一种基于模板的档案图像压缩新方法*)

杨 有
(重庆师范大学数学与计算机科学学院 重庆 400047)

摘 要 在大容量档案图像数据库中,不仅单页档案图像内部存在数据冗余,而且档案图像页之间存在大量集合冗余。本文提出了基于模板的压缩新方法,通过定义相似图像集合的模板,充分利用图像数据的先验知识,对档案图像的内容进行分析和理解,从图像内和图像间以二维模式压缩图像数据。实验表明,该方法能够大幅提高压缩性能。

关键词 档案图像,模板,图像压缩,集合冗余

A New Compression Approach to Document Image Based on Template

YANG You
(School of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract In super large document image database, data redundancy exists not only in individual image, but also between images. A new compression approach based on template was proposed. Through the definition of template, the approach utilized the image prophetic knowledge sufficiently. After the analysis and understanding of image content, it compressed image from intra-image and inter-image. The experiment showed that the approach could increase compression ratio greatly.

Keywords Document image, Template, Image compression, Set redundancy

档案数码化是信息化建设的一项基础工作,而信息化又是覆盖我国现代化建设全局的一项战略举措。基于网络的数字档案图像应用系统必须考虑三个问题:一是从存储空间方面考虑,即档案的大小。一页 ASCII 码档案只占 2～3kB,而一页典型扫描的数字档案需要 500kB～2MB,由此可见数字档案压缩的重要性。二是从时间方面考虑,即压缩图像的有效存取。传统压缩减少了存储容量,但没有提供压缩数据的有效存取方法,比如快速传输、压缩域处理和存取等。三是数字档案的可读性,即档案的质量。因此,在遵照《中华人民共和国档案法》和一些相应行规的前提下,对数字档案图像进行压缩,就成为这类应用系统的核心问题。

1 档案图像压缩技术

档案图像属于静态图像,但它又与诸如遥感和医学等静态图像有区别,其压缩方法也具有一定的特殊性。在众多的档案图像定义中,公认的观念是:档案图像是具有结构的图像,它包含了许多表征语言符号的元素,且这些元素之间具有明显的冗余性。由此,我们知道,在档案图像中,大多数有用信息在符号级而不是在像素级或纹理级中,因此纯粹的基于像素级或亚像素级的静态图像编码方法对档案图像是不适用的。同时,由于档案图像具有高对比度、高倾斜度、局部非均匀像素模式等特点,对其采用纹理编码也是不合适的。对于文本富裕档案,由于符号多次重复出现,而且图像具有较高程度的层次结构,因此这类档案的压缩可以采取去除符号级冗余的压缩方法,即 PM&S(Pattern Matching and Substitute,模式匹配与替代)技术[1,2]。

在数字档案图像应用系统中,数据压缩不仅要考虑单幅档案图像的编码,而且要考虑档案图像之间的冗余性,即用图像的集合统计特性来代替单个图像统计特性,从而降低整个图像集合的熵。比如,在工商档案、国土资源档案等政府资源类数字档案应用系统当中,一些申请书和登记表的内容都具有相似性,各户对应档案页之间存在极大的信息冗余。以某市的工商企业登记档案系统为例,该市共有约 20 万户企业,每户企业都有相似的档案页,如公司法定代表人履历表、公司章程等。由此可以看出,在这类应用系统当中,去除这类图像间的特定冗余性是十分必要的。

为消除图像间的这类冗余,降低整个图像集合的熵,K. Karadimitriou 等人针对医学图像数据,定义了相似图像(similar image)和集合冗余(set redundancy)的概念,首先提出了提取集合冗余的两种方法:MMD 方法(Min-Max Differential,最小-最大差分法)和 MMP 方法(Min-Max Predictive,最小-最大预测法)[3],随后又提出了提取集合冗余的质心方法(Centroid Method)[4]。这三种方法以及后来的改进方法——HCM(Hybrid Compression Model,混合压缩模型)[5]统称为 SRC(Set Redundancy Compression,集合冗余压缩)技术,它们利用图像间的像素统计特性,降低整个图像集合的熵。而且随着图像间的相似性增加,集合的熵就降低得越多。

