基于文档分割的图像兴趣域检测及编码方法

朱庆生,王 茜,傅鹤岗,吴中福
(重庆大学计算机学院,重庆 400044)

摘 要: 结合数字图书馆系统建设和应用的需求,文章提出一种基于文档分割的自适应文档图像兴趣域编码方法.文章针对数字化文档结构特征,详细描述了一种基于块缩图和涂染技术的快速图文分割算法,该算法分割图文时不受文本倾斜和插图区域不规则的限制;在将插图域和文本域进行准确分割的基础上,文章提出了一种自适应生成插图兴趣域屏蔽图和兴趣域位移法的压缩编码算法,最后给出采用该方法压缩含插图的扫描文档的示例.

关键词: 图像分割; 兴趣域编码; 涂染技术; 文档压缩
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12A-1943-04

An Approach to Detect and to Code ROI of Scanned Document Based on Segmentation

ZHU Qing-sheng, WANG Qian, FU He-gang, WU Zhong-fu
(*Dept. of Computer Science, Chongqing University, Chongqing* 400044, *China*)

Abstract: This paper proposes a new approach to adaptively detect and to code the region of interest of scanned document, which makes use of the result of a fast segmentation algorithm based on the bi-level reduced image in order to satisfy the needs of application of the digital library. The paper firstly describes in detail a block technique to reduce the original image, a modified smearing method to simplify computation and a fast segmentation algorithm. Based on the result of document segmentation, the paper then introduces a generating algorithm of ROI mark image and the max-shift method of document compression. Finally the paper shows an example of compressing scanned document, which includes both picture areas and text areas, by means of our scheme.

Key words: image segmentation; region of interest; smearing technique; document compression

1 插图兴趣域概念

传统图像压缩方法都假设图像信号在空间域上是密切相关的,但对于扫描文档图像而言,这种假设是不适当的.简单用传真标准 CCITT FG3 或 FG4 压缩和存储扫描文档,虽然能够获得高压缩比的文档,但丢失了颜色信息.在 ITU T.44 标准[1]中推荐的 MRC 标准描述了一种替代方法,该方法建议将文档划分为背景图、前景图和标记图三层分别用不同的编码器进行编码,但这个建议并没有定义编码过程和如何生成三层图像.

本文针对文档图像特征,引入兴趣域编码的概念.文档图像的兴趣域(ROI: Region of Interest)是指扫描文档中需要用高分辨率表示的插图部分,我们称文档中 ROI 以外的文字部分为背景域(BG: Background).在文档压缩编码技术中引入插图兴趣域概念后,有利于实现对插图区域的高质量编码,而且在恢复时允许用户对插图区域提出解压要求.当读者对文档中的插图区域感兴趣时,系统可对这块区域采用高质量、低压缩比,而对其它区域采用低质量、高压缩比,从而取得读者满意的效果.此外,在网络环境浏览文档时可以方便地实现文档图像的渐进传输(Progressive Transmission),即先传输文本和插图的轮廓信息,随后逐步传输插图的精度信息、不断细化插图的质量以满足用户需求.

在日益普及而传输带宽有限的网络应用中,兴趣域编码和渐进传输技术有非常广泛的应用价值.例如,在下载含有高精度图片的文档资料时,该技术可以使用户在看到插图轮廓后再决定是否继续下载它.兴趣域图像压缩技术能够在编码过程中结合用户主观判断,对图像感兴趣区域进行交互式传输编码,只有当接收方需要时,系统才使兴趣区域的图像更清晰.实现这一特性对普及数字化图书馆的应用有着十分重要的意义.本文提出了一种基于对扫描文档分析和图文分割技术来确定文档中的插图兴趣域的编码方法.

2 基于文档分割的兴趣域编码

在数字化图书馆、办公自动化等文档压缩技术中,通常对插图部分的质量要求都高于对文本部分的质量要求.为了使 ROI 图像域比 BG 文本域精度更高,在确保 ROI 质量的前提下

收稿日期:2002-06-10;修回日期:2002-09-19
基金项目:教育部优秀年轻教师基金(教人司[2000]11);重庆市科技攻关项目(No.2000-5672)

提高压缩比，必须区别 ROI 和 BG 的不同区域，以减少编码 BG 信息所需位数.基于这一需求，我们力图通过一种快速分割文本域和图像域的方法确定出文档中的图片区域，并将其视为文档编码的兴趣域.在人类可视效果失真不明显的前提下，实现文档图像的高压缩比.

