基于模板差分的档案图像集合冗余压缩研究

余 平[1] 杨 有[1] 尚 晋[2]
(重庆师范大学计算机与信息科学学院 重庆 401331)[1]
(重庆航天职业技术学院计算机工程系 重庆 400021)[2]

摘 要 档案图像信息系统中,页间信息冗余大量存在,对基于页间信息统计特性的压缩方法进行研究具有重要意义。集合冗余压缩正是利用图像之间的相似性降低整个图像集合的熵。基于模板差分的压缩方法是一种改进的集合冗余压缩技术,它通过模板建立相似档案图像集合,通过在最小-最大差分方法的编解码器中加入模板图像,来提高页间压缩性能。理论分析表明,模板差分压缩方法的压缩性能高于最小-最大差分方法。实验结果表明,模板差分方法和最小-最大差分方法均能有效提高图像集合的压缩比,而且模板差分方法比最小-最大差分方法更有利于提高压缩比。

关键词 档案图像信息系统,集合冗余,图像压缩,模板差分
中图法分类号 TP391.4 **文献标识码** A

Research on the Set Redundancy Compression of Document Image Based on Template Difference

YU Ping[1] YANG You[1] SHANG Jin[2]
(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)[1]
(Computer Engineering Department, Chongqing Aerospace Polytechnic College, Chongqing 400021, China)[2]

Abstract Much redundant information between document pages exists in the document image information system. The research on the compression method based on the page-page statistical features is significant. Set Redundancy Compression(SRC) is such a technique. It reduced the total entropy of the whole image set through utilizing the image page's similarity. Compression based on template differential(CTD) is an improved SRC. The similar image set was constructed by the template. The coding performance was improved by adding the template image to the Min-Max Differential (MMD) coding/decoding model. It was proved theoretically that CTD's coding performance is higher than MMD's. It was demonstrated by experiments that both the CTD and MMD are benefit to increase the compression ratio of image set, furthermore, and CTD is better than MMD.

Keywords Document image information system, Set redundancy, Image compression, Template differential

当前,越来越多的资源档案(Resource Document)正以数字形态在网络上被利用,这种利用不仅跨越了时空的限制,使档案信息获得了独立与自由,而且节省了成本,便于保密。但是纸质档案数码化需要图像处理技术的支持,支持的程度关系到以图像为主要内容的DIIS(Document Image Information System,档案图像信息系统)的生命。在图像处理技术中,图像压缩对DIIS最为重要,它的作用主要体现在如下两个方面:图像压缩的质量关系到信息的第一属性,即事实性,它决定了系统的利用价值;图像压缩可以减少存储空间,一方面节省存储成本,另一方面有利于提高传输效率,使DIIS的效能得到提高。

在DIIS应用中,图像压缩不仅要考虑单幅档案图像的冗余,还要考虑档案图像之间的冗余,即用图像的集合统计特性来代替单幅图像统计特性,从而降低整个图像集合的熵。比如,在工商企业登记档案、国土房产资源档案等政府资源类数字档案应用系统中,一些申请书和登记表的内容都具有相似性,各户对应档案页基本上都是格式文本,大部分区域的内容相同,只是在某些极小的区域有人工填写的信息,页与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[1]为例,该市共有约20万户企业,每户档案的平均页数为80页,每户企业都有相似的档案页,如公司设立登记申请书、公司法定代表人履历表、公司章程等,不完全统计表明可以认定相似档案的比例高达40%左右,即可以使用集合冗余压缩(Set Redundancy Compression,SRC)技术进行处理的档案页数量高达640万页。

SRC技术是由美国Louisiana国立大学的Karadimitriou K.等人于1996年首先提出的,他们针对医学图像数据,定义了相似图像(similar image)和集合冗余(set redundancy)的基本概念,并提出了提取集合冗余的3种方法:MMD(Min-Max Differential,最小-最大差分法)方法[2,3]、MMP方法(Min-

