doi:10.3969/j.issn.1673-4785.2010.02.015

小波的文本图像区分及其在文献信息数字化中的应用

陈 杰[1],孙忠贵[2],周书锋[2]
(1.聊城大学 图书馆,山东 聊城 252059; 2.聊城大学 数学科学学院,山东 聊城 252059)

摘 要:目前,OCR技术对文本图像区域自动区分的效果还不够精确,进而影响了OCR技术在文献信息数字化过程中的工作效率.针对这一局限,提出了一种基于小波的文本图像区分方法.方法首先对扫描区域进行小波分解,然后使用分解系数构建分解能量,最后依据分解能量大小对文本图像进行自动区分.结果表明,该方法对文本图像的区分效果较好,减少了在使用OCR技术进行文献信息数字化时的人为干预,有利于提高文献信息数字化过程的自动化水平.最后通过实验仿真验证了该方法的有效性.

关键词:数字化文献;OCR;小波;文本图像
中图分类号:TP18; TN911.72 **文献标识码**:A **文章编号**:1673-4785(2010)02-0185-04

Applying image classification using wavelets to digitization of document information

CHEN Jie[1], SUN Zhong-gui[2], ZHOU Shu-feng[2]
(1. Library of Liaocheng University, Liaocheng 252059, China; 2. College of Mathematics Science, Liaocheng University, Liaocheng 252059, China)

Abstract: The accuracy of optical character recognition (OCR) technology in distinguishing between text areas and image areas has remained relatively low. Unfortunately this reduces the efficiency of OCR in digitization of document information. After analyzing the main steps of OCR applied to a digital library, the authors evolved an image classification algorithm based on wavelets. Decomposing the scanning area with wavelet transform was the first step in the algorithm. The energy value of the area could then be derived from wavelet coefficients. The task of distinguishing between text and images was accomplished by analyzing their energy values. The algorithm proved fast and automatic, characteristics increasing the efficiency of the digitization of document information. It was clear that the simulation verified the new algorithm's feasibility.

Keywords: digitalizing document; OCR; wavelet; text image

文献信息数字化是信息资源建设的重要内容之一,OCR技术是目前进行文献数字化的主要技术手段[1-3].由于OCR(optical character recognition, OCR)技术不能很好地对文献的文本区域和图像区域进行自动区分,在使用OCR技术进行文献数字化时需要工作人员手工标记出文本区域和图像区域,以至降低了文献数字化的工作效率.针对这一问题,提出了一种基于小波的文本图像区分方法,该方法有助于实现文本区域和图像区域的自动区分,将其与OCR技术相结合,有利于提高文献数字化的自动化水平和工作效率.最后通过实验仿真验证了该方法的有效性.

收稿日期:2009-12-05.
基金项目:聊城大学青年教师科研基金资助项目(X0810029).
通信作者:孙忠贵. E-mail:altlp@vip.sina.com.

1 OCR技术使用步骤及局限

光学字符识别,也可简单地称为文字识别,实际上是计算机认字,是一种文字自动输入方法.在文献数字化中,其录入速度远远高于传统的手工输入,成为目前进行文献数字化的主要技术手段.虽然OCR软件多种多样,但在进行文献数字化时,它们的使用步骤基本相同[4-5]:

1)文献扫描:使用扫描仪对要进行数字化的文献进行扫描生成相应的图像;

2)图像处理:为了得到较好的识别效果,对1)得到的图像进行处理,主要包括倾斜校正、图像杂点及图像块的擦除、文献分析等;

3)文字识别:使用OCR软件对处理后的图像进行文字识别;

4)文稿校对:对3)中识别错误的文字进行校正.