基于缩图和涂染技术的自适应文档插图兴趣域检测及编码算法流程如图 1 所示.图像编码过程包括的主要步骤如下：

①采用灰度图像变换、亚抽样和涂染技术获得缩小的简化二值图像；

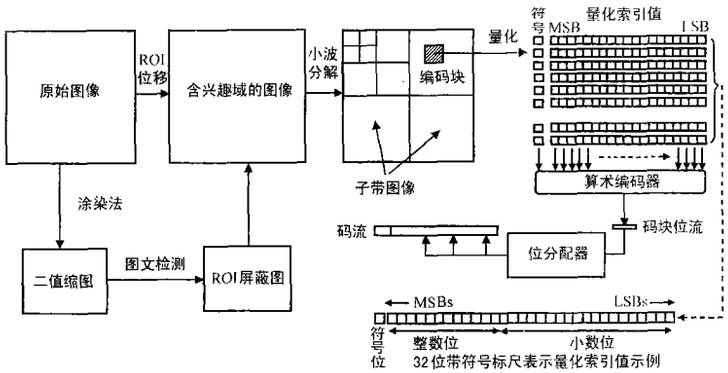


图 1 基于文档分割的图像兴趣域检测及编码流程图

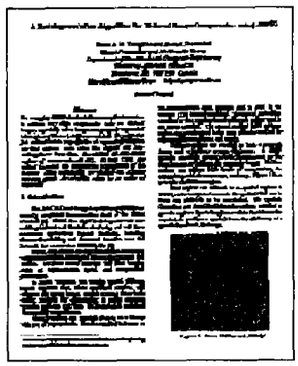


图 2 简化缩图涂染结果示例

②采用基于略图的快速文档分割算法，自动检测并确定文档中插图兴趣域，生成 ROI 屏蔽图；

③采用最大位移法突出 ROI 信息码，获得突出了兴趣域的图像；

④采用小波变换对含兴趣域的图像进行多层次图像分解，同时生成各层次上相应的 ROI 屏蔽码；

⑤将每个小波子带划分成编码块，对编码块的小波系数进行量化.图中给出了 32 位码块中带符号量化索引值的表示示例；

⑥采用基于位平面的算术编码器对每个码块的量化索引值进行独立的熵编码后形成位流；

⑦用多路位分配器从所获得的码块位流中生成最后压缩码流.

在解码端，根据所接收的位流信息，按位平面顺序对码块的小波系数进行渐进式解码、逆量化、逆变换和图像重建.显然，解码效果依赖于码块编码传送的顺序，在实际中可采用不同的渐进传输模型.可以采用分辨率递增模型(从低频子带的低分辨率码块开始解码)或质量递增模型(从失真最小的码块开始解码).

3 自适应兴趣域检测算法

借鉴文献[2]所提出的“涂染法”的思想，经算法改进后，采用块技术替代压抽样方法从原始图像中获得一张二值简化缩图，缩小图像的目的是为了减少生成门限值时的噪声影响和简化文档分割的计算复杂性.算法的具体思想方法如下：

第一步 将原始文档扫描图转换为一张灰度图，并用低通滤波器和块技术将原图缩小，使简化缩图中每个像素对应于原图中 $N \times M$ 的块，当且仅当原图中相应块的所有像素的灰度值均为“白”色(或其色值低于白色阈值)时，缩图中相应像素取值为“白”，其他所有像素均取值为“黑”色.采用这种方法进行化简的效果是对缩图进行涂染，一张用 200dpi 扫描的文档图像，用 8×8 块进行涂染的结果如图 2 所示，简化缩图中黑色部分在原图中是一张插图.