在数字档案应用系统中,由于档案图像本身具有高度结构性,相似档案图像可以划分为两类区域:一类区域的信息是固定不变的,即从区域的角度出发就可确定该部分为冗余信息,而不去考虑该区域内的像素相似性;另一类区域的信息具有变化性,它是各户档案图像页之间的真正区别,需要单独进行压缩处理。比如,在图 1 的示例中,人工填写的信息、照片、身份证、印章和签字属于信息变化区域,我们定义为 ROI 区域,需要根据这些区域的性质(图像、图形、文字)分别采取不同的压缩策略;而图像剩余的区域则属于信息非变化区域,我

*)本文受重庆市教委基金项目(KJ070805)资助。杨 有 博士生,主要研究方向为档案图像处理。

们定义为非 ROI 区域，记为 RROI，它是图像的相似区域，可以借鉴 SRC 的思想消除图像之间的冗余。此即本文提出的基于模板压缩的基本思想，我们简称为 COT（Compression Based on Template）方法。

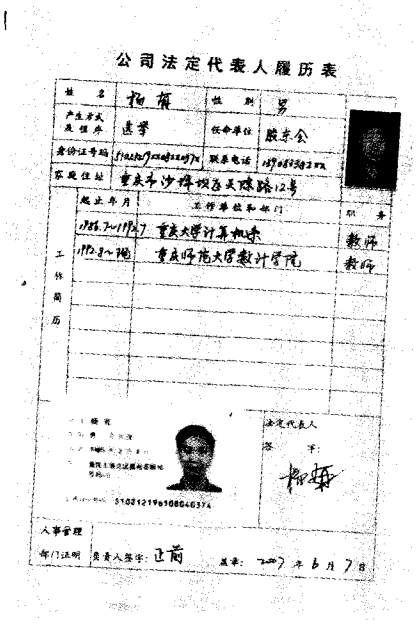

图 1　相似图像 I

2　COT 的基本思想

2.1　模板及其性质

COT 方法首先需要创建档案模板，一个模板对应一组图像，即一个相似图像集合。在该集合中，所有的图像具有某种程度的相似性，比如图像的相同区域具有类似的像素亮度，图像的直方图具有可比性，图像具有相似的边缘分布和特征分布等。在数字档案生产过程中，可以由文件名的前缀或后缀人工标识相似图像，不同的相似图像构成不同的相似图像集合，对应着不同的模板。

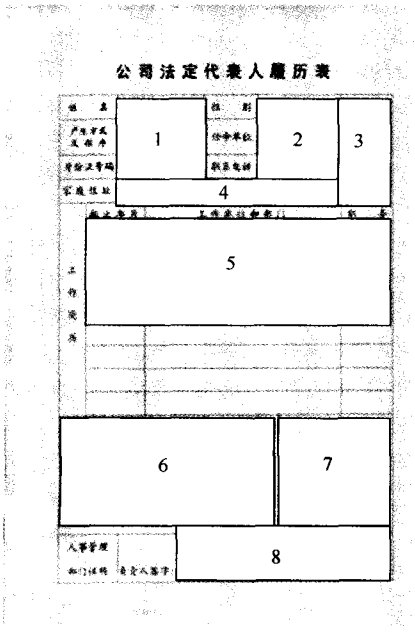

图 2　模板及其 ROI 区域

图 2 为图 1 所示图像对应的模板，其中共有 8 个 ROI 区域，即相似图像集合的变化部分，余下区域则为非 ROI 区域，对应图像集合的相似部分。由于模板和图像集合之间存在一一对应关系，因此可以通过定义模板的性质来表达各相似图像的本质特征。为此，我们定义：

1）$T \cdot A1$：模板的性质 1，表示 ROI 区域在模板图像中的相对位置。如果 ROI 区域为圆形（比如印章），则 $T \cdot A1$ 等于该区域的中心坐标；如果 ROI 区域为矩形，则等于该矩形的左上角坐标。

2）$T \cdot A2$：模板的性质 2，表示 ROI 区域的大小。如果 ROI 区域为圆形，则 $T \cdot A2$ 等于该圆的半径；如果 ROI 区域为矩形，则 $T \cdot A2$ 等于该矩形的宽度和高度。该性质与 $T \cdot A1$ 配合，可以指明相似图像中对应目标所处的大致区域，为区域分割缩小范围，降低计算量。

