一种优化的文档图像分割方法

朱庆生 林 杰 张 敏
(重庆大学计算机学院 重庆 400044)

摘 要 文档图像在数字图书馆、电子商务以及电子政务等工程中已获得广泛应用。如何对文档图像进行有效的转换、存储和传输,成为人们研究的焦点。将文档图像分割成不同的区域,根据不同区域的特点分别进行处理,成为一种有效的解决方案。本文在传统的块分割和图层分割方法的基础上,提出了一种优化的文档图像分割思路,对这两种方法进行了合理的综合处理,能够取得更好的效果。

关键词 文档图像,块分割,图层分割,图像压缩

An Optimal Document Image Segmentation Method

ZHU Qing-Sheng LIN Jie ZHANG Min
(Department of Computer Science, Chongqing Unniversity, Chongqing 400044)

Abstract In recent years, document images are widely used in OA, digital library, e-commerce, and electronic government, etc. How to efficiently convert, storage, and transmit the document image has become the focus of research. Dividing the image into difference regions, and processing them according to their characters, which is a better resolution. The paper proposes a optimal document image segmentation method based on the integration of the block segmentation and layer segmentation, which is proved a good outcome.

Keywords Document image, Block segmentation, Layer segmentation, Image compression

1 引言

近年来,随着信息技术的发展,文档图像在办公自动化、数字图书馆、电子商务、电子政务等项目中已获得广泛的应用[1,2],在这些应用中,数字化纸质文档通常采用扫描方式形成文档图像进行存储、传输、显示和打印。为了保证用户能够有效地处理文档图像,人们不仅希望文档图像存储量应尽可能小、处理速度应尽可能快,而且要求文档图像质量应尽可能好。

考虑一幅 400～600dpi(dots per inch)的彩色文档扫描图像,未压缩时其大小约为 45～100Mb,这么大的图像显然难于快速处理,这就要求对文档图像进行高倍压缩,在保证图像质量的前提下尽可能地提高压缩比。

文档图像的特点不同于一般自然图像,它由一些具有特定性质的区域块组成,包括文字、背景、线图、图片等。通常,文字和线图保存了较多的图像细节以及结构信息,具有较高的空间分辨率的特征,但对颜色分辨率的要求不是很高;背景和图片则保存了较多的颜色信息,通常要求较高的颜色分辨率,但对空间分辨率的要求不高。因此,在设计文档图像压缩算法的时候,如果根据不同区域块的特点,对不同的分割块采用不同的压缩技术,显然能够获得更好的压缩效果。传统的压缩技术(如 JPEG、JBIG 等)都是假定待压缩的图像在空域上具有相同的性质,因此简单地将它们用于文档图像处理时,难于获得最好的压缩效果。

为了对文档图像进行有效的压缩,首先必须将文档图像分割成不同的区域。传统的文档图像分割的方法主要有两种,基于块的分割和基于图层的分割。但是这两种分割方法都有一些不可克服的弊端,这一点将在文章的第 2 部分作进一步的讨论。

本文中提出了一种优化的文档图像分割方法,将这两种分割技术进行有效的综合,能够取得比传统的方法更好的效果。

2 块分割和层分割

传统的文档图像的分割主要是针对于二值图像,通常的做法是把文档图像分割成不同的块,然后对于每一个块进行分类(一般是分为文字块和图片块),这种方法称作块分割(或图文分割)。最早的应用是由 Wahl 等在 1982 年提出的 CRLSA(Constrained Run Length Smoothing Algorithm)算法[3]。该算法将二值文档图像划分为多个块,然后根据一些统计特性,如黑色像素的总数,图像水平方向白色和黑色像素的变化统计等,将这些块划分为文字区域或图片区域。1993 年,Chauvet 和 Coworkers 在 CRLSA 算法的基础上提出了一种递归的块分割算法,采用可变长度的结构元素来提取块分类的特性[4]。近年来,基于块的分割算法更多地应用于灰度图和彩色文档图像的处理中。在这些算法中,有的算法是直接基于图像特征,采用阈值化的方法来进行处理;有的则是利用离散余弦等变换系数的特征来分割文字区域和图片区域,例如,Murata 在 1996 年提出的分割方法就是一种基于离散余弦变换系数的方法[5]。