第二步 识别和分析二值缩图中的连接区域，对于高分辨率扫描文挡图像(大于 300dpi)采用 4 邻接连通性检测，对于低分辨率扫描文挡图像(200dpi)采用 8 邻接连通性检测.一旦找出连通域，用连通度门限值排除图符识别区域，然后抽取并检查每个图符是否为文本字符.二值缩图中最重要的一个特性是插图区域通常为一大片连通的黑块.因此，在缩图中图像域将以大块连通域的形式出现，其检查结果为非文本图符区域；而文本域将以一些小的连通域的形式出现，其检查结果为文本字符.为了确定缩图中连通块所对应的图像域是否为插图兴趣区域，我们引入连通符号的“权”的概念.这里“权”是指连通符号中所包含的像素数目，如果一个连通域的权值大于权的门限值，则认为它是插图区域；如果连通域的边界域大于某个门限值，则它可能是线图域.原则上讲，门限值可以根据具体情况来选取，其值取得较大时，不能检测出小的图域，其值取得较小时，又有可能把文字密集的文本域误分类为图片域.

在图文分割的具体实现过程中需要考虑门限值和参数.最重要的参数是块的尺寸 N 和 M 的选择，考虑二值图像按水平方向每个字节存 8 个像素，N 取 8 可以比较方便地实现按字节进行操作，以提高检测速度.尽管 M 取值没有限制，但取 $M = N$ 可以保持图像的方向性不变.通常用 8×8 块生成二值简化缩图可以获得很好的分割效果，但用更大尺寸的块生成缩图将进一步提高图像的检测速度，缩短插图检测所需要的时间.

在文本域的字符行之间通常保留有空格，但对行距极小的文本域在图像缩小后有时也可能会出现邻接行相连的现像，采用 4 连通法符号抽取可以减少这种现像的发生.

第三步 完成兴趣域屏蔽码生成.为区别插图区域 ROI 和文本区域 BG，引入 ROI 屏蔽图来指定需要高质量编码的插图兴趣区域.ROI 屏蔽图实际上是描述插图区域的一张位图，它用二元值 $M[X,Y]$($X \times Y$ 是原图像的尺寸)来指定被编码图像域中属于 ROI 的像素点.

$$M(x,y)=\begin{cases}1, & \text{当像素点}(x,y)\text{属于插图域时}\\ 0, & \text{当像素点}(x,y)\text{属于文本域时}\end{cases}$$

$M(x,y)$的初值均为"0",在对被涂染的缩图进行非文本连通域检测的过程中,将根据检测结果动态修改其值.一旦简化缩图的一个连通域被确认为图像域时,其相应像素集所对应原图的块集所包含的所有像素的屏蔽码均被置"1".

该方法所确定的图像兴趣域不受区域形状的限制,但兴趣域边界的划分精度与缩图尺寸的选择密切相关.实验表明,选 8×8 块对插图边界没有明显的影响,是一种实用的、具有自适应能力的兴趣域定界方法.

4 位移法兴趣域编码

兴趣域编码的基本思想是让插图 ROI 信息的编码精确度优先于文本信息的编码.基于动态生成的 ROI 屏蔽码,对插图区域的信息码进行放大后再编码.采用小波压缩编码技术进行编码时,基于尺度缩放方法的编码过程如下:

①计算小波变换系数;

②如果有插图 ROI 屏蔽码,导出子带中相应的 ROI 屏蔽码;

③按一定的比例尺缩小(降低)文本域的系数;

熵编码按位平面的重要性进行,即从最重要的平面开始,对图像的小波系数进行渐进式熵编码.要对插图域高质量编码,就应该突出插图域的信息码,尽量使插图 ROI 码在编码中成为具有重要意义的位平面 MSBs (Most Significant Bit-planes),而使文本域的信息码成为具有次要意义的位平面 LSBs (Least Significant Bit-planes).位移法编码是指在 ROI 屏蔽图的控制下,让文本码值向 LSB 方向移动 s 位,并在其首部补 0(相当于其值缩小 2^s 倍),同时将 ROI 码值尾部补 0(相当于其值放大 2^s 倍),从而突出 RIO 码,如图 3(b),(d).