3）$T \cdot A3$：模板的性质 3，表示 ROI 区域的类型，其值域为{IZ（图像区域）、GZ（图形区域）、TZ（文字区域）}。该性质用于选择区域编码方法。

4）$T \cdot A4$：模板的性质 4，用于图像配准。它具有两层含义：一是表达基图像的特征信息；二是指明提取浮动图像特征集的方法，包括特征类型和特征的参考坐标。

2.2　集合冗余压缩

考虑某个模板 T，设其相似图像集合的相似性为 S，则变化性定义为 $V=1-S$。比如，如果 $S=0.6$，则平均来讲，对于每个像素位置，集合中有 60% 的图像在该位置具有相同的像素值，而余下的 40% 则具有变化性。又假设存在 $n+1$ 种符号的字母表$\{a_0, a_1, \cdots, a_n\}$，且 a_k 是具有最高频率 S 的符号，即 $P(a_k)=S, 1\leqslant k\leqslant n,$，则各种符号出现的概率之和为

$$1=\sum_{j=1}^{n} P(a_j)=S+\sum_{j\neq k} P(a_j)$$

根据 Shannon 熵的定义，我们可以得到该相似图像集合的熵 H_T 为

$$H_T=-S\log S-\sum_{j\neq k} P(a_j)\log P(a_j) \tag{1}$$

如果除字母 a_k 以外其它字母出现的概率相等，即

$$P(a_j)=V/n, j\neq k$$

则(1)式可进一步简化为

$$H_T=-S\log(S)-V\log(V/n) \tag{2}$$

式子(1)和(2)说明：随着集合相似性 S 的增加，模板 T 对应的相似图像集合的熵 H 会明显减少。从而，在档案图像数据库应用系统中，所有模板对应的图像集合的熵 $H=\sum_{\forall T} H_T$ 也会降低。

3　COT 压缩方法

COT 压缩方法主要包括图像配准、图像分割和图像压缩三个步骤，如图 3 所示。以相似图像集中的某幅图像 S 为浮动图像，以模板 T 为基图像首先进行图像配准，得到配准后的灰度图像 R；R 在模板的控制作用下，快速进行图像分割，得到相似区域 RROI 和非相似区域 ROI，ROI 又分为 IZ，GZ 和 TZ 三个类型，然后根据类型分别采取不同的区域编码策略。

3.1　图像配准

图像配准是指同一目标的两幅或两幅以上图像在空间位置上的对准。由于可以将模板视为相似图像集合中的标准图像，因此将模板作为基图像，将所有的相似图像作为浮动图像并与模板图像配准是合理的。配准过程实现了两个目的：一

是解决了相似图像的倾斜问题，更有利于图像内的冗余去除[7]；二是使得该模板对应的集合冗余达到最大，有利于降低图像集合的熵。

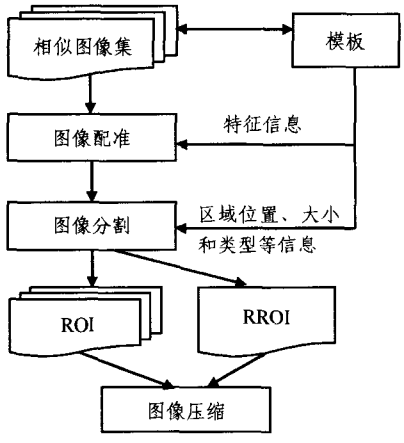

图 3　基于模板的图像压缩过程

由于基于灰度相关的图像配准能够获得较高的定位精度，但其计算量大，难以达到实时性要求，而且一旦进入信息贫乏的区域，便会导致误配准率的上升，因此基于特征的配准方法成为一种可选方案。该方案首先从相似图像中提取特征，用相似性度量和一些约束条件确定几何变换，最后将该变换作用于相似图像。配准中可以采用的特征包括边缘、轮廓、直线、兴趣点、颜色、纹理等。在 COT 压缩方法中，模板 T 创建之时就定义了模板图像的特征集合，即 $T \cdot A4$，它是已知的，无需计算；由于相似图像 I 和模板 T 具有相似性，因此根据 $T \cdot A4$ 的位置信息，可以在较小的图像区域内抽取相似图像 I 的特征，快速实现图像配准。