另一种文档图像的分割技术是基于图层的分割技术,它将文档图像的不同组成部分,如文本、背景、图片等看作具有

朱庆生 教授,博导,主要从事网络信息技术、文档图像处理、多媒体数据压缩、软件开发环境等方向的研究。林 杰 硕士研究生。张 敏 硕士研究生。

不同的纹理的图像，将它们分割成不同的图层。通常是将文档图像分割成前景、背景和标记层，分别表征文字的颜色、文字轮廓、背景颜色以及彩色图片。在图像重建的过程中，通过标记层来控制是从前景还是从背景中来选择像素点的值。当标记层的值为1时，从前景层中选择，值为0时从背景中选择。ITU建议的国际标准MRC图像模型是一种应用广泛的文档图像的图层表示模型[6]。1998年，Bottou、haffner等在MRC模型的基础上，提出了一种称为DjVu的文档图像的高质量压缩的技术，采用多尺度的二色聚类的方法将文档图像的前景、背景分离[7]。1999年，Hui Cheng和A. Bouman提出了一种多分辨率的文档图像的分割算法TSMAP，并且采用RDOS算法将扫描文档图像分割成单色块、双色块、图像块以及其他块，然后采用不同的压缩算法分别进行压缩，在相同的比特率下，能够取得比JPEG、DjVu、SPIHT等算法更好的压缩效果[8]。2000年，R. L. Queiroz采用优化的块域值的方法分割文档图像，用于多层文档图像的压缩，这种方法能够很方便地将一副复杂的文档图像进行分割[9]。

虽然块分割和图层分割的方法都获得了广泛的研究和应用，但是这两种分割方法都有各自不可克服的弊端。块分割的方法不能够将文字的颜色和文字区域的背景提取出来，而图层分割的聚类方法，虽然能够很好地将图像中的各部分都提取出来，但是图片中对比度较大的地方很容易被错误地分到不同的图层中去，这些问题都会造成压缩时的困难。显然这是我们所不希望得到的结果。

本文的作者试图将这两种分割方法合理地结合起来，运用到文档图像的分割中，提出一种优化的文档图像的分割方案。

3 提出的方法

要想获得较大的压缩效果，必须对图像进行有效的分割，对于不同的分割块，采用不同的压缩方法。本文中提出的分割方法对于以上两种分割技术进行了有效的综合。首先用块分割的方法将文字区域和图片区域分割开来，然后对文字区域用图层分割的方法进行处理，将文字、文字的颜色和背景分割开来。最后对于分割的结果进行逻辑运算，将图片和背景区域合并构成图像的背景层，文字的颜色作为图像的前景层，文字和线图轮廓作为图像的标记层。整个图像的分割示意图如图1所示。

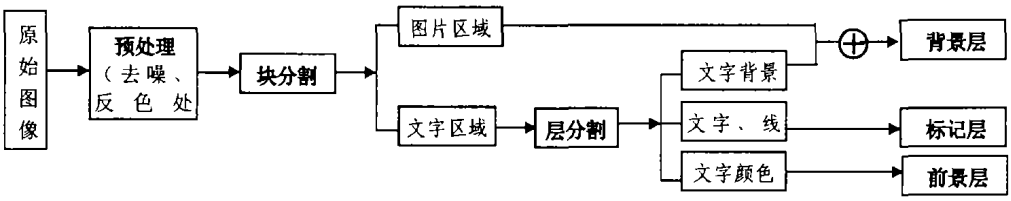

图1 文档图像的分割示意图

3.1 块分割

块分割也叫图文分割，在分割时，首先将图像划分成一个个小块，然后对每个小块进行分类，判断每个小块是属于文字块还是图像块。在进行文档图像的图文分割时，通常是将图像转化为灰度图像进行处理。块分割的方法读者可以参考Wahl的CRLSA算法[3]以及Chauvet和Coworkers的递归块分割算法[4]。

3.2 层分割

对于块分割得到的文字区域，通常还含有丰富的层次信息，包括文字以及线图的颜色、轮廓以及背景。我们采用层分割的方法进行进一步的分割，将它分割成前景层（文字颜色）、标记层（文字与线图轮廓）和背景层。

