一种有效的文本图像二值化方法

An Effedtive Binarization Algorithm For Document Image

(郑州信息工程大学) 庄　军　李弼程　陈　刚

Zhuang,Jun　Li,Bicheng　Chen,Gang

摘要：针对一般文本图像二值化方法--全局阈值法和局部阈值法所存在的不足，提出了一种整体与局部相结合的二值化方法，该方法根据文本图像的特点，自适应调整局部二值化的窗口大小，能有效避免断笔，伪影等现象，同时保持较快的处理速度。

关键字：文本图像；二值化；自适应；窗口大小

中图分类号:TP391.41　　**文献标识码**:A

文章编号:1008-0570(2005)08-0056-02

Abstract: This paper proposes a new method for binarization with global and local information combined, aiming at the weakness of the single global and local threshold. Because of the characteristics of document images, the new method extracts the window size for local binarization adaptively, which not only avoids disconnection and ghost but also remains a fast speed.

Keywords: document image; binarization; adaptive; window size

技术创新

1 概述

根据其运算的范围不同，文本图像的二值化方法可分为全局阈值方法和局部阈值方法。全局阈值法根据文本图像的直方图或灰度空间分布确定一个阈值，以此实现灰度文本图像到二值图像的转化。典型的全局阈值方法包括 Ostu 方法、迭代算法等。全局阈值法算法简单,对于目标和背景明显分离、直方图分布呈双峰的图像效果良好，但其对输入图像有噪声或不均匀光照等情况抵抗能力差，应用受到极大限制。局部阈值法通过定义考察点的邻域，比较考察点与其邻域的灰度值来确定当前考察点的阈值。非均匀光照条件等情况虽然影响整体图像的灰度分布却不影响局部的图像性质，使得局部阈值法较全局阈值法有更广泛的应用。常用的局部阈值法有 Niblack 算法、Bernsen 算法等。但局部阈值算法对文本图像进行二值化处理时，会出现伪影等问题。

本文根据文本图像特点，提出了一种整体与局部相结合的二值化方法。该方法自适应选取局部二值化时的窗口宽度，能有效消除一般二值化方法容易产生的断笔和伪影现象，同时吸收全局二值化方法的优点，保持较快的运算速度。

庄军：硕士研究生

基金项目：河南省教育厅基金（编号：sp200303099）资助项目

文章共分五个部分：第一部分概述；第二部分简单介绍两种常用的二值化方法：迭代算法和 Bernsen 算法；第三部分在前述算法的基础上提出了一种新的综合二值化方法；第四部分介绍实验结果；第五部分是总结。

2 全局迭代算法和 Bernsen 算法简介

2.1 全局迭代算法

全局迭代算法是一种全局阈值二值化方法。该方法首先选取一初始阈值，其值取为文本图像的最大灰度值与最小灰度值的均值，根据该阈值将图像二值化为目标与背景，然后以目标和背景的平均期望值作为新的阈值，对图像重新二值化，如此不断迭代。当阈值不再变化时，停止迭代。一般迭代几次后即可达到稳定状态。迭代算法具体过程如下：

首先计算初始阈值 $g_0=\frac{(g_{max}+g_{min})}{2}$，其中，$g_{max}$、$g_{min}$ 分别是文本图像的灰度最大值和灰度最小值。根据 g_0，把图像中的像素点分成大于 g_0 和小于 g_0 的两部分，分别 g_1 求它们的期望值，取为它们期望的平均值。如此反复迭代，当$|g_n-g_{n-1}|$足够小时，取 $T=g_n$，T 即为全局二值化的阈值。

$$T=\lim_{n\to\infty}\frac{m_f(g_n)+m_b(g_n)}{2} \tag{1}$$

其中

$$m_f(g_n)=\frac{\sum_{g=0}^{T_n} gp(g)}{\sum_{g=0}^{T_n} p(g)},\quad m_b(g_n)=\frac{\sum_{g=T_n+1}^{G} gp(g)}{\sum_{g=T_n+1}^{G} p(g)} \tag{2}$$

