类规则如下：

(1)查找所有连通元，若连通元的长、宽或面积大于给定阈值，则入选，记作 R_u；否则根据彩色信息 c_n 判别其是否属于图片类型，并修改其属性标志 z_f。

(2)查找连通元 R_u 所包含的全部子连通元，根据这些子连通元的中心点位置判别它们是否符合某种规律，若中心点几乎在一条直线上，则判定 R_u 为文本标题框，否则标记 R_u 为图片连通元或长条分割线连通元。

(3)若 R_u 为文本标题框，则标记其所包含的全部子连通元为文本属性，而 R_u 本身标记为文本框属性。这样就能很好地将又大又长的文字标题框去掉，而保留框内的文字元素，从而可以大大节省后面的分析时间。

图 2 示出了原文档图像连通元经过分类处理后的连通元提取情况。图片、长的分割线(以及文本框)不再属于感兴趣的连通元，只留下用矩形框来表示的文本属性连通元。

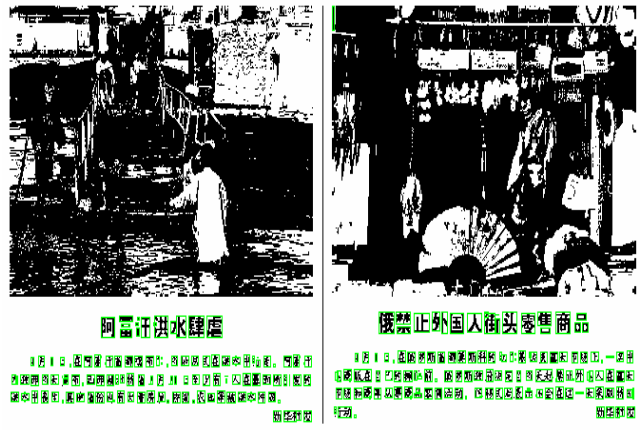

图 2 分类后的文档图像

3.2 合并文本属性连通元

合并文本属性连通元，使之成为独立的字符，是为了能够更加准确地提取文档图像中的特别字符。合并规则以及过程控制规则成为合并的关键。过程控制规则是规定如何使用所构造的合并规则的规则。

在合并过程中，难免会有误合并的情况，所以，合并规则中要提供相应的误合并处理规则，以避免误合并，一个有效的方法就是通过比较平均特征参数来确定当前是否进行合并。可以通过对文档图像中的汉字的大小(长与宽)进行统计计算，求出文档图像中汉字的平均特征参数，包括平均宽度、平均高度等。此外，还可以求出文档图像中字体的平均长宽比值。这是因为本文所分析的文档图像，其字体都是印刷字，而印刷字字体都是方块字，长宽比值一般在[0.8, 1.2]之间。采用平均特征参数还不能很好地处理版面的实际情况，尽管提高了合并速度，但仍会有连通元的误合并问题。对平均特征参数进行适当修正，即可进一步降低误合并。

下面给出几个实验实例。图 3 是版面分割成 2 个重叠的连通元，将 2 个字块合并起来，可以提取整个字。图 4 是包围结构的字块合并。图 5 是上下结构的字块合并。图 6 是左右结构的字块合并。

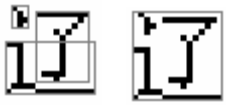

图 3 重叠连通元合并

图 4 包围结构连通元合并

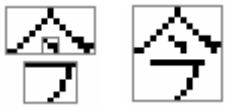

图 5 上下结构连通元合并

图 6 左右结构连通元合并

4 特别字符的目标检测

网络垃圾邮件中的彩色文档图像，特别是以电子报刊形式发送的垃圾邮件，都有醒目的主题，如“XX 周刊”。如果垃圾邮件过滤系统能够预先存储这些特别字符的信息，那么一旦检测到相关的特别字符，就可以将这些邮件定义为垃圾邮件。这些特别信息有时也可以是某种 Logo。

