DOI :10. 3969/J. ISSN. 1672-6693. 2010. 01. 012

数字档案图像的特点分析*

杨 有[1], 尚 晋[2]
(1. 重庆师范大学 数学与计算机科学学院, 重庆 400047 ;2. 重庆航天职业技术学院, 重庆 400021)

摘要 :随着档案图像信息系统越来越多地在 Internet 和 Intranet 上得到应用,对档案图像的处理变得越来越迫切,因此,分析和总结数字档案图像的特点就变得十分重要。本文以自然图像为对比项,以分析、总结和编程为手段,得到档案图像的固有特点、直方图特点和统计量特点。结果表明,档案图像具有较高的空间分辨率要求和较低的颜色分辨率要求,具有符号级冗余和版面高度结构化等特点 ;而且档案图像的直方图特点和统计量特点与自然图像存在明显差异。这些特点导致档案图像的增强和压缩应该有别于自然图像。由于图像特点众多,尚需进一步研究。

关键词 :档案图像 ;自然图像 ;图像特点 ;直方图 ;图像统计量
中图分类号 :TP391. 4 文献标识码 :A 文章编号 :1672-6693(2010)01-0057-04

档案是指过去和现在的国家机构、社会组织以及个人从事政治、军事、经济、科学、技术、文化、宗教等活动直接形成的对国家和社会有保存价值的各种文字图表、声像等不同形式的历史记录[1]。而档案图像是指纸质档案经数码化过程获得的电子档案,或实态档案经数字化虚拟而得到的虚态档案[2-3]。狭义的档案图像是指包含表征语言符号的元素的图像,档案图像的特点是其版面结构化程度非常明显,并且符号间具有十分明显的冗余度。然而,事实上,由于现代印刷技术的不断发展,今天的档案图像不仅包含文字,也包含图片、图形等。广义的档案图像是对狭义档案图像的扩展,它指包含了文字、图形、图片等区域并具有一定版面规范的图像。

由于诸多的原因,分析档案图像的特点,应用数字图像处理技术对档案图像进行处理是必要的。一方面,档案图像的获得经历了数字化或数码化过程[4] ,数字图像在成像过程和传输过程中必然引入噪声,如传感器或电子元件内部由于载荷粒子的随机运动所产生的内部噪声、电器内部一些部件的机械震动所导致的电流变化或电磁场变化产生的噪声、以及传输通道的干扰及量化噪声与解码误差噪声等,因此,数码化后的档案图像有必要进行图像增强,而且噪声抑制方法也有别于其它图像[5]。另一方面,档案图像在形成过程中可能产生图像倾斜,需要进行图像纠偏处理 ;档案图像在形成过程中,可能产生比较大的存储空间,不利于图像存储和传输,需要进行图像压缩处理。而且,对档案图像的深度检索与利用,涉及基于内容的图像挖掘技术,需要分析图像的版面,识别图像的内容,也必须分析档案图像的特点。

1 档案图像的固有特点

首先,数字档案图像与自然图像有很大差别。通常,数字档案图像包括文字区、图形区、图像区以及背景。很明显,文字需要较高的空间分辨率以供辨认,而对颜色的分辨率则要求不高。另一方面,自然图像则需要较高的颜色分辨率,而允许较低的空间分辨率。如图 1 所示,图 1(a)和图 1(c)分别为原始档案子图像和原始自然图像,图 1(b)的空间分辨率为图 1(a)的 1/4,图 1(d)的颜色分辨率(灰度级)为图 1(c)的 1/8,可以看出,档案图像越来越模糊,自然图像的灰度级越来越不丰富。反之,如果对图 1(a)适当降低颜色分辨率,而对图 1(c)适当降低空间分辨率,对原始图像的视觉效果影响不大。

其次,对于文本档案图像,它存在很多固有特点 :1)在文本档案图像中,符号会多次重复出现,因此,文本档案图像在符号级存在大量冗余信息,其图像压缩处理适合采用基于模式匹配的压缩方法,而非像素级和亚像素级的压缩方法 ;2)文本档案图像在层次方面高度结构化。这意味着,同一行的符号

* 收稿日期 :2009-06-18 修回日期 :2009-07-15
资助项目 :云南省 2009 年社会发展科技计划项目(No. 2009ZC128M)
作者简介 :杨有,男,副教授,博士研究生,研究方向为档案图像处理。