$m_f(g_n)$为目标期望值，$m_b(g_n)$为背景期望值。

2.2 Bernsen 算法

Bernsen 算法是一种经典的局部二值化方法。考虑以坐标(x,y)为中心的(2d+1)×(2d+1)模板，g(x,y)表示(x,y)处的灰度值，b(x,y)为 g(x,y)的二值化结果，则 Bernsen 算法可描述为：

(1) 计算每一点的阈值

$$T_1(x,y)=0.5\times\left(\max_{\substack{-d<k<d\\-d<l<d}} g(x+k,y+l)+\min_{\substack{-d<k<d\\-d<l<d}} g(x+k,y+l)\right) \tag{3}$$

(2) 逐点二值化　$b(x,y)=\begin{cases}0 & g(x,y)<T_1(x,y)\\ 1 & g(x,y)>T_1(x,y)\end{cases}$

Bernsen算法的阈值由考察点对应邻域的灰度值确定,算法不存在预定阈值,适应性较全局阈值法广。

3 整体与局部相结合的二值化方法

Bernsen算法以局部窗口内最大、最小值的均值作为对应考察点的阈值,当窗口内无目标点时,个别噪声点将引起阈值的突变。另外,背景灰度的非均匀性也将影响局部阈值的变化。当考察窗内均为目标点时,局部阈值被拉伸,这样势必使得宏观上本应同类的部分目标象素被强行二值化为背景,或者出现相反的情况,从而出现笔划断裂及伪影现象。所以窗口大小的选取在 Bernsen 算法中很关键。w 过小,容易造成断笔;w 过大,又会影响运算速度和二值化效果。

在文本图像的二值化过程中,局部窗口大小 w 的选取和笔划宽度密切相关,只要 w 能大于笔划宽度,就能避免窗口完全是目标的情况出现。而每幅文本图像中文字的字体大小可能差异很大,对应的文字笔划宽度也会有很大区别。所以必须找到一种自适应选取 w 的方法,才能满足局部二值化算法的需要。

文本图像有其自身的特点,在二值化过程中要求能较好地保持文字笔划信息。我们可以综合考虑图像的整体与局部信息,通过几项改进来避免二值化过程中出现断笔和伪影现象:自适应提取窗口宽度来消除断笔;引进新的阈值 T_3 和 T_4 来消除伪影,其中:T_3 是对由 Bernsen 算法得到的阈值 T_1 进行平滑处理的结果,T_4 是通过公式(1)得到的最终的目标和背景期望值计算得到:

$$T_4=\left(m_b(g_n)-m_f(g_n)\right)/2 \tag{4}$$

图 1 给出了新方法的流程框图。

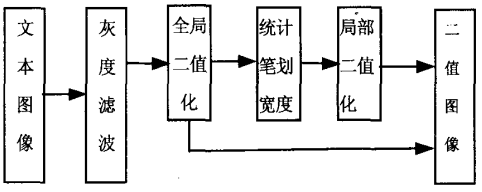


图 1 全局与局部相结合的二值化方法流程框图

新的二值化方法的具体步骤如下:

(1)首先对文本图像进行灰度滤波,滤除图象中的噪声。我们选用一种在均值滤波和中值滤波的基础上发展的一种边缘保持滤波器,该滤波器在滤除脉冲噪声的同时,又不至于使图象边缘十分模糊。具体说来,该算法首先对灰度图象的每一个象素点(x,y)取适当大小的一个邻域(如 3×3 邻域),分别计算(x,y)的左上角子邻域、左下角子邻域、右上角子邻域、右下角子邻域的灰度分布均匀度 v,然后取最小均匀度对应区域的均值作为该象素点的新的灰度值。计算灰度均匀度的公式为:

$$V=\sum(g(x,y)-\bar{g}(x,y))^2 \tag{5}$$

(2)全局二值化。用公式(1)得到全局二值化阈值 T,对图像进行二值化处理。

(3)计算笔划宽度。根据上步得到的二值化图像,统计得到笔划宽度 d。

笔划宽度按以下方式计算:在文本图像的全局二值化图像中随机取 100 个点,对其中的每一个点,若该点处于背景区,则不再考虑;若该点在目标区域中(黑色象素点),则从该点沿水平和垂直方向分别延伸,直到离开目标区域,统计得到水平和垂直方向上黑色象素点的长度。当垂直和水平方向的长度均超过一个阈值时,放弃该点,否则,取其中的较小长度值作为该点得到的字符笔划宽度;统计所有处于目标区域的点得到的字符笔划宽度,然后取其中的最大值为笔划宽度 d。

