文章编号：1672-2892(2010)05-0534-05

基于形态学和窗口特征的快速文本提取

谭 伟，方 超，杜建洪

(复旦大学 通信科学与工程系，上海 200433)

摘 要：提出了一种快速有效提取目标文本的方法。首先用边缘检测和数学形态学方法粗定位出所有文本区域和图形区域，然后通过多个窗口特征筛选出文本区域，最后设置判别规则提取出目标文本。由于减弱了背景区域的干扰，该算法具有较高定位准确度和较低时间复杂度。同时二值化时使用分块OTSU算法，增强了对复杂背景图像的适应性。通过对多类样本的测试，表明该算法具有较强的通用性，而较高的定位准确率证明了该算法的有效性。

关键词：二值化；形态学； 窗口特征；连通

中图分类号：TN911.73 **文献标识码**：A

Fast text extraction and its application based on morphology and window features

TAN Wei, FANG Chao, DU Jian-hong

(Department of Communication Science and Engineering, Fudan University, Shanghai 200433, China)

Abstract: This paper presents a fast and efficient way to extract the target text areas. First, the edge detection and mathematical morphology are used to realize coarse location of text and graphic areas. Then several window features are extracted to distinguish the text areas from graphic areas and thereafter certain rules are set to extract the target text areas. The algorithm has high location accuracy and low time complexity due to weakening the interference of background areas. At the same time, OSTU algorithm is applied to enhance the adaptability of images with complex background during binarization. Test results on several types of samples indicate that the algorithm has strong popularity. The high accuracy of the method has also proved the validity of the algorithm.

Key words: binarization; morphology; window features; connectivity

近年来，图像和视频已成为记录信息的主要载体，文本提取是该领域的重要应用之一。它可应用于视频检索[1]、版面分割[2]、目标检测[3]等领域。应用场景的多样性以及图像文字大小、字体、颜色、方向等的多样性，使文本提取算法成为研究热点。通过对已有文本提取算法的研究发现：Shen Qinghua,LI Shutao 和 James K[4]在文章中提出的基于数学形态学的页面分割方法，可以快速有效地定位出文本区域，但它只适用于对比度较高的图像，不具鲁棒性。Palaiahnakote 和 H Weihua[5]提出的基于滤波器和边缘特征的文本检测方法，适合于视频字幕等文字排列规范的版面，但对字符排列不规范的版面则缺乏适应性。Li Sun 和 Liu Guizhoug[6]提出的基于角点的文本检测方法，对图像分辨力和字符大小不敏感，且在筛选文字区域时需要色彩信息加以约束。Jaakko Sauvola 和 Matti Pietikainen[7]提出的基于窗口特征和连通性分析的版面分割方法，在简单背景和高对比度情况能有效区分文本和非文本区域。但该算法不适用于背景复杂图像，具有一定的局限性。

本文在 Jaakko 方法基础上进行了算法上的改进和新的应用研究实现，提出了基于数学形态学和窗口特征的快速文本提取算法。同时，本文对二值化过程也提出了相应改进，不针对整幅图像进行二值化，而只针对粗定位后的各区域进行二值化，增强了对复杂背景的适应性。

1 算法流程

本节及以下内容以信封为实例进行分析。

收稿日期：2010-01-26；修回日期：2010-03-30

算法流程如图 1 所示：首先通过光学设备获取窗口信函图像。通过样本分析，发现窗口信函的一些特点：

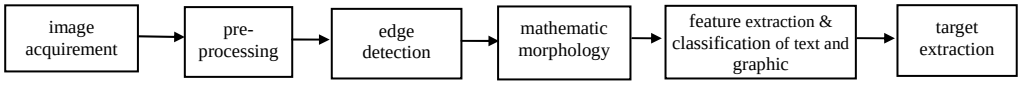


Fig.1 Algorithm flow
图 1 算法流程

1) 背景复杂：图案、花纹、邮戳、邮票、条形码都会对地址区域造成一定的干扰；

2) 字符大小不一，在信封上的位置分布不定。收件人的地址、姓名，发信人的地址、邮编及其他字符信息大小都不同。但字符排列有一定的特征，基本为横(纵)向排列，且同一文本行(列)的字符大小变化不大；

