基于 Gabor 滤波器的数字文档图像文字提取算法

付 平,李 孟,尹洪涛
(哈尔滨工业大学,黑龙江哈尔滨 150001)

摘 要: 本文提出一种在数字文档图像中自动检测和提取文字的算法. 首先对图像在不同方向和阶数上进行 Gabor 滤波,得到反映文档图像布局的滤波图像,然后在得到的滤波图像中直接提取候选文字区域,再利用几何特性和高频分量特性筛选准则从中剔除非文字区域. 最后选取了不同类型、不同语言和不同字体的文档图像进行实验,实验结果表明本算法对各种文档图像均能给出满意的结果.

关键词: 文字提取; Gabor 滤波器; 数字文档图像
中图分类号: TP309 **文献标识码**: A **文章编号**: 0372-2112 (2006) 12A-2387-05

Gabor Filter Based Text Extraction from Digital Document Images

FU Ping, LI Meng, YIN Hong-tao
(Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: This paper presents an algorithm that can automatically detect and extract text in digital document images. Firstly, we process and fuse Gabor filtered images at different orientations and scales and obtain an image that reflects the layout of the document image. Then, potential text regions are directly extracted from the resulting image. Finally, two criteria based on the geometrical property and high frequency content are adopted to kick-out those non-text regions. The experiments are performed on some representative images with different styles and with texts in different languages and fonts. Experimental results show that the algorithm works well on document images from a wide variety of source.

Key words: text extraction; Gabor filter; digital document images

1 引言

随着数字技术快速发展,数字图像日益增多. 对图像进行人工标注和检索既繁琐又浪费时间,这就需要对数字图像进行自动标注和检索,而文档分割在数字图像自动标注和检索中起着重要的作用.

已有文档分割方法大致可分为三类:从上至下的方法,从下至上的方法和基于纹理的分割方法. 经典的从上至下方法是基于周期滤波(RLS)算法[1,2]和轮廓投影[3]的. 从下至上的方法是由像素开始将小元素聚类成较大的接连元素,直到图像中的所有区域都被区分开[4]. 而基于纹理的方法利用图像中的文字具有明显纹理特征这一属性来提取文字[5]. 学者还提出一些新方法:文献[6]利用神经网络以及连通分量分析,在输入图像中自动提取文字,半色调和线条区域;利用多尺度特征向量和模糊局部抽取信息进行对格式独立的文档页面分割[7];对二值图像进行连通成份分析,进而用 Gabor 滤波器组分割文字区域和非文字区域[8];文献[9]可自动检测彩色图像中的文本,该方法对图像强度信息进行 Gabor 多通道滤波,对图像颜色信息进行图形理论聚类;文献[10]给出了 OCR 算法的若干评价准则.

这些文档分割算法都具有局限性. 从上至下的方法只适用于矩形块,不能处理倾斜的文字. 从下至上的方法会受到字体大小,扫描分辨率,分割线和字符间距的影响. 且这两种方法的输入图像必须为二值图像. 尽管文献[5]中的纹理分割方法得到结果较好,但没有考虑文字与图形交迭和混合的情况. 虽然在各个不同领域对文档分析和理解的研究很多,但在文档图像的 Gabor 变换域直接提取文字区域的方法很少.

本文提出一种基于 Gabor 滤波器的文字提取算法,可在各种文档图像中自动提取文字. 新算法利用如下两个特性:文字区域的中频和高频分量较丰富;文字区域通常具有类似矩形的边界. 第一个特性使我们可以设计适当的 Gabor 滤波器来捕捉文字区域的有效信息,进而提取出候选文字区域. 第二个特性可以用来在这些候选区域中找出真正的文字区域.

2 Gabor 滤波器

使用 Gabor 滤波器组进行图像分割在过去几十年里受到了广泛的关注[6~10],利用 Gabor 滤波器组可以得到大量的纹理特征,而且 Gabor 滤波器可以很好的模拟人类视觉系

收稿日期:2006-08-22;修回日期:2006-11-22

统[11,12].