到稿日期:2010-10-26 返修日期:2011-01-26 本文受重庆市教委科研项目(KJ100623),重庆市科委科研项目(CSTC,2009BB2389)和重庆师范大学博士基金(10XLB006)资助。

余 平(1975-),女,硕士,副教授,主要研究方向为数据挖掘和图像处理,E-mail:40465742@qq.com;杨 有(1965-),男,博士,副教授,主要研究方向为数字图像处理;尚 晋(1964-),女,硕士,副教授,主要研究方向为数字图像处理和信息安全。

Max Predictive，最小-最大预测法)[2,4]和质心方法(Centroid Method)[5]。这3种方法均利用图像间的像素统计特性来降低整个集合的熵，而且随着图像间相似性的增加，集合的熵会降低。

随后的时间里，少数学者对SRC技术进行了进一步的研究。2002年，Jiann-Der Lee等人[6]提出混合压缩模型(Hybrid Compression Model，HCM)，其利用区域生长技术将医学图像分块，然后再应用质心方法降低集合冗余，实现了比常规质心方法高5.6%～134.9%的压缩比。2006年，Ait-Auodia S等人[7]使用CT和MR医学图像集合，对MMD，MMP和质心方法等SRC技术的性能进行了对比分析，实验结果表明：在使用诸如Bzip2，Gzip，Huffman，RAR和ZIP等编码技术之前，使用SRC技术比不使用SRC技术可以大幅提高压缩比，其中SRC技术对Huffman编码的压缩比改进最大，对Bzip编码的压缩比改进最小。2006年，Nielsen C等人[8]利用最小扩张树(Minimum Spanning Tree，MST)这个数据结构来组织和压缩相似图像集合，有力地支持了基于内容的图像检索。

但是，很难检索到SRC技术在DIIS中应用的参考文献。一方面，DIIS中存在大量的集合冗余，另一方面SRC又可以大幅降低集合冗余，因此，将SRC技术用于相似档案图像集合成为可能。值得注意的是，医学相似图像和扫描档案图像的获取方式存在差别，前者是单模图像，后者是多模图像。由于纸质档案数码化过程的批量生产特殊性，使得相似档案图像集合的建立较医学图像更为困难。限于篇幅的限制，本文只讨论SRC技术在降低相似档案图像集合方面所起的作用，并不强调相似档案图像的建立过程。

本文第1节对SRC技术进行介绍，说明SRC编解码流程，阐述SRC降低集合冗余的理论基础，并介绍后续章节将要使用的MMD方法；第2节介绍DIIS中提取相似档案图像集合的方法，并分析DIIS中存在的集合冗余；第3节针对相似档案图像集合，阐述一种基于模板差分的集合冗余压缩方法；第4节为实验，将测试第3节提出的方法，并对比分析它的性能；最后得出结论。

1 集合冗余压缩技术

所谓集合冗余压缩(Set Redundancy Compression)就是降低集合冗余的方法。Karadimitriou K等人给出了这种技术的编解码模型[2]，如图1所示。整个压缩过程包括集合冗余提取和单页图像压缩两个步骤，第一个步骤使单页相似图像从相似图像集合中完成解相关操作，第二个步骤是采用任意压缩方式对单页图像进行压缩。该模型包含两个方面的含义，一是这两个步骤是独立的，第一个步骤使用的任何一种SRC技术都不会影响到第二个步骤的单页压缩方法，因此从这个意义上讲，可将第一个步骤视为第二个步骤的预处理过程；二是两个步骤降低了两种冗余，它们分别是图像间和图像内的冗余，因此这种基于SRC技术的压缩模型比一般的压缩模型增加了一种压缩途径。