这里我们采用一种改进的多尺度的K-均值法来进行图像的层分割。K-均值法[10]是一种常用的图像分割聚类算法，用来将一幅图像分割成K个区域。在我们的图层分割中，只需要将图像分成两个区域，前景和背景。

2色图像均值聚类算法 对于一幅2色的图像，观察它的直方图，我们可以得出这样的结论，图像的前景和背景分别对应着直方图中的两个尖峰。因此，我们可以通过对图像像素颜色进行聚类的方法，很容易地将图像的前景和背景提取出来。下面给出基于传统的K-均值法的2色图像的聚类算法。

1）设定文档图像的前景颜色为黑色（暗色），背景颜色为白色（亮色）；

2）对图像中像素的颜色进行迭代，在第i次迭代的时候，根据下面的原则来确定每个像素X是前景F还是背景B：$X \in F$，如果 $\| g(x)-g(F)' \| < \| g(x)-g(B)' \|$；否则 $X \in B$；其中g(x)，g(F)，g(B)分别代表该像素的颜色值，以及当前迭代时的前景和背景的颜色值；上面的公式表明图像点的像素值离哪一个聚类中心的位置近，我们就将它归为哪一类；

3）计算所有的前景（背景）像素的平均值，更新当前的前景（背景）；

4）重复第2）步，直到前景和背景的颜色收敛。

在这里，我们根据经验，将图像中的亮色（白色）作为图像的背景，而将暗色（黑色）作为图像的前景，通常来说是正确的。对于部分图像出现的反色情形，我们可以在预处理中对它们进行处理。

块内2色聚类 实际上，我们通常需要处理的图像都不仅仅是2色。因此我们需要对上面提到的2色聚类算法进行有效的扩充，采用基于块的2色聚类算法来对它们进行处理。

首先，我们将图像分割成一个个矩形的小块，然后在每个小块内运用上面的2色聚类算法，在每个小块内得到它的前景和背景色。这样，我们可以获得两幅分辨率较低的图像。一幅用各个小块的前景色来绘制，另外一幅用各个小块中的背景色来绘制。

采用这种基于块的2色聚类算法虽然能够解决图像多色分割的问题，但是在块的大小的选择上却有很多的问题需要考虑。

- 块不能太大，否则不能够提取所有字体的颜色，所以块的最大值不应该超过图像中的最小字体的大小；
- 按照上面的原则选取的图像块，通常会比较小，很多会完全包含在非文字区域，因此只包含图像的背景色；而如果图像中有些字体很大，小的图像块也会完全包含在文字的笔划中，这样，图像块就会只包含图像的前景色。

多尺度的块2色聚类 为了解决好上面的问题，这里我

们采用一种多尺度的思想。我们不是单独地考虑某一个块，而是采用连续分割的方法，不断提高图像的分辨率。先将图像分割成较大的块，然后将这些块分割得更小，其中每一个小块的大小是前一级块大小的因数[7]。

首先在较大的块上运用块内2色图像聚类算法，获得每一个大块的前景和背景颜色，然后将它们作为下一级的小块的前景和背景的初始值，开始下一级的块内2色聚类。不同之处在于，计算下一级块的前景(背景)时，除了考虑本级的块内像素的均值外，还要加上前一级对应块的前景(背景)的加权值。算法可以描述如下：

1)将图像分割成较大的块，在这些块中进行块内2色聚类，得到各个块的前景和背景的颜色；

2)将图像分割成较小的块，这些小块是大块的一部分。将对应大块的前景(背景)作为小块的前景(背景)的初始值；

3)在小块内运用块内2色聚类算法，确定每个像素是属于前景还是背景；

4)更新当前的前景(背景)颜色，其值为所有属于前景(背景)的像素值的和，和对应大块的前景(背景)像素值的加权平均；

5)在小块内重复第3)步直到图像的前景(背景)颜色收敛。

这样一种多尺度的方法可以有效地解决前面提到的分割块大小难以确定的问题。首先逐步提高其分辨率的方法，可以保证分割块可以足够的小，从而可以捕捉到所有文字的颜色。如果当前块只包含前景或者背景中的一种颜色，那么它相应的背景或者前景值就是对应大块的背景或者是前景的颜色，对于那种完全在背景区域或者文字区域的块可以有效地解决。更重要的是这样加权平均的方法对于那些可以较好地提取前景和背景颜色的块并没有太大影响。这里，加权系数通常取的是0.8。