Gabor 函数 $g(x,y)$ 及其傅立叶变换 $G(u,v)$ 为:

$$g(x,y)=\left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi\mathrm{j}Wx\right] \quad (1)$$

$$G(u,v)=\exp\left\{-\frac{1}{2}\left[\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right]\right\} \quad (2)$$

Gabor 滤波器的非正交性决定了滤波图像中会含有一些冗余信息,下面的改进可以消除这种冗余[13]. 令 U_l 和 U_h 分别表示最低和最高中心频率, K 表示多分辨分析的方向数, S 表示多分辨分析的阶数. 消除这些冗余,就要保证各个滤波器响应在频域中相互接触又互不重叠,如图 1 所示. 于是可得到计算 σ_u 和 σ_v(σ_x 和 σ_y)的公式:

$$a=(U_h/U_l)^{\frac{1}{S-1}},\sigma_u=\frac{(a-1)U_h}{(a+1)\sqrt{2\ln 2}},$$

$$\sigma_v=\tan\left(\frac{\pi}{2k}\right)\left[U_h-(2\ln 2)\frac{\sigma_u^2}{U_h}\right]\left[2\ln 2-\frac{(2\ln 2)^2\sigma_u^2}{U_h^2}\right]^{-\frac{1}{2}} \quad (4)$$

令式(1),(2)中 $W=U_h$, $m=0,1,\cdots,S-1$.

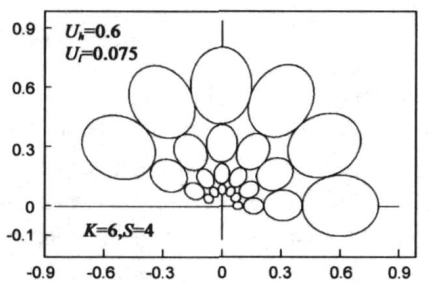

图 1 Gabor滤波器响应示意图

3 文字提取算法

3.1 图像滤波

首先对文字图像(图 2(a))和非文字图像(图 2(b))使用图 1 中的 Gabor 滤波器进行滤波. 图 2(c)和图 2(d)分别给出了图 2(a)和图 2(b)滤波后的系数谱,可以看到文字图像的中、高频系数明显比非文字图像大. 这是因为相对于文字来说,自然图像比较平滑.

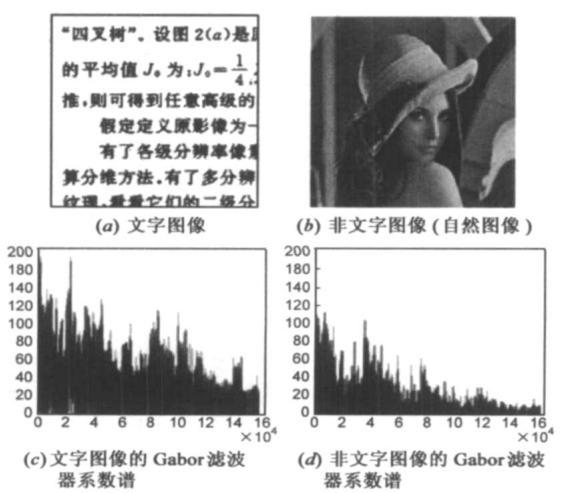

图 2 两类图像及其 Gabor滤波器系数谱

为了更清楚的表现文字区域和非文字区域的不同特性,在图 3(a),(b)分别给出了一幅文档图像和它的 Gabor 滤波图像(使用的 Gabor 滤波器如图 1). 由滤波图像可以看到:文字区域含有较丰富的中频和高频信息;在后两阶滤波图像中,文字区域具有类似矩形的边界.

由图 3(b)可以看出,文字区域和非文字区域在滤波图像前两阶难以区分,本算法选取参数为 $U_h=0.6$, $U_l=0.3$, $K=6$, $S=2$ 的 Gabor 滤波器组,即图 3(b)中下面两排滤波图像所使用的滤波器. 令 $O_{s,k}$, $s=0,1$, $k=0,\cdots,K-1$ 表示这 12 个 Gabor 滤波图像.