3.2　图像分割

图像分割是为后续工作有效进行而将图像划分为若干有意义区域的一种技术，它将图像细分为构成它的子区域或对象，分割的程度取决于要解决的问题。也就是说，在应用中，当感兴趣的对象已经被分离出来时，就停止分割。

在 COT 方法中，$T \cdot A1$ 和 $T \cdot A2$ 两个性质为我们指示了 ROI 的大致区域，因此配准后的图像 R 根据 $T \cdot A3$ 性质直接确定 ROI 的大致区域。在大致区域中，再根据区域类型分别采取具体的图像分割技术确定 ROI 的精确位置。该过程如图 4 所示。

对于 IZ 区域，我们有如下先验知识：其相应的二值图像区域通常是具有一定高度和宽度的、以黑色为主的连通方块，因此通过数学形态学中的开运算，去除图像中的微小连接、毛刺和凸出部分，就会产生明显的方块图像，从而达到分割 IZ 的目的。对于 GZ 性质的区域，我们也有如下的先验知识：在档案图像中，常见的图形主要是直线。因此 GZ 区域的提取主要是检测各个方向的直线段。Hough 变换[8]是直线检测的常用技术，HMT(Hit-Miss Transform，击中不击中变换)[9]是形状细化、识别和定位的有效工具，两者结合使用就很方便地分割出 GZ 区域。对于 TZ 性质的区域，我们也可以使用如下的先验知识：文本图像在层次方面高度结构化，这意味着同一行的符号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，涂污(smearing)技术[10]不失为一种可取的分割方案。

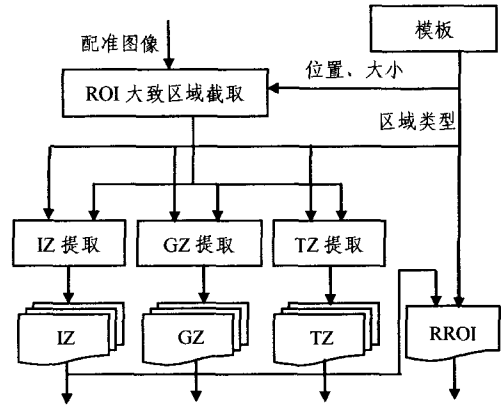

图 4　档案图像分割过程

3.3　图像压缩

图像压缩是在保证一定的图像质量和满足任务要求的条件下，减少原始图像数据量的处理过程。特定的应用应该选择特定的压缩方法，这依赖于应用的特定需求和能够使用的资源数量。

在 COT 方法中，$T \cdot A3$ 让我们对档案图像的内容有所理解，区域编码方法可以据此来决定。首先，RROI 是相似图像的信息非变化区，图像集合内所有相似图像的 RROI 区域都可以用模板的对应区域来代替，从而大幅节省存储空间。其次，对于 ROI，其三种不同性质的区域(IZ，GZ 和 TZ)决定了它们应该采取的编码方法。对于 IZ 性质的 ROI 区域，EZW(Embedded Zerotree Wavelet，嵌入式零树小波)算法是一个不错的选择，它简单有效，不需要任何训练。对于 GZ 性质的 ROI 区域，可以采用矢量描述编码。对于 TZ 性质的 ROI 区域，由于文字最终是供人阅读的，因此其轮廓(包括中空部分)十分重要。传统的二值文字图像容易出现笔画的粘连或缺失，在低分辨率情况下尤为突出。而灰度的文字图像存在过多的灰度等级的冗余信息，使其压缩率受到影响。因而采用基于灰度缩减的方法对文字图像进行编码，既可以最大限度地保持文字的轮廓形状，又可降低由图像灰度等级带来的冗余性。

3.4　图像解码

图像解码是图像编码的逆过程，相似图像的重构就是 ROI 和 RROI 两个区域的重组。由于其区域编码数据和位置参数均已存储，因此各区域数据解压后直接进行图像或运算，即可实现整个图像的解码。