实际中，我们对于这种多尺度的K-均值法进行了一定的改进，使之更适合于文档图像的分割。首先，在大块进行2色聚类的时候，我们不是简单地将黑色作为图像的前景，白色作为图像的背景，而是通过图像的直方图，获得图像中最常用的两种颜色，将其中暗色作为图像的前景，而将其中的亮色作为图像的背景，这样可以极大地提高图像聚类时的收敛速度，减少算法执行的时间；另外，我们还采用一种自适应的方法来获得加权平均的权重系数，对于大多数的文档图像都能取得更好的分割效果。

3.3 区域合并

通过块分割和图层分割，我们将文档图像的文字颜色、文字轮廓、背景以及图片提取出来，然后我们需要将文字区域背景和图片合并起来构成背景层。这样做的好处有两个：一是便于图像的处理和压缩，减少冗余数据的编码；另外，图像在网络上传输的时候，我们可以根据用户的需要进行渐进传输，将文字背景区域和图片合并起来构成背景层，也便于文档图像的传输。

区域的合并通过逻辑运算来实现。在分割的时候，不在该层中的像素用缺省的颜色进行填充。这里，我们将图层分割中的图片区域用白色(也可以用其他的颜色)填充。因此，将文字背景和图片合并构成背景层，可以通过逻辑与操作完成。该过程可以用下面的公式表示：文字背景⊕图片=背景层。

4 实验结果

下面给出我们采用优化的文档图像分割方法所得到的结果。

原始图像是一幅大小为792×464的256色的、采样为300dpi的灰度图，作者首先通过块分割将文字区域和图片区域分割开来。然后对文字区域图像进行图层分割，先将其划分为9×9大小的块，在这个尺度上进行第一级的颜色聚类。在这个基础上，将图像分割成3×3大小的块，进行第二级的聚类。得到100dpi的前景和背景图像，以及300dpi的标记层。最后，将背景图像和块分割的图片区域合并构成背景层。

结论 文档图像的广泛应用，也就提出了文档图像压缩的问题，传统的文档图像的压缩技术都是基于像素空域的一致性的假定，因此很难取得最好的压缩效果。根据不同区域的特征采用不同的压缩技术成为一种合理的解决方案。本文中所提出的优化文档图像的分割方法结合了块分割和图层分割的优点，能够在实际的应用中获得更好的效果，为基于分割的文档图像的压缩做出了很好的准备工作。

参考文献

1 Witten I H, Moffat A, Bell T C. Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, New York, 1994

2 Phelps T, Wilensky R. Towards active, extensible, networked documents: Multivalent architecture and applications. In: Proc. of the 1st ACM Intl. Conf. on Digital Libraries, 1996

3 Wahk F M, Wong K Y, Casey R G. Block Segmentation and Text Extraction in Mixed Text/Image Documents. Computer Graphics and Image Processing, 1982, 20

4 Chauvet P, Lopez-Krahe J, Tafin E, Maitre H. System for an intelligent office document analysis, recognition and description. Signal Processing, 1993

5 Murata K. Image Data Compression and expansion apparatus, and Image Area Discrimination Processing Apparatus Therefor. US Patent, 1996

6 MRC. Mixed raster content (MRC) mode. ITU-T Recommendation T.44. 1999

7 Bottou L, et al. High Quality Document Image Compression with DjVu. Journal of Electronic Imaging, 1998

8 Cheng H. Document image segmentation and compression. A thesis submitted to the faculty of Purdue university. Aug. 1999

9 de Queiroz R L, Fan Z, Tran T. Optimizing block-thresholding segmentation for multilayer compression of compound images. IEEE Trans. Image Proc, Oct. 2000

10 章毓晋. 图像分割. 科学出版社, 2001