(4)对文本图像进行逐点二值化。根据点(x,y)的灰度 g(x,y)的大小,将所有的点分为两类:一类是灰度满足 $T(1-\alpha)\le g(x,y)\le T(1+\beta)$ 的点,对它们进行局部二值化处理;其余的点则进行全局二值处理。

①考察图像中每个点,利用下式进行二值化处理

$$b(x,y)=\begin{cases}0 & g(x,y)<T(1-\alpha)\\ 1 & g(x,y)>T(1+\beta)\end{cases} \tag{6}$$

其中,$\alpha,\beta\in(0,1)$一般可取 02~0.4。

②对灰度值不在(6)式范围内的点,即满足 $T(1-\alpha)\le g(x,y)\le T(1+\beta)$ 的点,采用改进的 Bernsen 算法进行二值化处理:

a 利用前面第二步得到的笔划宽度 d,选取窗口宽度 W=2d+1。用公式(2)计算灰度值在 $T(1-\alpha)\le g(x,y)\le T(1+\beta)$ 内所有点的局部阈值 $T_1(x,y)$。

b 引入阈值 $T_2(x,y)$,用以确定局部考察窗口内极大极小值的变化

$$T_2(x,y)=\max_{-d\le k,l\le d} g(x+k,y+l)-\min_{-d\le d,l\le d} g(x+k,y+l) \tag{7}$$

c 对由 $T_1(x,y)$所确定的阈值曲面进行平滑处理,以消除光照不均等噪声所造成的阈值突变。

作为避免伪影出现的一个手段,引入阈值 $T_3(x,y)$:

$$T_3(x,y)=\operatorname{avg}_{-d\le k,l\le d} T_1(x+k,y+l) \tag{8}$$

d 逐点二值化。如果 $g(x,y)>T_3(x,y)$ 和 $T_2(x,y)>T_4$ 同时满足(其中$T_4=\left(m_b(g_n)-m_f(g_n)\right)/2$),则 b(x,y)=0,否则 b(x,y)=1。$T_4(x,y)$成立将保证考察点所在窗口一定跨越目标和背景,这是消除伪影的另一个手段。

4 实验结果

下图为分别使用经典方法与新方法对一张文本图像进行二值化得到的结果,由结果图 2 可以看出,综合二值化算法和全局迭代算法相比较,保持了较多的文字笔划的完整性,例如:原图中的“记”、“者”、“国”等字,使用全局迭代算法得到的结果中明显有断笔现象;而 bernsen 算法的结果中则伪(见第 124 页)

技术创新

VPN的基础,它可以为路由器之间、防火墙之间或者路由器和防火墙之间提供经过加密和认证的通信。虽然它的实现会复杂一些,但其安全性比其他协议都完善得多。

3.2 安全电子邮件

电子邮件的安全需求也是机密、完整、认证和不可否认,而这些都可以利用 PKI 技术来获得。利用数字证书和私钥,用户可以对他所发的邮件进行数字签名,这样就可以获得认证、完整性和不可否认性,用加密的方法还可以保障信息的机密性。目前发展很快的安全电子邮件协议是 S/MIME, 这是一个允许发送加密和签名邮件的协议。该协议的实现需要依赖于 PKI 技术。

3.3 电子商务

PKI 技术是解决电子商务安全问题的关键, 综合 PKI 的各种应用, 我们可以建立一个可信任和足够安全的网络。在这里,我们有可信的认证中心,典型的如银行、政府或其他第三方。在通信中,利用数字证书可消除匿名带来的风险,利用加密技术可消除开放网络带来的风险,这样,商业交易就可以安全可靠地在网上进行。

参考文献

[1]李明柱,PKI 技术及应用开发指南,http://www-900.ibm.com/

[2]Matt Bishop,Computer Security:Art and Science,清华大学出版社

[3]Bruce Schneier ,Applied Cryptography:Protocols,Algorithms,and Source Code in C,John Wiley & Sons

作者简介:史创明,男, 1963-,汉族,安阳师范学院计算机科学系副教授,籍贯:河南;研究方向:计算机网络,中文信息处理;email: scmscmscm@163.com;

Author brief introduction:Shi,Chuangming.Sex:male. Nnationality:Han.Title of technical post: Associate professor.Native place:Henan.Studying subjects:networks and Chinese information processing.Email: scmscmscm@163.com .Unit: Department of Computer and Science, AnYang Teacher´ s College.