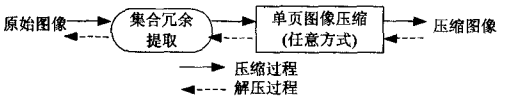

图1 集合冗余压缩技术编解码模型

SRC技术之所以能够降低集合冗余，是因为它根据图像间的统计特性，降低了整个图像集合的熵。设存在$n+1$种符号的字母表$\{a_0,a_1,\cdots,a_n\}$，a_k是具有频率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对熵的定义，可以得到该相似图像集合的熵H_T为

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

为简化计算，可假设除字母a_k以外其它字母出现的概率相等，即

$$P(a_j)=\frac{V}{n},\ j\neq k$$

则式(1)可进一步简化为

$$H_T=-S\log S-V\log\frac{V}{n} \tag{2}$$

式(1)和式(2)说明：随着集合相似性S的增加，相似图像集合的熵H_T会明显减少。

Karadimitriou K等人提出的3种SRC具体方法中使用了“最小图像”、“最大图像”和“平均图像”的概念，它们代表了图像间的统计特性，分别由相似图像集合中每个像素位置的最小值、最大值和平均值组成。基于这3类图像，SRC技术通过集合映射操作缩小相似图像的灰度值动态范围，减小每幅图像内像素分布的方差，从而降低每个像素点编码所需要的比特数，实现页间压缩的目的。

其中MMD方法的编解码公式定义为[3]

$$D_x^{MMD}=\begin{cases}value(P_x)-\min_x, & \text{if } (value(P_x)-\min_x)\\ & <(\max_x-value(P_x))\\ \max_x-value(P_x), & \text{otherwise}\end{cases} \tag{3}$$

$$value(P_x)=\begin{cases}D_x^{MMD}+\min_x, & \text{if } (value(P_x)-\min_x)\\ & <(\max_x-value(P_x))\\ \max_x-D_x^{MMD}, & \text{otherwise}\end{cases} \tag{4}$$

式中，x代表像素点的位置，$value(P_x)$代表像素点x处的灰度值，D_x^{MMD}为需要存储到差值图像中的差值，$\min_x$为最小图像中位置x的值，$\max_x$为最大图像中位置x的值。

2 基于模板差分的页间压缩方法

2.1 档案图像信息系统中的集合冗余

首先，DIIS中图像之间的相似性是大量存在的。在DIIS中，1)“户”是信息组织的基本单位，它由许多档案页组成。比如，在某市工商档案信息系统中[1]，档案页共分为“企业证照”、“董事资料”、“批文”和“处罚决定”等18个类别。2)不同“户”的相同属性档案页之间具有相似性，这种相似性是大量存在的，是集合冗余压缩的基础。比如第i户和第j户的“企业证照”档案页，不同之处在于名称、编号和地址等少量信息，大量背景信息是相同的。3)相同“户”的类似档案页之间也具有相似性，具有这种相似性的档案页难以形成相似档案图像集合，不利于SRC技术的应用。

其次，对于不同“户”的相似档案页，可以通过模板建立DIIS的相似图像集合[9]。模板是不同“户”相似档案页中的参考页或标准页，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。

最后，通过 SRC 技术实现 DIIS 中图像存储空间的降低。每个模板对应的相似档案图像集合，在像素统计特性上具备了 SRC 技术处理的理论基础，基于 SRC 的压缩会降低整个图像集合的熵，而且图像间的相似性越高，图像集合的熵越小。

2.2 基于模板差分的页间压缩

基于模板差分压缩（Compressed Based on Template Difference，CTD）的思想来源于灰度等级缩减方案，图像像素的灰度等级越低，所需编码的比特数就越少。MMD 方法将处理图像像素的灰度值与最大图像和最小图像对应像素值进行比较，以实现待处理图像像素灰度值的减小，从而减少编码比特数。而 CTD 方法不仅要与 MMD 方法进行比较，而且还要将待处理图像像素的灰度值与模板图像对应像素值进行比较，以进一步降低待处理图像的灰度值，实现更高的压缩比。