占用的空间位置大致相等,行与行之间的距离或者段落与段落之间的距离基本固定;3)档案图像最直观、最典型的特征是图像中存在大量的空白区域,即非文字区域,不仅文档四周存在大量的空白,行与行之间、字与字之间、甚至笔画与笔画之间均存在着大量的空白。基于此,假设图像为黑底白字,G 为全局图像平均灰度,L 为局部图像平均灰度,则当 $L < G$ 时,局部区域为前景的可能性大;而当 $L > G$ 时,局部区域为背景的可能性大;4)除此之外,文本档案图像的笔画之间存在着极大的几何形态相关性和局部稳定性;图像的明暗对比明显,层次分明,即亮度变化不平滑;图像的像素具有成块不变性和块间跳变明显的特点;图像的低频带(文字或空白)需要保真,消除噪声干扰;图像的高频带(文字轮廓边缘和尖峰噪声点)需要提升边缘,降低噪声;对档案图像的处理,客观上要求对图像实施低通滤波和高通滤波,既要去掉高频噪声,又要提升文字边缘,改善文字视觉效果;而且由于打印和扫描等诸多原因,图像一般均存在严重的噪声干扰,如背影、文字边缘变暗和细小斑点等。

图1 档案图像和自然图像在分辨率方面的区别

2 档案图像的直方图特点

图像直方图概括了图像中各灰度级的含量,一幅图像的明暗分配状态,可以通过直方图反映出来,它是图像增强的常用技术之一[6]。一幅均匀量化的自然图像其灰度直方图通常在低值灰度区间上频率较大,使得较暗区域的细节不清楚;而一幅均匀量化的档案图像其灰度直方图通常在高值灰度区间上频率较大,使得较亮区域的细节容易被忽视。下面通过图2的两幅原始输入图像来分析档案图像的直方图特点。

图2(c)和图2(d)分别为自然图像和档案图像对应的直方图,可以看出:1)自然图像的直方图分布图2(c)比档案图像的直方图分布图2(d)更加平坦,在灰度级50~220之间均有较多的像素,灰度级范围较大。而档案图像只在灰度级为220~240之间有较多的像素,灰度级范围非常窄小。这说明:表示自然图像像素所用灰度级较多,具有较多的灰度变化,即纹理;而表示档案图像像素所用灰度级较少,没有多少灰度变化,即没有纹理。2)档案图像直方图2(d)只在220~240灰度级范围内据有较多的分布,可以理解为档案图像的白色背景,而相对于背景的前景,其灰度分布微不足道,这符合档案图像中存在大量空白区域的固有特点。

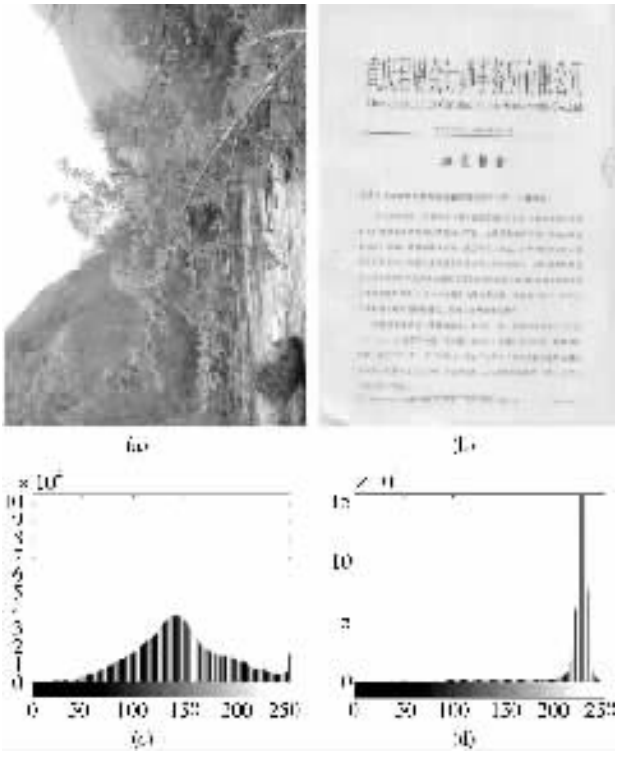

图2 原始自然图像和档案图像

直方图均衡化(Histogram equalization,HE)通过增加像素灰度值的动态范围来达到增强图像整体对比度的效果,是常见的图像增强方法之一[7]。图3为图2所示原始图像经HE处理后的结果。可以看出:1)将图3(a)、(b)和图2(a)、(b)进行比较,图3

(a)和图 3(b)具有更高的对比度,而且图 3(a)和图 3(b)出现了"过暗或过亮"的现象,即无论是自然图像还是档案图像,其 HE 增强后的图像都显得很粗犷。2)相对于原始输入图像直方图图 2(c)和图 2(d),HE 处理以后的图像直方图图 3(c)和图 3(d)更为平坦,且接近于均匀分布,即处理后的图像具有更大的动态范围。①比较图 3(b)和图 2(b),不难发现:HE 处理后的档案图像噪声明显地被放大,从主观视觉的角度来看,达到了不可接受的程度。