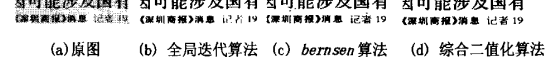
(455000 安阳师范学院计算机科学系)史创明 王立新

(Department of Computer and Science, AnYang Teacher´ s College Henan China,455000) Shi, Chuangming Wang,Lixin

(投稿日期:2005.4.22) (修稿日期:2005.4.30)

技术创新

(接第 57 页)影现象比较严重,并且在较大字体中出现了较多的断笔现象;和 bernsen 算法相比,综合算法在大大提高运算速度的同时, 明显地减少了伪影,断笔现象也大大减少。综合算法在三种方法中,实验效果最佳。



(a)原图　(b) 全局迭代算法　(c) *bernsen* 算法　(d) 综合二值化算法

图 2　算法效果比较图

5 总结

本文根据文本图像特点,综合全局二值化算法与局部二值化算法的思想提出了一种综合二值化算法。该方法自适应调整局部二值化时的窗口大小 W,并引入了新的判决阈值, 有效地消除了断笔和伪影现象。此外,新算法的运算速度较传统局部二值化算法也有较大提高。实验表明,这一新的文本图像二值化方法是有效的。

参考文献

[1]Bulent Sankur,Mehmet Sezgin. Image Thresholding Techniques:a Survey Over Categories[J]. Journal of Electronic Imaging January 2004,13(1).

[2]N.Otsu,A Threshold Selection Method From Gray Level Histograms,IEEE Transactions on System,Man and Cybernetics,SMC-9(1979)62-66.

[3]T.W.Ridler,S.Calvard,Picture thresholding using an iterative selection method,IEEE Trans.System,Man and Cybernetics,SMC-8(1978)630-632.

[4]W.Niblack,An Introduction to Image Processing,Prentice-Hall,1986,pp: 115-116.

[5]Bernsen, Dynamic Thresholding of Gray level Image, ICPR`86:Proc.Int. Conf.on Pattern Recognitio- n,Berlin,Germany,1986,pp:1251-1255.

[6] 叶芗芸, 戚飞虎, 吴健渊. 文本图像的快速二值化方法[J]. 红外与毫米波学报, October,1997 ,Vol 16.No.5:344-350.

作者简介:庄军,男,汉族,1969 年,硕士研究生,研究方向为图象处理, 模式识别 ;email: zhuangjun200303@163.com;李弼程,男,汉族,1970 年,教授,主要研究方向为图像分析与处理、目标识别与数据融合。

About author:Zhuangjun, Male, Born in 1969. Research Area: Image processing, pattern recognition; Li Bicheng, Male, Born in 1970, professor, Research Area: Image processing and analysis, Object recognition and data fusion;

(450002 河南郑州信息工程大学信息工程学院信息科学系)庄 军 李弼程 陈 刚

(Depart of Information Science, Information Engineering Institute,Information Engineering University, Zhengzhou, 450002 ,China) Zhuang,Jun Li, Bicheng Chen,Gang

通讯地址:(450002 郑州市 1001 信箱 835 号) 庄军

(投稿日期:2005.4.8) (修稿日期:2005.4.20)

一种有效的文本图像二值化方法

作者： 庄军， 李弼程， 陈刚， Zhuang,Jun， Li,Bicheng， Chen Gang
作者单位： 450002,河南郑州信息工程大学信息工程学院信息科学系
刊名： 微计算机信息
英文刊名： CONTROL & AUTOMATION
年，卷(期)： 2005,21(8)
被引用次数： 21次