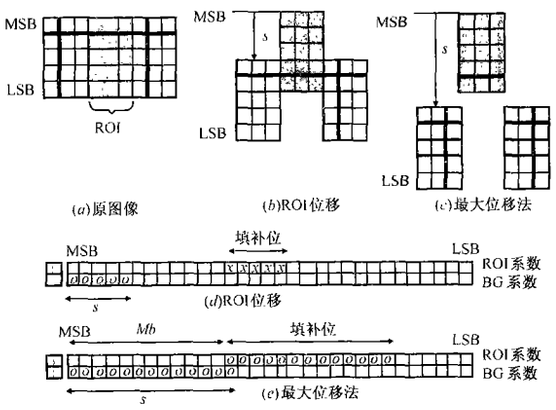


图 3　基于位移法和最大位移法图示

在解码端,通常需要 ROI 屏蔽码来区别 ROI 信息和 BG 信息.为了能够从码流本身的信息中区别 ROI 和 BG 系数,采用最大位移法(Max-shift Method),即取标尺移动位数 s(缩放倍数)为最大,使得再 BG 和 ROI 的所有码块的位平面之间没有覆盖,见图 3(c),(e).

确定尺度 s 的方法是由编码器先扫描所有量化系数,选择适当尺度 s,使得 $s\geq\max(M_b)$,这里 M_b 表示当前图像域编码块中任意 BG 系数 $q_{BG}(x,y)$的最大位平面.换句话说,所选择的尺度 s 位移后能够足以使 ROI 的最小非零系数 $q_{ROI}(x,y)$大于 BG 的最大系数 $q_{BG}(x,y)$.

最大位移法不需要对 ROI 形状进行编码和解码,它通过门限值 2^s 与系数本身比较便可区别 ROI 或 BG 的系数,因此,解码时不再需要 ROI 屏蔽码.其编码和解码步骤如下:

在编码端:生成 ROI 屏蔽码 $M(x,y)$;确定最佳量化尺度 s;用 s 降低由 $M(x,y)$所指定的 BG 域小波变换系数的量化尺度;用 RGN 方法将 s 写入码流中;用常规方法对量化变换系数进行熵编码.在解码端:将所收到的量化系数与 2^S 进行比较,所有系数小于 2^S 均认为是 BG,并提高这些系数的尺度为 2^S.

小波逆变换实际上是把 L 和 H 子带合二为一,显然,对高质量图像同时需要这两个子带的系数;如果再对这两个子带继续逆变换跟踪,则需要将四个子带合并为二个,这就需要四个子带的小波系数.在解码端,可以根据逆变换对系数的要求,采用 W5×3 或 W9×5 滤波器来生成子带中的 ROI 屏蔽码.

设 $X(2n)$和 $X(2n+1)$分别表示在尺度 s 上的图像码块网格中偶数和奇数位置上的系数,$L(n)$和 $H(n)$分别表示与它对应的低通 L 和高通 H 子带的第 n 个位置上的系数. $W5\times3$ 滤波器中原码块与相应子带的对应关系有:

$$X(2n)=L(n)-\frac{H(n-1)+H(n)}{4}$$

$$X(2n+1)=\frac{L(n)+L(n+1)}{2}+\frac{-H(n-1)+6H(n)-H(n+1)}{8}$$

即 $X(2n)$有 3 个相关点 $L(n)$、$H(n-1)$和 $H(n)$,$X(2n+1)$有 5 个相关点 $L(n)$、$L(n+1)$、$H(n-1)$、,$H(n)$和 $H(n+1)$.如果 $X(2n)$或 $X(2n+1)$在 ROI 中,它们在 L 和 H 子带中相关位置的系数就应该在 ROI 中.在每一步分解的过程中,由综合滤波器在垂直方向和水平方向扩展当前子带(或原图像)的 ROI 屏蔽码.