由使用OCR技术进行文献数字化的基本步骤可以看出:2)和4)是人为干预部分,这2部分的存在破坏了整个过程的自动化特性,降低了文献数字化的处理速度,这2部分也成了使用OCR技术进行文献数字化的效率瓶颈.由于现有的OCR软件只能识别文字不能识别图像,故在2)中需要对图像块进行处理.为了提高识别精度,在使用OCR软件进行文献数字化时,对图像块的处理手段主要有2个:一是通过工作人员的人眼视觉找到图像块,然后进行擦除,其缺陷是速度慢、且工作人员容易疲劳;另一个手段是不擦除图像块,将其和文本一样进行处理,这样做虽然避免了手工擦除、提高了处理速度,但图像块的存在会对文本识别造成干扰,且把处理文本的手段用来对图像块处理有时还会得到意想不到的结果.比如,在文献数字化时,往往要对文本部分进行二值化[6],由于没有对图像块和文本区域进行区分,结果图像块也被二值化,本来颜色、灰度丰富的图像被简单地用黑、白2种颜色来表达,造成大量有用细节信息的丢失.为了对图像和文本区域进行区分,一些OCR软件虽然已具有了版面自动分析功能,但分析效果还不够精确,有待提高.

上面所说的图像块的二值化和图像块的缺失现象在一些数字化文献中时有发生,这造成一些数字化文献不能很好地忠于原始文献.出于维护知识产权等原因,现在学术期刊越来越多地倾向印刷作者的肖像照片,在对这部分学术期刊进行数字化处理时,肖像区域与文本区域的区分尤为重要.由于肖像与一般图像相比,具有纹理更简单、灰度变化更缓慢的特点,故肖像与文本区分的可操作性更强、更有可能达到好的区分效果.算法主要针对作者肖像和文本区域进行区分.

2 基于小波的文本图像区分

2.1 小波变换原理

Mallat提出了小波变换的分解与重构的快速算法[7].对二维信号(如图像)$f(x,y)$,其分解公式为

$$C_{n,m}^{j}=\frac{1}{2}\sum_{k,l\in Z}\bar{h}_{k-2n}\bar{h}_{l-2m}C_{k,l}^{j-1},$$

$$d_{n,m}^{j1}=\frac{1}{2}\sum_{k,l\in Z}\bar{h}_{k-2n}\bar{g}_{l-2m}C_{k,j}^{j-1},$$

$$d_{n,m}^{j2}=\frac{1}{2}\sum_{k,l\in Z}\bar{g}_{k-2n}\bar{h}_{l-2m}C_{k,l}^{j-1},$$

$$d_{n,m}^{j3}=\frac{1}{2}\sum_{k,l\in Z}\bar{g}_{k-2n}\bar{g}_{1-2m}C_{k,l}^{j-1}$$

式中:$C_{k,l}^{j-1}$代表图像的第$j-1$阶近似分量,$C_{n,m}^{j}$代表图像的第j阶近似分量($\mathrm{LL}^{(j)}$),$d_{n,m}^{j1}$代表图像的第j阶垂直方向细节分量($\mathrm{LH}^{(j)}$),$d_{n,m}^{j2}$代表图像的第j阶水平方向的细节分量($\mathrm{HL}^{(j)}$),$d_{n,m}^{j3}$代表图像第j阶对角方向的细节分量($\mathrm{HH}^{(j)}$).

图像的重构公式为

$$C_{n,m}^{j-1}=2\sum_{k,l\in Z}(C_{k,l}^{j}h_{n-2k}h_{m-2l}+d_{k,l}^{j1}h_{n-2k}g_{m-2l}+d_{k,l}^{j2}g_{n-2k}h_{m-2l}+d_{k,l}^{j3}g_{n-2k}g_{m-2l}).$$

式中:h、g分别为低通滤波器和高通滤波器系数.

小波变换是空间(时间)和频率的局部变换,通过伸缩和平移等运算功能可对信号进行多尺度的细化分析,最终达到高频处时间细分、低频处频率细分、从而可以聚焦到信号的任意细节,因而被称为“数学显微镜”.小波变换已被广泛应用于文本处理、信号分析等领域.文献[7]指出通过分析文本和图像的小波系数可以达到区分文本和图像的目的.本文在文献[8]的基础上,通过定义小波分解不同分量的能量比,得到了一个文本区域和肖像区域的区分方法.