目前还只是根据连通元的面积来查找特别字符，因为大部分醒目的标题都会是文本属性连通元中面积最大的。对于与该连通元处于同一直线上，且特征信息(如面积、长宽比)相似的所有文本属性连通元，标识这些连通元属性为特别字符。同时还可以根据特别字符的连通元数据，计算出相关的特征信息，如位置信息、几何关系结构。

4.1 字符归一化

在提取到特别字符的特征信息后，为了清除因各版面字符不同而带来的大小、形状等方面的变化，还需要进行归一化处理，以利于生成样本库和检测的进行。本文采用线性归一化的方法，把不同大小、不同字号的字体映射到一个 64×64 的方格模板内，以达到标准化的目的。

假设新的标准模板中，坐标为 (i,j)，其中，$i=0,1,\cdots,63$，$j=0,1,\cdots,63$。而原始字体大小为 $m\times n$，则新坐标 (i,j) 对应到原图中的坐标为 $x=\frac{n}{64}\times i$，$y=\frac{n}{64}\times j$，为了使所对应的值更加精确，对 x,y 进行修正，采用四舍五入的方法，即：

$$x=\frac{n}{64}\times i+0.5 \tag{3}$$

$$y = \frac{n}{64} \times j + 0.5 \quad (4)$$

在经过上面的映射变换之后，可以生成特别字符的样本库。根据检测需要，归一化后的信息都应保存在样本库中，包括印刷字体的点阵信息等。

4.2 生成字符目标样本库

在提取到特别字符的特征信息并且归一化之后，可以让系统生成字符目标样本库。系统会提示使用者输入所提取的特别字符名称，并连同其图像信息一起存储在样本库中。此后，文档图像中一旦含有这些特别字符的信息，就可以检测出来，并提示使用者。

4.3 特定字符的匹配

特别字符的检测建立在有相关字符目标样本库的基础上。在系统生成样本库之后，就可以开始特别字符的检测。在这之前，系统同样会先进行测试文档图像的版面分析，提取出该版面的特别信息，经过归一化处理从样本库中读取数据，进行检测。检测时，系统会将这些从样本库中读取到的信息与在测试文档图像中提取到的特别字符的特征信息相对比，如果两者有很高的相似度，则可以认为它们是表示相同的特别字符，说明需要关注该文档图像。

本文检测工作的主要思想是将 2 个二维矩阵信息进行一一对比。这个检测过程只是简单地告诉使用者，在测试文档图像中是否包含字符目标样本库中，需要关注哪些特别字符，以及可能会是什么字符。

在特别字符的目标检测结果中，所提取出的特别字符用红色矩形框表示，将这些特别字符信息分别与样本库中的特别字符信息相匹配，其相似度高于系统所设定的阈值，表明该测试文档图像含有样本库中的特别字符。

5 结束语

由于垃圾邮件过滤目前还仅仅依靠关键字的语意过滤，因此迫切需要研究一种可行的网络彩色文档图像检测方法。本文针对上述问题，提出一种依靠检测彩色文档图像的特别字符的方法，实现对不同颜色深度的彩色文档图像进行版面分析。

按照上述方法开发了一个实验系统并进行了测试。实验结果表明，上述方法可以有效地对不同颜色深度和不同几何结构的彩色文档图像进行版面分析，检测出是否含有特别字符，具有较好的实用性和应用价值。但是还是有很多不足的地方，如合并文本属性连通元、处理复杂结构的文档图像。

参考文献

[1] 王浩军, 赵南元, 邓钢轶. 藏文识别的预处理[J]. 计算机工程, 2001, 27(9): 93-96.

[2] Ohva J, Shin A, Akamalsu S. Recognizing Characters in Scene Images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(2): 214-220.

[3] Lopresti D, Zhou Jin. Document Analysis and the World Wide Web[C]//Proc. of International Workshop on Document Analysis Systems. Malvern, PA, USA: [s. n.], 1996.