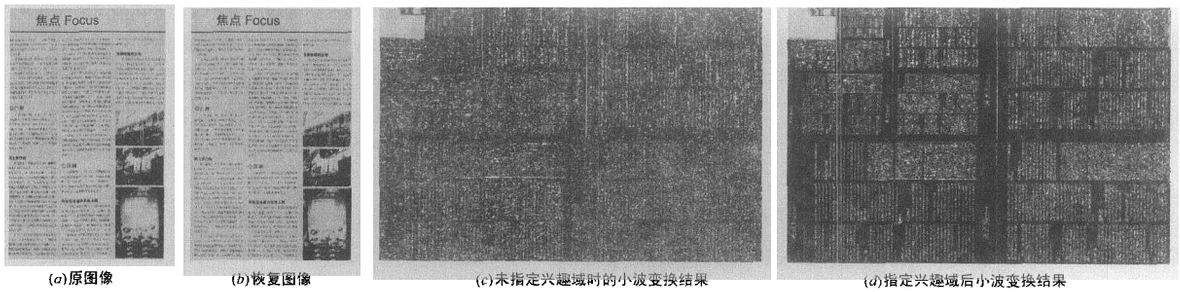
5 结束语

本文针对日益普及的文档数字化建设和应用的要求,提出了一种基于图文分割方法把文档图像中插图作为图像兴趣域进行压缩编码的方法.文章针对数字化文档可视化特征,详细描述一种基于块缩图和涂染技术的快速图文分割算法,该算法不受文本倾斜和插图区域不规则的限制.在对数字文档页中图像区域和文本区域进行有效分割的基础上,文中介绍了一种自适应确定图像兴趣域的压缩编码算法.考虑人类视觉敏感度等因素,数字化文档中的文字部分通常允许适度的失真.因此,我们在处理文档压缩编码的过程中,系统一旦检测到文档图像中含有高质量的插图(尤其是彩色插图)区域时,则将插图区域视为文档中要求高保真的区域,并自动生成兴趣域屏蔽码,然后采用兴趣域位移法和位平面渐进编码技术对其进行压缩编码.

图 4 展示了采用该方法压缩文档的实际效果.对照图 4(a)和图 4(b),不难发现,插图部分经压缩和解压缩后图像质量没有任何变化,但文字部分的背景上有明显的"失真".通常

由于受文档页背面印刷符的影响,扫描时往往产生人们所不希望的背景噪声,当文档区域的像素位移后,这些背景噪声已明显消失,可见,这种人为使文本区域"失真"的结果既可以提高压缩比,又可以提高文档质量.

文章详细描述了在压缩编码中如何自动确定图像兴趣域的几个步骤及其算法思想.关于小波变换压缩算法读者可参考 JPEG2000 等相关文献[3].该方法与目前流行的压缩标准相比,在网络环境中可获得更高压缩比和质量的数字文档压缩和传输效果.图 4(*c*)和图 4(*d*)分别给出了当采用三层小波滤波器"零树法" 进行压缩时,采用或不采用"插图兴趣域编码"对不同频域图像结果的明显影响.



(*a*)原图像 (*b*)恢复图像 (*c*)未指定兴趣域时的小波变换结果 (*d*)指定兴趣域后小波变换结果

图 4 插图兴趣域压缩编码结果

参考文献:

[1] ITU Recommendation T.44. Mixed Rarer Content (MRC) Mode[S].

[2] Saitoh T, Pavlidis T. Page segmentation without rectangle assumption [A]. Proc. of the 11th Int. Conf. on Pattern Recognition[C]. Saint Malo, France: CPR, 1991.

[3] ISO/IEC JTC1/SC29 WG1 N1890R. JPEG 2000 Part 1 Final Committee Draft[S].

[4] Zhu Q S, Li B. An algorithm of VBR coding for video transmission[A]. See: M. H. Hamza. Computer Graphics & Imaging[C]. Canada: IASTED/ACTA Press, 1998.

[5] Li B, Zhu Q S. VBR coding algorithm based on wavelet transform[J]. Computer Engineering & Science. 1999, 21(1): 4 - 7.

[6] Jain A, Yu B. Document representation and its application to page decomposition[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(3): 294 - 312.

[7] Melas D E, Wilson S P. Double markov random fields and bayesian image segmentation[J]. IEEE Trans. on Signal Processing, 2002, 50(2): 357 - 365.

作者简介:

朱庆生 男,1956 年生于重庆,曾为英国伦敦大学、美国伊利诺大学、香港浸会大学访问研究员,先后发表论文 50 余篇,现为重庆大学计算机学院教授,中国计算机学会常务理事,主要研究方向为多媒体数据压缩技术、网络信息系统及软件开发环境等.

王 茜 女,1964 年生于重庆,重庆大学计算机学院副教授,主要研究方向为电子商务安全技术、网络数据传输技术.

傅鹤岗 男,1950 年生于重庆,重庆大学计算机学院副教授,主要研究方向为多媒体技术及应用、软件工程.