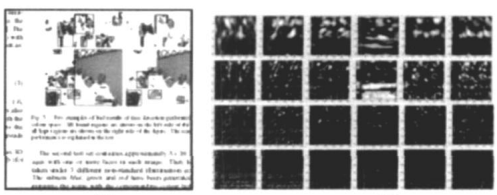

图 3 文档图像及其滤波图像

3.2 文字提取算法

新算法的主要思想是首先处理 Gabor 滤波图像,得到一幅反映原始图像高频分量的图像;然后在该图像中找出候选文字区域;最后使用文字区域的两个特性(类似矩形的边界和高频分量的丰富性)去除非文字区域. 新算法与传统算法相比主要有以下两点创新:

1. 直接在 Gabor 滤波图像中提取文字区域,而传统方法是通过在空域中分析原始图像的布局来确定文字区域的.

2. 引入了一个新特征:文字区域在滤波图像中具有类似矩形的边界.

3.2.1 滤波图像预处理

滤波图像预处理的目的是得到一幅反映文档图像布局的图像. 为了去除滤波图像中的噪声从而更好的提取候选文字区域,首先将 12 个 Gabor 滤波图像 $O_{s,k}$, $s=0,1$, $k=0,\cdots,K-1$ 进行二值化. 由于不同的文档图像滤波系数大小不尽相同,为了得到更能反映当前所处理文档图像的二值化图像,我

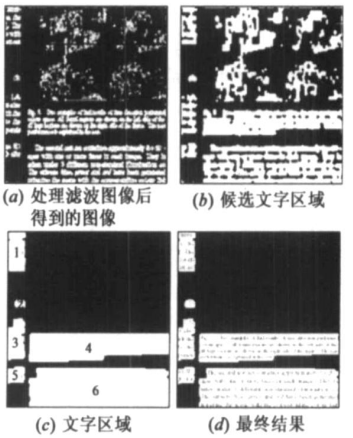

图 4 图 3(a)所示图像的文字提取过程

图 5 确定文字区域的算法示意图

们选取二值化的阈值为 $T_{s,k} = Q_{s,k} \times A_{s,k}$,其中 $A_{s,k}$为滤波图像 $O_{s,k}$中所有像素的均值,$Q_{s,k}$为根据经验选取的加权值. 在本文的实验中选取 $Q_{0,k} = 1.4$, $Q_{1,k} = 1.7$, $k = 0, \cdots, K-1$. 为了融合各方向的信息,将得到的二值图像 $O'_{0,k}$, $k = 0, \cdots, K-1$ 和 $O'_{1,k}$, $k = 0, \cdots, K-1$ 进行平均,得到两幅图像. 然后以 1∶9 的比例对这两幅图像进行加权平均,这样就得到了融合各方向和阶数信息的图像,它将用于后面区域分割的操作. 以图 3(*a*)为输入图像,图 4(*a*)为上面操作后得到的图像.

3.2.2　分割候选文字区域

滤波图像预处理后的图像可能含有非文字区域和噪声,可通过二值化和一些修补操作去除. 二值化操作可去除一些非文字区域,而且可以将文字区域与背景区域分离. 因为 Gabor 滤波器具有很好的局部分辨能力,所以在本应属于同一区域的字符和词组之间可能会存在间隙,可以通过修补操作将这些小区域连接起来. 对二值化和修补操作后的图像使用八邻域边界追踪算法[14]得到候选的文字区域. 图 4(*b*)给出了由图 4(*a*)中提取出的 44 个候选文字区域.

3.3.3　确定文字区域

为了将非文字区域从候选文字区域中剔出,使用如下两个度量准则:

1. 标准率(Standard Rate,简称 SR)准则:对于候选文字区域,可以找到面积最小的矩形将其包围. 定义某个区域的标准率 SR 为该区域中像素数与它对应的面积最小矩形中像素数的比值,它表示的是该候选区域与矩形的相似度. 如果一个区域的标准率大于阈值 A,就认为该区域为文字区域. 否则,利用下面的高频分量参数进行测试.

2. 高频分量(High Frequency Content,简称 HFC)准则:候选区域的 HFC 为该区域中像素值为 255 的像素数与该区域中像素数的比值. 它反映了该区域中含有高频分量的多少. 文字区域的高频分量要比非文字区域的高频分量丰富,所以 HFC 可作为候选区域是否是文字区域的判据. 如果某区域的 HFC 大于 B,就认为该区域为文字区域.