基于上述思想，CTD 方法的编码器定义为：

$$D_x^{CTD}=\min\{\max_x-value(P_x),value(P_x)-\min_x,|value(T_x)-value(P_x)|\} \tag{5}$$

解码器定义为

$$value(P_x)=\begin{cases}\max_x-D_x^{CTD}, & \max_x-value(P_x)=D_x^{CTD}\\ \min_x+D_x^{CTD}, & value(P_x)-\min_x=D_x^{CTD}\\ value(T_x)\pm D_x^{CTD}, & \text{otherwise}\end{cases} \tag{6}$$

式中，x 表示像素位置，D_x^{CTD} 代表经 CTD 方法处理需要存储到差值图像中的差值，$value(P_x)$ 和 $value(T_x)$ 分别代表相似档案图像和模板图像在位置 x 处的像素灰度值，$\min_x$ 为 MIN 图像中位置 x 处的像素灰度值，$\max_x$ 为 MAX 图像中位置 x 处的像素灰度值。在式(6)中，当 $value(T_x)>value(P_x)$ 时，取减号“－”，否则取加号“＋”。

由于模板档案图像也是某相似档案图像集合中的一员，因此它的像素灰度值介于 MIN 图像和 MAX 图像之间。将 MIN 图像、MAX 图像、模板图像和待处理的相似档案图像以一维数组的方式绘制出如图 2 所示的灰度分布图，其横坐标为像素位置，纵坐标为像素灰度值。从图中可以直观看到，待处理的相似图像与模板图像似乎距离更近时，其像素灰度差值更小。

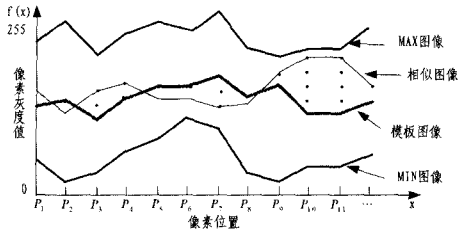

图 2 基于模板差分的集合冗余压缩方法

上述直观猜测可以从如下的推理中获得验证。因为 MMD 技术中需要存储的差值为[9]

$$D_x^{MMD}=\begin{cases}value(P_x)-\min_x, & \text{if } (value(P_x)-\min_x)<(\max_x-value(P_x))\\ \max_x-value(P_x), & \text{otherwise}\end{cases} \tag{7}$$

而从图 2 可知，在任意像素位置都有

$$\min_x\leqslant value(T_x)\leqslant\max_x$$

所以恒有

$$D_x^{CTD}\leqslant D_x^{MMD}$$

即 CTD 方法存储的图像差值比 MMD 小，理论上其编码性能高于 MMD 技术。

从算法的时间复杂性来讲，CTD 方法和 MMD 方法属于同一级别，计算 MIN 图像和 MAX 图像的时间复杂性为 $O(N\log N)$，编解码的时间复杂性为 $O(M)$，其中 M 和 N 分别表示 SSDI 中相似图像的数量和相似图像的大小。但由于 CTD 包含计算最小差分的最小值运算，而 MMD 方法只包含比较操作，因此 MMD 方法运算时间少于 CTD 方法。

3 实验与分析

以某市的工商企业登记档案图像信息系统[1]为例，选用具有相同档案页属性的不同企业登记户的对应档案页为一个相似图像集合，即

$$T=\{P_i\mid P_i\in DIIS\wedge[P_i]=\text{"}\times\times\text{"}\}$$

式中，T 代表一个相似图像集合，对应着一个模板；P_i 和 $[P_i]$ 分别代表 DIIS 中的某页档案及其属性值。在 T 中，P_i 可能来自不同规模的登记企业，也可能来自不同地域的登记企业，还可能来自不同时间的登记企业。