2.2 基于小波的文本图像区分算法

用小波对图像进行分解,所得j阶分量的能量

是指该分量系数的平方和[9],其具体定义如下:

$$\mathrm{ELL}^{(j)} = \sum_{n,m} (C^{j}_{n,m})^2,$$

$$\mathrm{ELH}^{(j)} = \sum_{n,m} (d^{j1}_{n,m})^2,$$

$$\mathrm{EHL}^{(j)} = \sum_{n,m} (d^{j2}_{n,m})^2,$$

$$\mathrm{EHH}^{(j)} = \sum_{n,m} (d^{j3}_{n,m})^2.$$

为了对图像变化的缓慢程度进行严格数学度量,下面给出图像的第 j 阶细节分量与近似分量能量比的定义:

$$\mathrm{EC}^{(j)} = \frac{\mathrm{ELH}^{(j)} + \mathrm{EHL}^{(j)} + \mathrm{EHH}^{(j)}}{\mathrm{ELL}^{j}}.$$

显然,若图像的灰度变化缓慢,则其细节分量的能量小,细节分量与近似分量的能量比也小;若图像的灰度变化较快,则细节分量的能量大,细节分量与近似分量的能量比也大. 由于文字的笔划特征,在整个文本区域内,灰度不停地在最大值(白色)和最小值(黑色)之间快速跳变;而对于人脸肖像来说,由于其灰度变化比较缓慢,往往看上去比较柔和. 因此肖像区域细节分量和近似分量的能量比远远小于文本区域细节分量和近似分量的能量比. 这就保证了把细节分量和近似分量的能量比作为区分文本和肖像的分类特征的合理性.

2.3　仿真实验

对图 1 的 4 个采样区域进行区分处理,在图像分解与重构时为了能够精确地重构原图像,滤波器系数需要满足对称性要求,为此选用 Bior3-7 小波进行实验,这里只进行一层小波分解.

采用 4 幅图像进行实验,其中图 1(b)是随机抽取自 Manchester 大学人脸库中编号为 1a034 的一幅肖像[10]. 采用 Matlab 仿真编程[11],得到图 1 中 4 个采样的细节分量与近似分量的能量比依次为:0.002 0、0.000 1、0.012 3 和 0.006 3. 在这 4 个样本中,文本区域的最小能量比是肖像区域最大能量比的 3 倍还多. 为了进一步说明用细节分量和近似分量的能量比能区分肖像和文本区域,进行文献数字化的可行性,随机选取 5 种不同的学术期刊,对每种期刊又随机采样 20 幅肖像和 20 幅文本区域构成图像库. 计算图像库中全部 200 幅图像的细节分量和近似分量的能量比,所得结果如图 2 所示.

(a)肖像1　(b)肖像2　(c)中文字符　(d)英文字符

图 1　4 个不同的图像区域采样

Fig. 1　Four image region samples

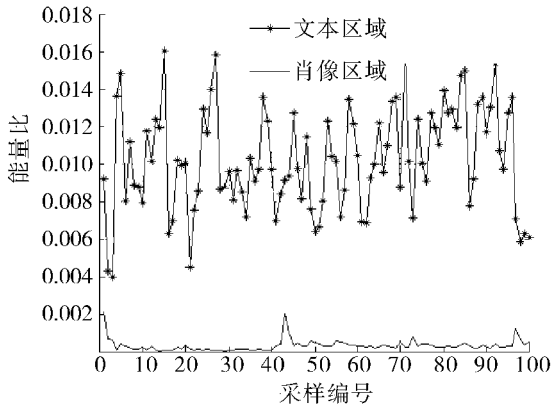


图 2　基于小波的文本图像区分仿真

Fig. 2　Simulation of image classification based on wavelet

由图 2 可知,在随机选取的 100 幅文本区域中,只有 4 幅的能量比小于 0.006,且最小值为0.004 3;在随机选取的 100 幅肖像区域中,只有 3 幅的能量比大于 0.001,且最大值为 0.002 1. 通过上面分析可知,以能量比作为分类指标,这个仿真图像库是线性可分的. 取能量比为 0.003,能量比大于 0.003 的样本划分为文本区域,能量比小于 0.003 的样本划分为肖像区域,便可对仿真图像库进行合理分类.

本文仿真图像库采用的样本均为灰度图像,对彩色图像可对其颜色空间进行处理. 基于小波文本图像区分方法主要利用了人脸肖像灰度变化缓慢的特点,对于其他灰度变化缓慢的图像这一算法同样有效.