文字区域的确定过程见图 5,其中 A 和 B 需根据当前处理的文档图像的字体、文字大小等情况进行选取.

对图 4(*b*)中的所有候选区域使用上面的算法(本文取 $A = 0.8$, $B = 0.25$),得到图 4(*c*)中的六个文字区域. 它们的 SR 分别为 0.803,0.810,0.631,0.852,0.852 和 0.944. 区域 1,2,4,5 和 6 是由第一个准则确定的,而区域 3 是由第二个准则确定的,其 HFC 为 0.342. 图 4(*d*)给出了最终的文字提取结果.

4　实验结果

实验选取了六幅具有代表性的文档图片. 它们大体可以分为四类:包含文字和自然图像,且二者不重叠(图 6(*a*));包含文字和人工生成图片(图 6(*b*)和图 6(*c*));包含文字和人工生成图表(图 6(*d*));自然图像中含有文字(图 6(*e*)和图(*f*)). 实验图像为大小为 256×256,灰度级为 256 的文档图像. 实验图像和提取结果如图 6.

图 6(*a*)为含有汉字和自然图像的图像,由实验结果可以看出本算法可有效的处理汉字文档. 图 6(*b*)为含有文字的卡通图片,本算法对这类图像的效果很好. 如果文档图像是由扫描或者照相机得到的,那么它有可能会倾斜. 图 6(*c*)给出了

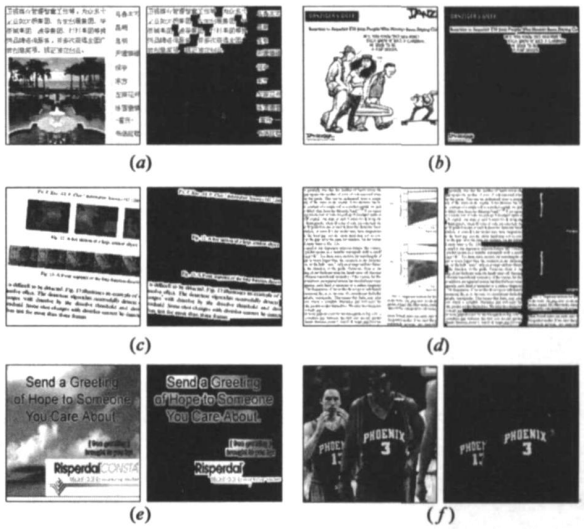

图 6　实验结果

倾斜的文档图像和它的提取结果,可以看到本算法受文档图像倾斜的干扰很小. 图 6(*d*)中的文字被成功提取,而且页眉、页码和图标也被成功提取. 但是与图表距离很小的数字被漏检,这是因为这些数字被错误的分到了非文字区域. 图 6(*e*)为覆盖有文字的自然图像,这类图像在商业广告中经常出现,新算法可以处理这类图像. 需要说明的是,"CONSTA"并没有被提取出来,这是因为"CONSTA"的灰度值与周围背景的灰度值相差太小. 这类问题可以通过增加图像的对比度来解决. 最

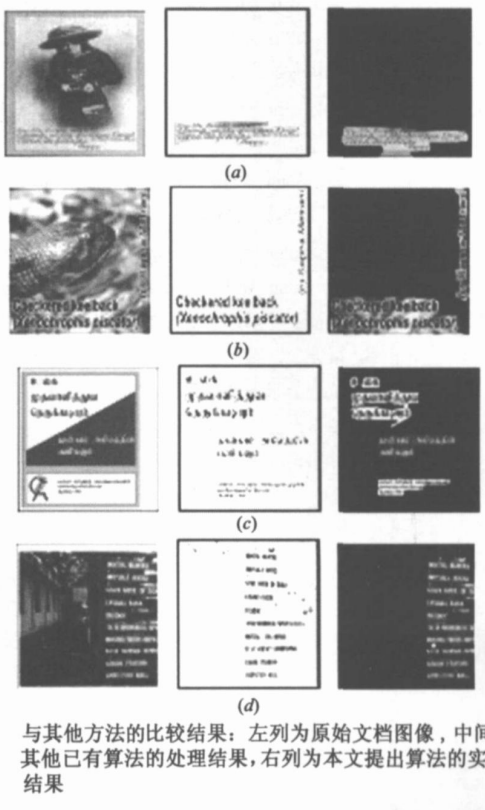

图 7　与其他方法的比较结果:左列为原始文档图像,中间为其他已有算法的处理结果,右列为本文提出算法的实验结果

后一幅图像中篮球运动员球衣上的队名和号码被提取出来.由实验结果可以看出,本文提出算法可对不同语言,不同风格和不同字体的文档图像进行有效处理.