采用 CTD 方法，对某 T 中的 10 幅相似图像进行压缩，其结果如表 1 所列。表中体现了两种压缩效果，分别是 $PSNR=33.9115$dB 和 $PSNR=29.5835$dB，它们对应着低压缩比和高压缩比。高压缩比下，每页档案图像的空间占用不超过 25kB，这是一种实际应用的要求。从表 1 可以看出：在 PSNR 值不变的情况下，采用集合冗余压缩（SRC）技术比不采用 SRC 技术能够大幅提高压缩比；而且使用 CTD 的 SRC 技术比使用 MMD 的 SRC 技术更有利于提高压缩比。

表 1 基于 CTD 压缩方法的性能

压缩方法	相似图像平均大小（Byte）	平均压缩比	压缩比改进 RI(%)	平均 PSNR（dB）
未压缩	959274	—	—	—
JPEG2000	95927	10.00	—	33.9115
MMD+JPEG2000	84073	11.410	14.10	33.9115
CTD+JPEG2000	82200	11.670	16.70	33.9115
JPEG2000	25600	37.47	—	29.5835
MMD+JPEG2000	22786	42.098	12.35	29.5835
CTD+JPEG2000	22419	42.787	14.19	29.5835

表 1 中的压缩比改进定义如下，

$$R=\frac{R_{SRC}-R}{R} \tag{8}$$

式中，R 为仅仅使用标准压缩模型所获得的压缩比，而 R_{SRC} 为结合使用 SRC 和标准压缩模型所实现的压缩率。

当 PSNR 值为 33.9115dB 时，基于模板压缩方法的效果如图 3(b)所示。与单纯的 JPEG2000 压缩方法相比，如果放大两幅图中的细节进行观察，没有发现两幅图像存在明显的视觉差异。当 PSNR 值增大到 29.5835dB 时，此时正对应着 DIIS 对每页档案平均占用空间为 25kB 的要求，无论是 JPEG2000 方法，还是使用了 SRC 技术的方法，放大压缩图像后均会观察到明显失真。

图 3　基于 CTD 的相似图像压缩效果(PSNR=33.9115)

结束语　集合冗余压缩(SRC)技术实现图像压缩不仅要利用页内信息的统计特性,而且要利用页间信息的统计特性,图像相似性越大,SRC 实现的压缩比越高。基于模板差分的压缩(CTD)方法依赖模板建立相似档案图像集合,是一种改进的最小-最大差分(MMD)方法,它比 MMD 方法具有更佳的压缩性能,适合在档案图像信息系统中使用。

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。另一方面,定量评测一个档案图像压缩算法是否优秀是非常困难的,因为在档案图像信息系统中,某些评测参数不仅仅要有利于传统的空间减少,还要有利于图像的处理、图像的检索和图像数据的传输。

参 考 文 献

[1] 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

[2] Karadimitriou K. Set redundancy, the enhanced compression model, andmethods for compressing sets of similar images[D]. Department of Computer Science, Louisiana State University, Baton Rouge, La, USA, August 1996

[3] Karadimitriou K, Tyler J M. The min-max differential method for large-scale storage and compression of medical images[C]// Proceedings of of Annual Molecular Biology and Biotechnology Conference. Baton Rouge, La, USA, 1996

[4] Karadimitriou K, Tyler J M. Min-Max Compression Methods for medical image databases[J]. ACM SIGMOD Record, 1997, 26(1):47-52

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

[6] Lee J-D, Wan S-Y, Ma C-M, et al. Compression Sets of Similar Images Using Hybrid Compression Model[C]// Proceedings of 2002 IEEE International Conference on Multimedia and Expo (ICME'02). Lausanne, Switzerland. Aug. 2002, 1:617-620

[7] Ait-Auodia S, Gabis A. A comparison of set redundancy compression techniques[J]. EURASIP Journal on Applied Signal Processing, 2006:1-13

[8] Nielsen C, Xiaobo L. MST for lossy compression of image sets [C]// Proceedings of the Data Compression Conference(DCC'06). Vienna, Austria, March 2006:463