3) 对比度低，且部分区域存在由光照不均引起的灰度带，影响二值化效果。

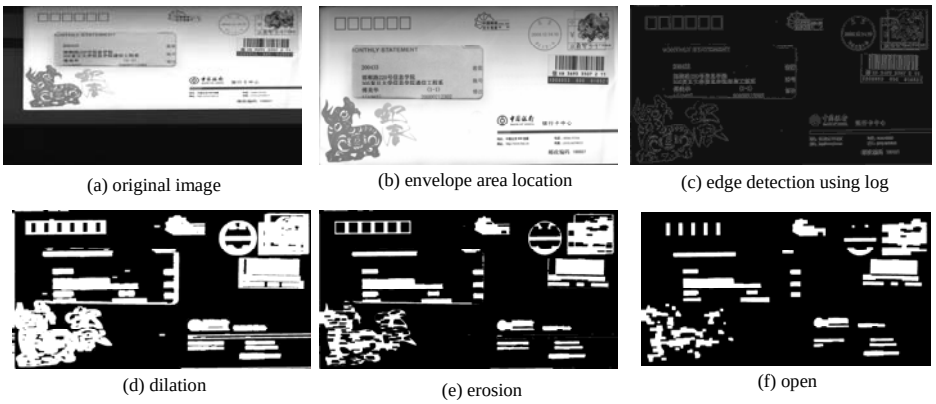
预处理包括灰度化、滤波等。在低对比度情况下，需要进行对比度拉伸。粗定位包括边缘检测和数学形态学，目的是定位出所有文本和图形区域。特征提取及文本图形区域判定属于文本定位阶段，通过窗口特征区分文本区域和图形区域，实现文本区域快速提取。最后阶段为目标提取，通过一定规则筛选出所需信息。

2　算法实现

算法主要包括 3 部分：粗定位、文本定位、目标提取。

2.1 信封区域提取

由于拍摄样本图像包含除信封区域以外的其它信息，如图 2(a)所示，因此粗定位之前需要提取信封区域。通过对图像水平和垂直投影图的分析，利用行与列灰度的阶跃变化可以快速提取出分界位置。提取后效果如图 2(b)所示。



(a) original image　(b) envelope area location　(c) edge detection using log

(d) dilation　(e) erosion　(f) open

Fig.2 Process of edge detection and mathematical morphology
图 2 边缘检测和数学形态学处理

2.2 粗定位

粗定位主要用边缘检测算法和数学形态学方法来定位所有文本和图形区域,这主要是基于文本和图形区域富含较多边缘信息的思想。由于信封背景复杂，对比度低，全局二值化效果不好，用边缘检测方法可以取得较好效果：先检测出边缘(包括强边缘和弱边缘)，然后用数学形态学方法进行连通，标记所有文本和图形区域。具体算法如下：

1) 用边缘检测算法获取边缘图像 I。由于图像对比度较低，如果只检测强边缘，则进行膨胀腐蚀之后文本区域会断裂。由于噪声及背景纹理的影响，如果检测过多弱边缘，形态学处理后会形成大面积的连通域，造成粗定位错误，因此边缘检测算子的选取非常重要。经多次试验，发现选取 log 算子能取得较好的效果(如图 2(c)所示)。

2) 用一定大小的结构算子 Str1 对 I 进行膨胀操作,得到图像 I_1(如图 2(d)所示)。

3) 用一定大小的结构算子 Str2 对 I_1 进行腐蚀操作，得到图像 I_2(如图 2(e)所示)。

4) 为了使得到的区域更加平滑，可以用一定大小的结构算子 Str3 对 I_2 进行开操作，得到图像 I_3(如图 2(f)所示)。膨胀操作时，结构算子选取太大，会把相邻区域连通到一起；如果太小，字符之间不能连接而造成断裂。同样，腐蚀操作时，结构算子过大会将造成字符断裂甚至完全抹去该区域；太小则达不到所需的分割效果。根据