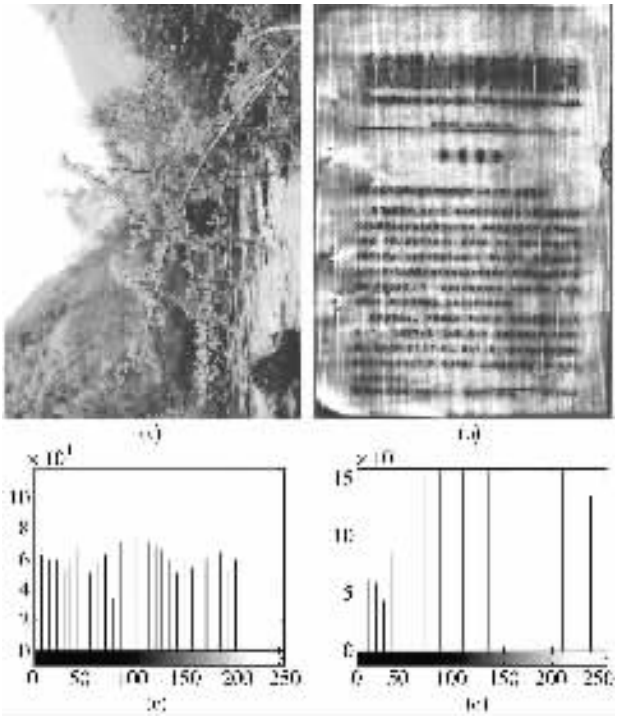

图 3　直方图均衡化处理

自适应直方图均衡化(Adaptive HE,AHE)是通过滑动窗口技术对 HE 的改进,旨在克服 HE 方法的缺点[8]。图 4 为图 2 所示原始图像经 AHE 处理后的结果。可以看出:1)比较图 3(a)、(b)和图 4(a)、(b),可以发现图 4(a)、(b)的视觉效果要好于图 3(a)、(b),即 AHE 增强后的图像没有 HE 方法所产生的"过暗或过亮"现象。2)比较图 4(d)和图 4(c),可以发现档案图像的直方图仍呈单峰状,完全有别于自然图像对应的直方图。3)比较图 4(c)和图 2(c),可以发现,AHE 使得原始图像的直方图变得更加平滑和平坦;而比较图 4(d)和图 2(d),可以发现,AHE 对档案图像的直方图整体形状改变不大。因此,AHE 对档案图像和自然图像所起的图像增强作用,其效果有明显区别。4)比较图 4(b)和图 3(b),可以发现,图 4(b)的噪声明显小于图 3(b),即 AHE 方法在抑制噪声方面比 HE 方法要好。

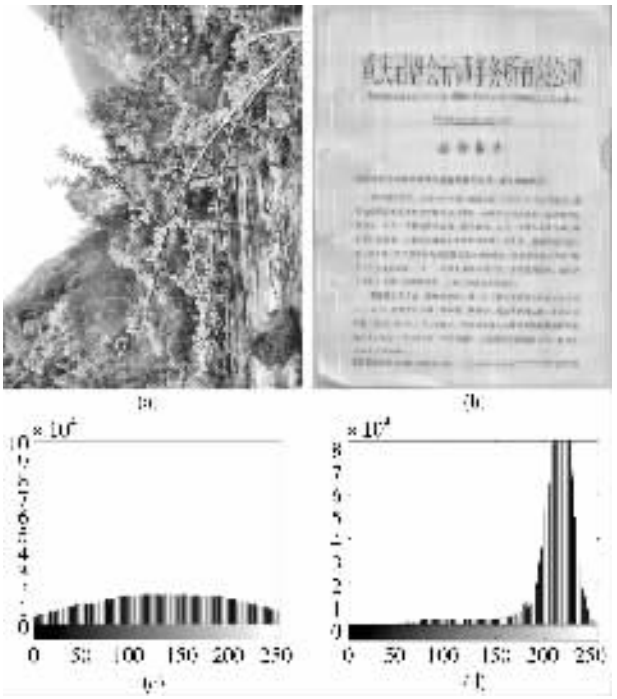

图 4　自适应直方图均衡化处理

3 档案图像的统计量分析

数字图像中的原点距和中心矩是图像处理中被广泛使用的数字特征。其阶中心矩定义为

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k, k = 2,3,4,\cdots$$

在静态图像中,一阶中心矩就是图像变量的数学期望,即亮度均值,它大致描述了灰度概率分布的中心;二阶中心矩就是图像变量的方差,它为概率分布的离散程度提供了一种度量;三阶中心矩描述了概率分布的非对称性,即灰度分布峰值相对于均值的偏离程度;四阶中心矩描述了分布曲线的尖削或平坦程度[9]。其中三阶中心矩可用偏态系数(S, skewness)进行描述