[9] 杨有. 一种基于模板的档案图像压缩新方法[J]. 计算机科学,2008,35(6):265-267

[10] 梁枫,王平. 基于角点特征的高精度图像配准算法[J]. 重庆理工大学学报:自然科学版,2010,24(2),87-90

(上接第 282 页)

配准结果;图 6(c)是将图 6(a)、图 4(b)和图 6(b)作为 R,G,B 通道合成的 BMP 图像;图 6(d)是图 4(a)(b)(c)融合得到的 BMP 图像。表 1 给出本文精匹配方法与文献[11]的比较,从数据可以看出,采用自适应阈值的空间距离约束方法与文献[11]的固定阈值方法相比,能够得到更高的配准精度、更好的配准效果。

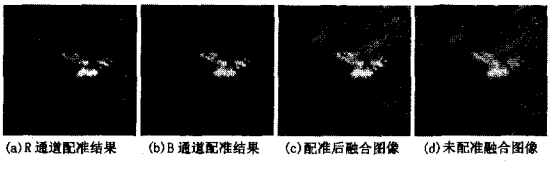

图 6　配准结果网格及特征点

表 1　本文方法与文献[11]比较

待配准图像	运行时间(s)	RMSE	
		文献[11]方法	本文方法
图 4(a)	0.22	0.58	0.4
图 4(c)	0.2	0.38	0.27

结束语　本文提出一种基于特征点的多光谱遥感图像自动配准算法。其通过特征网格选取步骤,减小计算量,同时确保了特征点的均匀分布。在粗匹配步骤中,依据多光谱图像的特点构造搜索窗口,提高了算法效率及初匹配的正确率。改进了精匹配的方法,提出了自适应阈值的空间距离约束方法,避免了人工阈值选取,使算法具有通用性。实验结果表明,该方法能够实现多光谱遥感图像的自动、快速、亚像素级配准。

参 考 文 献

[1] Zitova B, Flusser J. Image registration methods: a survey[J]. Image and Vision Computing, 2003, 21(11):977-1000

[2] 谌安军,陈炜,毛士艺. 一种基于边缘的图像配准方法[J]. 电子与信息学报,2004,26(5):679-684

[3] Pluim J P W, Antoine J B, Viergever M A. Mutual Information based Registration of Medical Images: A Survey [J]. IEEE Transactions on Medical Imaging, 2003, 22(8):986-1004

[4] Imam, Yetik S. A novel non-rigid registration method[C]// IEEE Southwest Symposium on Image Analysis & Interpretation(SSIAI2010). Austin, TX, 2010:45-48

[5] Lin Hui, Du Pei-jun, Zhao Wei-chang, et al. Image Registration Based on Corner Detection And Affine Transformation[C]//3rd International Congress on Image and Signal Processing (CISP2010). Yantai, China, 2010:2184-2188

[6] Trias-Sanz R, Pierrot-Deseilligny M, Louchet J, et al. Methods for Fine Registration of Cadastre Graphs to Images[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2007, 29(11):1990-2000

[7] Wang Wei-an, Liu Yi, Zheng Bo, et al. A Method of Shape Based Multi-Sensor Image Registration [C] // 2009 Urban Remote Sensing Joint Event. Shanghai, 2009:1-5

[8] Eastman R D, Moigne J L, Netanyahu N S. Research issues in image registration for remote sensing[C]//IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA, 2007:1-8

[9] 夏德深,傅德胜. 计算机图像处理及应用[M]. 南京:东南大学出版社,2004

[10] 郑明玲,刘衡竹. 遥感图像配准中特征点选择的高性能算法研究及其实现[J]. 计算机学报,2004,27(9):1284-1289

[11] 胡明昊,任明武,杨静宇. 一种快速实用的特征点匹配算法[J]. 计算机工程,2004,30(9):31-33

[12] Yu Le, Zhang Deng-rong, Holden E-J. A Fast and Fully Automatic Registration Approach Based on Point[J]. Computers & Geosciences, 2008, 34(7):838-848