4　实验与分析

以某工商档案图像应用系统为例，诸如图 2 所示的模板和图 1 所示的相似图像大量存在，我们采用 COT 方法来对其进行压缩。

首先，我们定义模板图像 T 的性质。经过对图像 T 的分析，可人工定义该模板共包含 8 个 ROI 区域，分别标记为 ROI1，ROI2，…，ROI8，其性质定义如表 1 所示。$T \cdot A1$ 的定义包括“坐标值”和“中心坐标”两项，“中心坐标”=“Y”表示该 ROI 区域为圆形，其“坐标值”代表原点位置；否则“坐标值”代表矩形框的左上角位置。$T \cdot A2$ 定义为矩形框的宽和高；$T \cdot A3$ 说明除 ROI3 和 ROI6 为图像区域 IZ 外，其它 6 个区域均为文本区域 TZ；$T \cdot A4$ 定义为 4 个控制点的坐标值，

(下转第 271 页)

参 考 文 献

[1] Tang Y Y, Sune C Y, Yan C D, et al. Financial document processing based on stuff line and description language [J]. IEEE Transactions on Systems, Man and Cybernetics, 1995, 25(5)

[2] 通用表格 SDK [Z]. http://www.wintone.com.cn/prod/10/detail200.aspx

[3] 张猛,余仲秋,姚绍文. 手写体数字识别中图像预处理的研究. 微计算机信息,2006,22(6-1)

[4] Trier ΦD, Taxt T. Evaluation of binarization methods for document images [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1995, 17(3): 312-315

[5] Lam S W, Javanbakht S W, Srihari S N. Anatomy of a form reader [C]// Proc. Second Int'l Conf. Document Anal Recog. Tsukuba, Japan: [s. n.], 1993: 506-509

[6] Wang X R. Two adjacent angles dot method in the form recognition [C]. Wavelet Active Media Technology and Information Processing, 2006: 251-258

[7] Chen G Y, Bui T D, Krzyzal. Contour-based Handwritten Numeral Recognition Using Multi Wavelets and Neural Networks [J]. Pattern Recognition, 2003, 36(7): 1579-1604

[8] Malaviya A, Peters L. Fuzzy Feature Description of Handwriting Patterns [J]. Pattern Recognition, 2007, 30(10): 1591-1604

[9] Daijin K, Bang Sungyang. A Handwritten Numeral Character Classification Using Tolerant Rough Set [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(9): 923-937

[10] Zhang R, Ding X. Offline Handwritten Numeral Recognition Using Orthogonal Gaussian Mixture Model [C]. IEEE International Conference on Image Processing, 2001, 1(1): 1126-1129

[11] Hu J, Yan Y. Structural Primitive Extraction and Coding for Handwritten Numeral Recognition [J]. Pattern Recognition, 1998, 31(5)

[12] 吴志勇,蔡莲红. 语音合成技术的原理 [Z]. http://www.ctiforum.com/technology/tts/tts0301.htm

[13] 李强,贾云霞. Visual C++项目开发实践. 中国铁道出版社[M] ISBN: 7-113-05381-5/TP·993, 2003: 447-455

(上接第 267 页)

代表表格外框的 4 个角,用于图像配准。

表 1 模板图像的性质示例

区域	T·A1 坐标值	T·A1 中心坐标	T·A2	T·A3	T·A4
ROI1	(502,635)	N	(500,465)	TZ	(502,305)
ROI2	(500,1434)	N	(476,465)	TZ	(3220,305)
ROI3	(499,1910)	N	(295,598)	IZ	(499,2203)
⋮	⋮	⋮	⋮	⋮	(3217,2208)
ROI8	(2940,983)	N	(1225,277)	TZ	

其次,我们进行图像配准过程。由于 T 的特征集合已经定义,可以视为已知,因此问题的关键就是提取 I 中的对应特征,然后选择几何变换,并将该变换作用于 I。在实验中,我们先将 I 二值化[6],得到二值图像 B,然后分别以 $T \cdot A4$ 的 4 个坐标点为中心,在一定范围内(B 的子图像)提取 I 的特征集合。考虑到相似图像的失真主要来自扫描仪,因此配准过程中使用投影变换来纠正这种失真。