图 7 给出了本算法与其它已有算法的实验结果比较.图 7(a)为本算法与文献[8]的处理结果比较,可看出本文算法与文献[8]中算法可达到相同的结果.图 7(b),(c),(d)为与文献[9]算法的比较,由(c),(d)结果可以看到,本算法受噪声干扰明显降低.

5 结论

基于文字区域含有丰富的中\高频成分和文字区域在 Gabor 滤波图像中具有类似矩形的边界两点特性,本文提出了一种新的文字提取算法.我们首先设计了能够捕获文字区域有效信息的 Gabor 滤波器,进而提出了在候选文字区域中确定文字区域的准则.本文采用五幅不同语言,不同字体,不同风格的文档图像进行实验,实验结果证明了新算法的有效性.

参考文献:

[1] F M Wahl, K Y Wong, R G Casey. Block segmentation and text extraction in mixed text/image document [J]. Computer Graphics and Image Processing, 1982, 20(4): 375 - 390.

[2] K Y Wong, R G Casey, F M Wahl. Document analysis system [J]. IBM Journal Res. Dev, 1982, 26(6): 647 - 656.

[3] D Wang, S N Srihari. Classification of newspaper image blocks using texture analysis [J]. Computer Graphics and Image Processing, 1989, 47(3): 327 - 352.

[4] L O'Gorman. The document spectrum for page layout analysis [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1993, 15(11): 1162 - 1173.

[5] A KJain, S Bhattacharjee. Text segmentation using Gabor filters for automatic document processing[J]. Machine Vision and Applications, 1992, 5(3): 169 - 184.

[6] A K Jain, Y Zhong. Page segmentation using texture analysis [J]. PR, 1996, 29(5): 743 - 770.

[7] K Etemad, D Doermann, R Chellappa. Multiscale document page segmentation using soft decision integration [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(1): 92 - 96.

[8] S S Raju, P B Pati, A G Ramakrishnan. Gabor filter based block energy analysis for text extraction from digital document images [A]. Proceedings of the First International Workshop on Document Image Analysis for Libraries [C]. Palo Alto, California, USA: IEEE, 2004. 233 - 243.

[9] S S Raju, P B Pati, A G Ramakrishnan. Text localization and extraction from complex color images [J]. International Symposium on Visual Computing 2005, LNCS-3804: 486 - 493.

[10] S Mao, T Kanungo. Emprirical performance evaluation methodology and its application to page segmentation algorithms [J]. PAMI, 2001, 23(3): 242 - 256.

[11] M Clark, A C Bovik, W S Geisler. Texture segmentation using Gabor modulation/demodulation [J]. Pattern Recognition Letters, 1987, 6(4): 261 - 267.

[12] J G Daugman. Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression [J]. IEEE Trans. Acoust. Speech Signal Process, 1988, ASSP-36: 1169 - 1179.

[13] I Fogel, D Sagi. Gabor filters as texture discriminator [J]. Biol Cybernet, 1989, 61(2): 103 - 113.

[14] M Sonka, V Hlavac, R Boyle. 图像处理,分析与机器视觉 [M]. 北京:人民邮电出版社, 2002. 142 - 148.

作者简介:

付 平 男,1965 年生于黑龙江哈尔滨,博士学位,现为哈尔滨工业大学电气工程及自动化学院自动化测试与控制系研究员,博士生导师,中国电子学会及中国计量测试学会高级会员.主要研究方向为计算机自动测试与控制、图像处理. E-mail: fuping@hit.edu.cn

李 孟 男,1982 年生于黑龙江大兴安岭,2005 年获得哈尔滨工业大学硕士学位,现为哈尔滨工业大学自动化测试与控制研究所博士研究生.研究方向为图像处理、模式识别、三维纹理. E-mail: lemon820910@126.com