参考文献(6条)

1. Bulent Sankur;Mehmet Sezgin Image Thresholding Techniques:a Survey Over Categories 2004(01)

2. N Otsu A Threshold Selection Method From Gray Level Histograms[外文期刊] 1979

3. T W Ridler;S Calvard Picture thresholding using an iterative selection method 1978

4. W Niblack An Introduction to Image Processing 1986

5. Bernsen Dynamic Thresholding of Gray level Image 1986

6. 叶芟芸;戚飞虎;吴健渊 文本图像的快速二值化方法[期刊论文]-红外与毫米波学报 1997

本文读者也读过(3条)

1. 方敏.徐俊艳.王建平.刘泓.FANG Min.XU Jun-yan.WANG Jian-ping.LIU Hong 一种新的文本图像二值化方法[期刊论文]-合肥工业大学学报(自然科学版)2001,24(2)

2. 李冠一 灰度文档图像的直接局域二值化方法[学位论文]2002

3. 陈艳.孙羽菲.张玉志.CHEN Yan.SUN Yu-fei.ZHANG Yu-zhi 基于连通域的汉字切分技术研究[期刊论文]-计算机应用研究2005,22(6)

引证文献(22条)

1. 尹翔 现代档案管理如何发挥图像处理的最大功效[期刊论文]-兰台世界 2012(29)

2. 梁英宏.王知衍 高噪声条件下基于投影的二维条码倾斜检测[期刊论文]-微计算机信息 2006(22)

3. 李文博 一种基于SVM的数字仪表显示值识别方法[期刊论文]-现代电子技术 2011(4)

4. 苑全兵.黄福 数字字符识别算法研究[期刊论文]-电子测试 2010(4)

5. 王红霞.程艳芬 改进的EM算法在分块灰度图像二值化中的应用[期刊论文]-武汉理工大学学报（交通科学与工程版） 2011(4)

6. 刘婀娜.罗予频.华成英 变形文档图像的矫正方法研究[期刊论文]-微计算机信息 2007(3)

7. 吴强.贾传荧.李冰梅 图像处理技术在档案文本数字化中的应用[期刊论文]-兰台世界 2006(7)

8. 代小红 基于图像模式识别的数字图书资料修复及应用[期刊论文]-图书情报工作 2009(3)

9. 蓝炳伟 一种基于特征点的虹膜匹配算法[期刊论文]-计算机安全 2009(1)

10. 田自君.刘艺 基于LoG算子边缘检测的图像二值化处理[期刊论文]-中国测试技术 2007(6)

11. 赵天雪.孙光民.许爽 视频文本图像增强算法研究[期刊论文]-微计算机信息 2007(33)

12. 王非.赵强.唐定勇 基于多层次信息的可视化研究[期刊论文]-微计算机信息 2006(19)

13. 刘焱.李敏勇 靶面目标图像识别算法[期刊论文]-微计算机信息 2006(36)

14. 何颖.王玲 基于特征块与小波变换的图像拼接算法[期刊论文]-计算机工程与设计 2010(9)

15. 代小红.王光利 基于模式识别的零件表面瑕疵图像提取的设计与实现[期刊论文]-表面技术 2011(5)

16. 陈龙.雷小娟.金艳.陈根方 基于投影技术的手写体工尺谱乐谱分割研究[期刊论文]-福建电脑 2011(11)

17. 梁松涛. 吕学强. 程涛. 施水才 基于分块和Lab颜色模型的字幕提取方法[期刊论文]-微计算机信息 2010(17)

18. 杨硕. 尚振宏 一种新的二维条码图像二值化算法[期刊论文]-昆明理工大学学报（理工版） 2008(1)

19. 彭兴邦. 蒋建国 一种基于亮度均衡的图像阈值分割技[期刊论文]-计算机技术与发展 2006(11)

20. 嵇新浩. 姚金良 复杂图像中文本定位的研究现状[期刊论文]-微计算机信息 2007(33)

21. 章慎锋 基于USB口汉字识别研究[学位论文]硕士 2005

22. 彭兴邦 基于机器视觉的电子桩考系统[学位论文]硕士 2006