$$S = \frac{E(X - E(X))^3}{\sigma^3} = \frac{\mu_3}{\sigma^3}$$

其中 σ 为图像的标准差。四阶中心矩可用峰度系数(K, kurtosis)进行描述

$$K = \frac{E(X - E(X))^4}{\sigma^4} = \frac{\mu_4}{\sigma^4}$$

表 1　档案图像和自然图像的统计量

统计量	档案图像 D	自然图像 N
均值 μ	226.425 1	162.630 7
方差 σ^2	22.022 1	56.414 1
偏态系数 S	-432.337 9	116.545 0
峰度系数 K	1.012 0E+004	6.973 2E+003

通过对图2所示的档案图像和自然图像进行计算,得到它们的统计量如表1所示。从表中可以看出:1)μ_D明显大于μ_N,说明档案图像的亮度明显大于自然图像,这从图2的视觉主观效果可以得到验证。2)σ_D^2明显小于σ_N^2,说明档案图像的概率分布离散程度小于自然图像,这符合图2(c)、(d)所示的直方图分布特点。3)档案图像的偏态系数S_D为负值,自然图像的偏态系数S_N为正值,说明档案图像为右偏态,自然图像为左偏态,且由于$|S_D| > |S_N|$,说明档案图像的右偏程度比自然图像的左偏程度严重。4)K_D明显大于K_N,说明档案图像的灰度分布相对于自然图像比较陡峭,这也可从图2(c)、(d)得到验证。

4 结语

根据以上实验与分析,可以得到如下结论:档案图像具有高度结构化的版面,而自然图像具有丰富的纹理,它们在满足主观视觉方面对空间分辨率和颜色分辨率的要求各不相同;档案图像和自然图像的直方图特性也迥然不同,档案图像直方图的单峰状特性明显,而自然图像的直方图分布相对平坦;档案图像和自然图像的统计特性也有差异,统计特性的值也从另外的角度印证了它们的灰度分布特性。

数字图像的处理涉及增强、恢复、压缩等诸多方面,与之相适应的图像特点也名目繁多,既有空域的也有频域的,既有图像的全局特征也有图像的局部特征。本文仅从档案图像的结构化特征和直方图特征进行了详细的讨论,并与自然图像进行对比分析和实验,两者之间更多的区别与联系有望进一步研究和总结。

参考文献:

[1] 吴宝康. 档案学概论[M]. 北京:中国人民大学出版社,1988.
[2] 丁海斌. 档案虚拟论[J]. 档案学通讯,2004(2):25-28.
[3] 丁海斌. 档案管理虚拟论[J]. 档案学通讯,2004(3):23-26.
[4] 杨有. 工商档案数码化[J]. 重庆师范大学学报(自然科学版),2004,21(2):31-34.
[5] 龙兴民. 基于贝叶斯神经网络先验模型的图像去噪研究[J]. 重庆师范大学学报(自然科学版),2009,26(3):65-68.
[6] Gonzalez R C, Woods R E. Digital image processing[M]. 2nd ed. Beijing: Publishing House of Electronics Industry, 2002.
[7] 王炳健,刘上乾,周慧鑫,等. 基于平台直方图的红外图像自适应增强算法[J]. 光子学报,2005,34(2):209-301.
[8] 孙即祥. 图像处理[M]. 北京:科学出版社,2004.
[9] 刘学华,王立静,吴洪宝. 中国近40年日平均气温的概率分布特征及年代际差异[J]. 气象与环境研究,2007,12(6):779-787.

Analysis in the Characteristics of the Digital Document Image

YANG You[1], *SHANG Jin*[2]

(1. School of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047;
2. Chongqing Aerospace Polytechnic, Chongqing 400021, China)

Abstract: With the application development of more and more document image systems used on Internet and Intranet, more and more document image processing requirements are increased. So it's necessary to analyze and summarize the characteristics of the document image. By the comparition of the natural images, through analyzing, summarizing and programming, three kinds of characteristics, including document potential attributes, image histogram attributes and image statistical attributes are pointed out. The study results show that the document image has many potential attributes: high spatial resolution and low color resolution requirement, character or symbol redundancy and high hierarchical page structure. The study results also showed that there are obvious differences between the document image histogram attributes and the natural image statistical attributes. From these characteristics, the document image processing including enhancement and compression was different from the natural image. A further study of needed to continue because there are many types of digital image characteristics.

Key words: document image; natural image; image features; histogram; statistical variables of image

(责任编辑 游中胜)