3 结束语

小波变换与经典的傅里叶变换相比,在时域和频域均据有较强的局部化功能,在图像分割中得到较为广泛的应用[12]. 实验表明,基于小波的文本图像区分方法,能够很好地区分文献的文本区域和肖像区域. 算法只需要计算小波分量的能量比,不需要过多人为干预,这在一定程度上解决了 OCR 软件不能自动区分文本区域和图像区域的局限,把该方法与 OCR 技术相结合,有助于快速标记文献的图像区域,从而提高文献数字化的效率. 需要说明的是,该算法只对区分文本和灰度变化缓慢的图像有效,对于一些灰度变化较快的图像,算法并不适用,这也是目前图像处理领域的难点之一.

参考文献:

[1]孙洪睿. 高校数据信息平台的研究与设计[J]. 应用科技,2009(7):41-46.
SUN Hongrui. The research and design platform of college information[J]. Applied Science and Technology, 2009(7):41-46.

[2]孙 萍,苏东出. 基于 OCR 的电子图书目录自动生成算法的实现[J]. 现代情报,2004(9):151-155.
SUN Ping, SU Dongchu. An algorithm based on OCR for e-book directory automatically generated[J]. Modern Information, 2004(9):151-155.

[3]梁 红. 高校数字图书馆信息资源建设探析[J]. 图书馆工作与研究,2005(4):55-57.
LIANG Hong. The analysis about the construction of university digital library information resources[J]. Library Work and Study, 2005(4): 55-57.

[4]上海中晶科技有限公司. OCR 软件使用经验谈[J]. 电子出版,2002(6):10.
Shanghai Microtek Technology Co, Ltd. The applying experience of OCR software[J]. Electronic Publishing, 2002(6):10.

[5]张成昱,赵 仪,邹 荣,等. 中文电子图书系统开发和应用研究[J]. 大学图书馆学报,2002(4):19-23.
ZHANG Chenyu, ZAO Yi, ZHOU Rong, et al. Study on the development and application of a Chinese e-book system [J]. Journal of Academic Libraries, 2002(4):19-23.

[6]苏东出. 一种改进的黑白二值化方法—谈文献扫描图像的数字化处理[J]. 情报杂志,2003(5):69-70.
SU Dongchu. An improved binarization method: introducing the digitization of documents scanning[J]. Journal of Information, 2003(5): 69-70.

[7]MALLAT S. 信号处理的小波导引[M]. 北京:机械工业出版社,2002:193-199.

[8]SCHETTINI R, BRAMBILLA C, CIOCCAA G, VALSASNA A, De PONTI M. A hierarchical classification strategy for digital documents[J]. Pattern Recognition, 2002(35): 1759-1769.

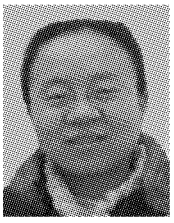
[9]唐远炎,王 玲. 小波分析与文本文字识别[M]. 北京:科学出版社,2004:269-277.

[10]The University of Manchester. Face image library [EB/OL]. [2005-9-1]. http://images. ee. umist. ac. uk/danny/face. tar. gz.

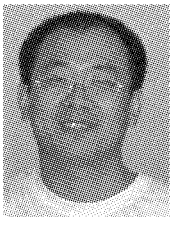
[11]GONZALEZ R C, WOODS R E, EDDINS S L. Digital image processing using Matlab [M]. Beijng: Publishing House of Electronics Industry, 2004:181-186.

[12]张晓威,郑雄波,郭 健. 小波域内背景图像的文本信息提取研究[J]. 哈尔滨工程大学学报,2008(3):314-318.
ZHANG Xiaowei, ZHENG Xiongbo, GUO Jian. Extracting text information from a background image using wavelet domains[J]. Journal of Harbin Engineering University, 2008(3):314-318.

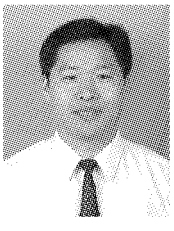
作者简介:



陈 杰,女,1974 年生,主要研究方向为图书馆资源建设、信息检索等. 发表学术论文 10 篇.



孙忠贵,男,1971 年生,副教授,主要研究方向为信息处理、机器学习等. 发表学术论文 15 篇.



周书锋,男,1973 年生,讲师,主要研究方向为机器学习、计算机应用等. 发表学术论文 13 篇.