随后,在 ROI 大致区域内,根据 $T \cdot A3$ 性质分割出精确的三类区域。比如,对于 ROI3,通过其 $T \cdot A1$ 和 $T \cdot A2$,按照下式计算其大致区域的左上角坐标(x_{ul}, y_{ul})和右下角坐标(x_{br}, y_{br}):

$$\begin{cases} x_{ul} = TA1(1) \pm TA2(2)/2 \\ y_{ul} = TA1(2) \pm TA2(1)/2 \end{cases}$$

由此得到比 T 定义的 ROI3 更大的区域,以保证 IZ 区域能够从此大致区域内分割出来。由于该大致区域是图像 R 面积的 1/12,从而大幅降低了区域分割的计算量。

表 2 COT 方法的压缩性能比较

压缩方法	文件大小	压缩率	PSNR
JPEG	461580B	18.73	27.40
小波变换	426302B	20.28	30.11
COT 方法	135755B(ROI)	63.68	33.45

最后,我们对相似图像进行压缩,其实验结果如表 2 所示。其中原始图像为 300dpi 扫描的灰度图像,尺寸为 3503×2468,占用字节数是 8645404B;分割后的 IZ 区域采用改进后的 EZW 技术,即 MTZ(Multi-Threshold Zerotree,多阈值零树)小波编码算法[11]进行压缩,TZ 区域采用灰度缩减方法进行压缩。值得说明的是,表中 COT 方法显示的压缩结果(文件大小和压缩比)只包含 ROI 区域的数据,没有包括非 ROI 区域,这是因为整个相似图像集合只需要一个模板图像,由模板定义的 RROI 区域占用的空间可以忽略不计。可以看出,COT 方法的压缩性能得到了大幅提高。

结束语 通过模板的定义,充分挖掘图像内和图像间的数据冗余,开辟了图像二维压缩的有效途径,大幅降低了图像数据的存储空间,为大容量图像数据的应用系统提供技术支持。在 COT 方法当中,如何快速进行图像配准和图像分割是值得进一步研究的课题。针对相似图像中的同类 ROI 区域(比如示例中的照片和身份证区域),研究更加高效的编码方法也十分必要。

特定的应用应该选择特定的压缩方法。如果数据库中存在包括法律和历史档案的图像,则不允许有任何视觉信息的损失,因此压缩方案必须选择无损的。如果在基于网络传输的系统中,受到诸如带宽、功率和物理存储器的限制,则只需要保证文本可读性,使用有损压缩来提高压缩比。COT 方法提供了有损和无损自由选择的灵活度。

参 考 文 献

[1] Ye Yan. Text Image Compression Based on Pattern Matching [D]. University of California, 2002

[2] Kia O E, Doermann D S, Rosenfeld A, et al. Symbolic Compression and Processing of Document Images [J]. Computer Vision and Image Understanding, 1998, 70(3): 335-349

[3] Karadimitriou K, Fenstermacher M. Image compression in medical image databases using set redundancy [J]. IEEE Proceedings of Data Compression Conference, DCC '97. March 1997: 445

[4] Karadimitriou K, Tyler J M. The centroid method for compressing sets of similar images [J]. Pattern Recognition Letters, 1998, 19: 585-593

[5] Lee Jiann-Der, Wan Shu-Yen, Ma Cherng-Min, et al. Compression Sets of Similar Images Using Hybrid Compression Model. Proceedings [J]// 2002 IEEE International Conference on Multimedia and Expo, 2002. ICME '02. Vol 1. Aug. 2002: 617-620

[6] 杨有,尚晋. 一种政府资源档案图像的二值化方法[J]. 计算机科学,2007,34(3):227-229

[7] Inglis S J. Lossless Document Image Compression [D]. New Zealand: University of Waikato, 1999

[8] Hough P V C. Method and Means for Recognizing Complex Patterns [P]. US Patent 3,069,654, Dec. 1962

[9] Gonzalez R C, Woods R E. Digital Image Processing Second Edition [M]. Beijing: Publishing House of Electronics Industry, 2006: 532-534

[10] Wong I Y, Casey R G, Wahl E M. Document Analysis System. IBM J Research Develop, 1982, 26(6): 647-656

[11] 杨波,汪同庆,叶俊勇. 带噪图像的多阈值零树编码方法[J]. 光电工程,2004,31(3):60-63