窗口信函字符排列的方向性(基本为横向排列)，在选取膨胀算子时，可将结构算子取为长方形。

5) 标记所有连通区域，用矩形框 Rect $(i)=[L(i),T(i),R(i),B(i)]$ 表示第 i 个区域 $\Omega(i)$。$L(i)=\min(Col(i))$，$T(i)=\min(Row(i))$，$R(i)=\max(Col(i))$，$B(i)=\max(Row(i))$。其中 $(L(i),T(i))$ 为矩形框左上角像素点坐标，$(R(i),B(i))$ 为右下角像素点坐标，$Col(i)$ 为第 i 个区域所有像素点纵坐标的集合，$Row(i)$ 为第 i 个区域所有像素点横坐标的集合，max/min 为求最大/小值函数。

2.3 文本定位

在本算法中，二值化是至关重要的步骤。二值化效果决定各窗口特征，进而影响最终文本提取结果。全局二值化通常不能正确对背景与前景进行分割，虽然最大区间类算法(OSTU)[8]通过自适应确定阈值，能取得较好的分割效果，但在本应用中，如果对整幅图像采用 OSTU 算法，由于受信封灰度带的影响，部分区域二值化效果不好。如果采用局部二值化算法，即便使用积分图像来减少运算量[9-10]，但如果图像分辨力较大，也会导致处理时间过长而不具实时性。而且使用局部二值化，图形区域也会由于前景像素点过于稀疏而影响判决结果。

本文对二值化算法进行了改进，能适应复杂环境并实现快速定位。具体操作为：先用边缘检测和数学形态学的方法进行粗定位，定位出所有文本和图形区域，然后对每个区域独立使用 OSTU 法求最佳阈值，最后对每个区域进行特征提取，并通过特定判决条件判断某区域属于文本区域还是图形区域。分别对每个粗定位区域进行二值化的好处在于能适应于背景复杂的信封图像，获得良好的分割结果。另外，只对粗定位出来的区域进行窗口特征提取，减少了计算量，具有实时性。

粗定位后将所得区域划分成 $n\times n$ 像素的小窗，n 的大小由图像的大小及扫描解析度决定，在试验中一般取 10~20。这样分割的目的是为了获得用于特征提取的小窗，这些小窗必须足够小，以区分不同的区域(文本或图形)；也要足够大，以获取可以依赖的区域特征。本文使用的特征有黑白比和垂直互相关。黑白比是小窗中黑色像素点与像素点总数之间的比值，图形小窗的黑白比接近于 100%。在本文中，设定 85%或更高为图形，2%~85%为文本，2%以下为背景，85%和 2%为噪声和扫描误差。垂直互相关通过垂直像素行的快速变化判定该小窗是文本还是图形，文本区域的相关性较小而图形区域的相关性较大，一般而言，阈值取为 0.97。实验发现，阈值的选取要考虑到环境因素和图像大小差异，存在偏差 ± 0.03。

这些特征的具体定义及文字图形判别规则如下：

1) 黑白比 BW_ratio=NBP/$(n\times n)$，其中 BW_ratio 为黑白比，NBP 为 $n\times n$ 小窗中黑素像素点的个数，n 为小窗的大小；

2) 信号互相关：信号互相关计算垂直距离为 d 的两列像素点之间的关系：

$$\mathrm{Cr}(d,y)=1-\frac{2}{M}\sum_{k=0}^{M-1}p(y,k)\mathrm{XOR}(p(y+d),k) \tag{1}$$

式中：M 是小窗的高度；$p(y,k)$ 是第 y 列的第 k 个像素；d 是两列像素之间的水平距离，一般取 1 和 5；$\mathrm{Cr}(d,y)$ 取算术平均 $\overline{\mathrm{Cr}(d,y)}$；XOR 为异或运算。

判决规则如表 1 所示。

表 1 文本和图形判决规则

Table1 Rules to classify text and graphic

text	BW_ratio ⩾ 0.02	BW_ratio ⩽ 0.85	$\overline{\mathrm{Cr}(1,y)}$<0.97	$\overline{\mathrm{Cr}(5,y)}$<0.97
graphic	BW_ratio>0.85	$\overline{\mathrm{Cr}(1,y)}$ ⩾ 0.97	$\overline{\mathrm{Cr}(5,y)}$ ⩾ 0.97	