[4] Victor W, Manmatha R, Riseman E M. Finding Text in Images[C]// Proc. of the 2nd ACM International Conference on Digital Libraries. Philadelphia, PA, USA: ACM Press, 1997: 23-26.

[5] Jain A K, Yu Bin. Automatic Text Location in Images and Video Frames[J]. Pattern Recognition, 1998, 31(12): 2055-2076.

（上接第 230 页）

征变化较大的攻击事件各节点之间相关程度较低，而本文引入的 BIC 评分函数补偿的相关性出现了过度拟合的情形。为了克服此缺陷，可以加大 BIC 函数中的惩罚项，使补偿的相关性与实际情形更接近，以提高此类攻击行为的识别率。

4.4 性能评价

本文主要研究网络攻击的漏报率和误报率，在推理过程中主要考察对 DoS 攻击(如 land 攻击、neptune 攻击、namp 攻击、pod 攻击、smurf 攻击、teardrop 攻击)、经常作为 DoS 攻击前奏的刺探攻击(如 ipsweep 攻击、portsweep 攻击、nmap 攻击、satan 攻击)及正常数据包的识别能力。

为了进行性能评价，本文引用文献[7]的研究成果进行比较。文献[7]提出基于信息增益原理的改进贝叶斯网络模型，对贝叶斯网络中的特性节点进行特征选择，并删除一些冗余属性，以达到优化贝叶斯网络的目的。

表 3 基于相同推理数据集检测的识别率比较

攻击名称	BIC 评分贝叶斯网络	信息增益贝叶斯网络	神经网络
ipsweep	0.980 392	0.815 6	0.243 2
land	0.555 556	0.977 0	0.952 9
neptune	0.987 948	0.926 0	0.864 0
nmap	0.952 381	0.834 0	0.886 0
normal	0.992 764	0.926 4	0.901 1
pod	0.908 046	0.780 0	0.260 0
portsweep	0.923 729	0.963 2	0.917 8
satan	0.853 464	0.826 0	0.794 0
teardrop	1.000 000	0.985 6	0.935 4

表 3 列出了基于 BIC 评分函数的贝叶斯网络模型、基于信息增益学习的贝叶斯网络模型及基于神经网络的异常检测模型对相同推理数据集的检测结果。由表 3 可以看出，本文 BIC 评分贝叶斯网络模型对大部分 DoS 攻击和刺探攻击的识别率高于文献[7]基于信息增益的贝叶斯网络模型对相应攻击的识别率。

5 结束语

改进的贝叶斯网络模型对行为特征变化较大的未知攻击进行检测时，由于其 BIC 函数存在强拟合特性，因此降低了它对未知攻击的识别能力。为了克服上述缺陷，可以加大 BIC 函数中的惩罚项，使补偿的相关性与实际情形更接近，从而提高对此类攻击行为的识别率。

参考文献

[1] Acherenson J. The Sprite Network Operation System[J]. IEEE Computer, 1998, 21(2): 23-36.

[2] 刘 勃, 周荷琴. 基于贝叶斯网络的网络安全评估方法研究[J]. 计算机工程, 2004, 30(22): 111-113.

[3] 李柏生, 林亚平, 鄢喜爱. 基于朴素贝叶斯网络的入侵检测分析[J]. 网络安全技术与应用, 2007, 1(9): 23-25.

[4] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006.

[5] The UCI KDD Archive. KDD99 Cup Dataset[EB/OL]. (1999-10-28). http://kdd.ics.uci.edu/databases/ kddcup99/kddcup99.html.

[6] 李国和. 基于对象分布的连续属性离散化方法[J]. 计算机应用研究, 2006, 23(9): 258-260.

[7] 何 慧, 苏一丹, 覃 华. 基于信息增益的贝叶斯入侵检测模型优化的研究[J]. 计算机工程与科学, 2006, 28(6): 38-40.