根据上述特征及判别规则，可以将各区域内小窗标志为 3 类：T 代表文本块，G 代表图形块，而 B 代表背景块。由于后续运算不需对原始图像各像素点进行，只对标记了 T,G,B 的区域块 $\Omega(i)$ 进行，减少了运算量。文献[7]用滤波的方法来进行连通性分析，并通过角点来检测文本区域，需要不停地对整幅图像进行迭代滤波，运算量较大。本文由于在粗定位阶段已确定了各区域，只要统计各类型(文本，图形，背景)的比例，并设置简单的判别规则，就可以剔除图形区域，获得所需的文本区域。

对第 i 个区域 $\Omega(i)$ 设置文本判定规则：

1) T_ratio(i) >0.5

其中 T_ratio(i) 和 G_ratio(i) 分别为第 i 个区域中文本块和图形块的比例，B_ratio(i) 为背景块的比例，T_ratio(i) + G_ratio(i) +B_ratio(i) =1。理论而言，如果 2.2 中判决条件完全成立，则对文本区域：T_ratio(i) +B_ratio(i) =1 且 G_ratio(i) =0。对图形区域：G_ratio(i) +B_ratio(i) =1 且 T_ratio(i) =0。但由于某些区域图形和文本共存或噪声的影响，实际情况并不如此。经分析发现当文本窗口比例大于 0.5 时可将该区域判定为文本区域。反之，当图形窗口比例大于 0.5 时可将该区域判定为图形区域。

2) $Width(i)/Height(i)$

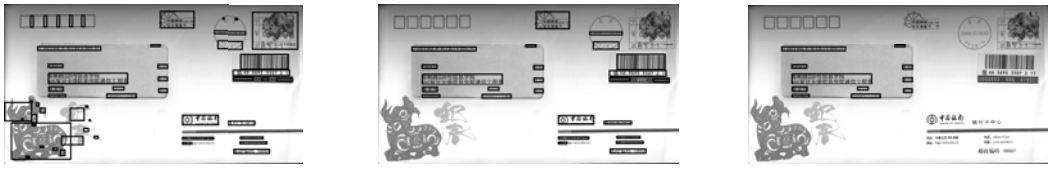
宽高比，其中 $Width(i)=R(i)-L(i)$, $Height(i)=B(i)-T(i)$。通过大量样本分析可知，地址区域的宽高比在一定范围内：$Width(i)/Height(i)>1$。

3) min *Width* 和 min *Height*

min *Width*,min *Height* 分别为区域最小宽度和最小高度。通过分析，发现地址区域都满足一定宽度和高度条件。$Width(i) \geqslant \min Width$ 或者 $Height(i) \geqslant \min Height$。

2.4 目标提取

经过以上处理可以定位信封所有文本区域。就信封分拣系统而言，目标文本主要是收件人信息，如姓名、地址等。所用到的特征包括：区域位置、水平或垂直投影以及宽高比。设定相应的判决条件就可以获得所需地址区域文本，如图 3(c)所示。



(a) coarse location　(b) text location　(c) target extraction

Fig.3 Coarse location,text location and target extraction
图 3 粗定位、文本定位和目标提取

3　结果及分析

对本文方法用 MATLAB R2007B 进行仿真，并用 VC 平台编程实现。测试样本一共为 150 幅信封图像，包括不同光照环境和背景纹理。地址块总数为 542。样本分辨力为 2 048×2 682。表 2 和表 3 为本文算法与文献[7]以及文献[4]的比较结果。

表 2 提取文本数比较
Table2 Comparison of text number extracted

	correct detected blocks	missing blocks	misdetection blocks	correct detected samples
proposed method	496	39	14	134
paper[7]	457	81	35	128
paper[4]	484	71	38	130

表 3 提取性能比较
Table3 Comparison of extraction performance

	correct detected rate/(%)	missing rate/(%)	misdetection rate/(%)	whole correct rate/(%)
proposed method	91.51	7.20	2.58	89.33
paper[7]	84.32	14.94	6.46	85.33
paper[4]	89.30	13.10	7.01	86.67

其中：正确率(correct detected rate)=正确提取文本块数(correct detected blocks)/文本块总数(total text blocks)；漏检率(missing rate)=漏提取文本块数(missing blocks)/文本块总数(total text blocks)；错检率(misdetection rate)=错误提取文本块数(misdetection blocks)/文本块总数(total text blocks)；整幅信封定位率(whole correct rate)=整幅信封正确提取数(correct detected samples)/样本总数(total samples)。

Fig.4 Algorithm applied to text extraction of business card and container code images
图 4 算法用于名片和集装箱图像文本提取

从表 2 和表 3 可以看出,本文算法在整幅图像定位率以及地址块正确率上较文献[4]和文献[7]都有所改进。文献[7]采用全局二值化及简单的窗口特征提取和连通算法,处理背景复杂及高分辨力图像时效果不佳。文献[4]利用边缘信息及启发式条件来进行文本提取，处理速度较快，但容易受噪声影响，适应能力不强。在试验中，只有当某样本正确提取文本块数量占地址区域文本块总量 80%以上时，整幅图像才能算正确提取。

图4为部分名片和集装箱图像的文本提取结果。由于图像分辨力及提取目标不一致，形态学结构算子的选取及定位判决条件应做相应调整，算法思想及流程不变。

4 结论

本文提出了一种基于数学形态学和窗口特征的快速文本提取方法，具有较高的定位准确率。综合使用边缘检测和数学形态学方法，对低对比度和较小角度倾斜情况不敏感，具有较强的鲁棒性，同时由于减弱了背景区域的影响，能减少处理时间。用多个窗口特征来区分文本和图形区域，能适应背景复杂的图像，抗干扰能力强。通过对多类样本的测试表明该算法有较强的通用性。

参考文献：

[1] Sato T,Smith M,Satoh S,et al. Indexing digital news libraries by recognition of superimposed captions[J]. ACM Multimedia Systems Special Issue on Video Libraries, 1999,7(5):385-395.
[2] 郭丽,黄元元,杨静宇. 基于连通域的版面分割研究[J]. 南京理工大学学报, 2003,27(1):16-19.
[3] 王飞,李在铭. 视频动目标标识文本检测与识别技术[J]. 信息与电子工程, 2003,1(1):25-30.
[4] SHEN Qinghua,LI Shutao,James K. Page segmentation using mathematical morphology[C]// Proc. of 2005 International Symposium on intelligent Signal Processing and Communication Systems. Hong Kong:[s.n.], 2005.
[5] Palaiahnakote S,HUANG Weihua,Tan ChewLim. Efficient video text detection using edge features[C]// 19th International Conference on Pattern Recognition. Tampa,FL:[s.n.], 2008.
[6] LI Sun,LIU Guizhong,QIAN Xueming,et al. A novel text detection and localization method based on corner response[C]// IEEE International Conference on Multimedia and Expo. New York:[s.n.], 2009.
[7] Jaakko S,Matti P. Page segmentation and classification using fast feature extraction and connectivity analysis[C]// Proc. of the 3rd International Conference on Document Analysis and Recognition. Montreal Que.,Canada:[s.n.], 1995.
[8] Ostu N. A threshold selection method from gray-level histogram[J]. IEEE Transactions on Systems,Man and Cybernetics Society, 1979,9(1):62-66.
[9] Faisal S,Daniel K. Efficient implementation of local adaptive thresholding techniques using integral image[EB/OL]. [2007-11]. http://www.dfki.uni-kl.de/~shafait/papers/Shafait-efficient-binarization-SPIE08.pdf, 2007.
[10] Derek B,Gerhard R. Adaptive thresholding using the integral image[J]. Journal of Graphics Tools, 2007,12(2):13-21.

作者简介：

谭 伟(1986-)，男，湖南省益阳市人，在读硕士研究生，研究方向为图像处理.email: tanwei8699@126.com.

方 超(1985-)，男，安徽省安庆市人，在读硕士研究生，研究方向为图像处理.

杜建洪(1960-)，男，南昌市人，教授，研究方向为图像处理、无线通信.