ISSN 1000-0054 清华大学学报（自然科学版）2006年 第46卷 第7期 17/41
CN 11-2223/N JTsinghua Univ(Sci&Tech), 2006, Vol.46, No.7 1247-1249,1253

基于OCR信息的JBIG2编码算法

尚俊卿， 刘长松， 丁晓青

（清华大学 电子工程系，智能技术与系统国家重点实验室，北京 100084）

摘 要：二值图像编码在文本存储、图象检索中有广泛的应用。为了提高二值图像的压缩比，提出了一种利用OCR结果的JBIG2(joint bi-level image group)编码算法。它在对二值文本图像进行基于模式匹配的压缩时，利用了OCR识别结果和识别置信度的信息，从而更好地完成了字模重建和模式匹配的处理，提高了JBIG2算法的性能。图像中所有识别结果可信的字符被重建字模代替，编码器只需编码字符的位置。实验结果表明：该算法优于以往JBIG2算法的效果，它可以获得高于以往有损压缩算法的图像质量，并在实验图像上得到高于以往无损压缩算法14.3%的压缩比。

关键词：模式识别；二值图像编码；文本图像压缩；OCR；模式匹配

中图分类号：TP 391.4 **文献标识码：A**

文章编号：1000-0054(2006)07-1247-03

Lossy JBIG2 based on optical character recognition

SHANG Junqing, LIU Changsong, DING Xiaoqing

(State Key Laboratory of Intelligent Technology and System, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Bi-level image coding is useful for document storage and archiving, image searches on the Internet and digital libraries. The JBIG2 (joint bi-level image group) standard for lossless and lossy coding of bi-level images is a very flexible encoding strategy which allows researchers to design their own encoders. OCR processing of text images is one encoding technique that gives measurable recognition and the confidence results. We propose a lossy JBIG2 encoding method which uses OCR processing results to improve text image compression based on pattern matching. All the credible recognized characters in the image are replaced by representative character images so that the encoder only needs to mark the positions of these characters. Experiment results show that this method gives better results than previous JBIG2 encoding methods with 14.3% less storage compared to previous lossless methods while preserving relatively good text image quality.

Key words: pattern recognition; bi-level image coding; text image compression; OCR; pattern matching

二值图像是图像中很重要的一类，它在传真、文件资料的数字化存储、数字图书馆的建设和图文的www检索等领域中都有广泛的应用，因而如何存储二值图像就成了研究人员广泛关注的研究课题。为了对二值图像进行有效的压缩，JBIG (joint bi-level image group)委员会在1999年提出了二值图像压缩标准JBIG2[1]。JBIG2将二值图像分为文本区域，半调区域和普通区域，并且采用不同的方法对不同区域进行压缩。对二值文本区域，它采用了基于模式匹配的压缩方法。JBIG2的一个重要的特点是它只定义了一个明确的码流格式和不太严格的解码流程；它并没有定义一个明确的编码器，编码的方法和流程是开放的，允许设计者灵活掌握。因此不同的编码器会有不同的复杂度、速度和压缩性能，相同的原始图像会产生不同的压缩码流。

基于JBIG2标准的这个特点，设计了一种利用OCR (optical character recognition)识别信息的JBIG2编码算法。该方法处理的对象是二值图像的文本区域，它利用OCR识别结果和该识别结果置信度的信息进行字模重建，从而更好地完成了模式匹配的处理。本文主要讨论对其中的文本图像的处理，而对其他二值图像区域采用算术编码的方法进行压缩。实验表明，与以往的JBIG2编码算法相比，该方法既能得到较高的压缩比，又能尽量减少“替换错误”，使解压后的图像仍有比较好的视觉质量。

1 JBIG2压缩算法

JBIG2将二值图像分为不同的区域进行处理。文本图像是一类很重要并且很特别的图像，其中包

收稿日期：2005-05-12

基金项目：国家自然科学基金资助项目 (60472002)

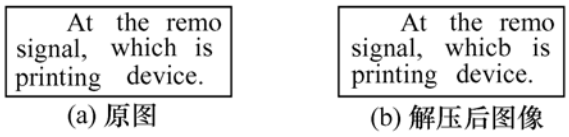
作者简介：尚俊卿(1981-)，男(汉)，山西，硕士研究生。

通讯联系人：刘长松，副教授，

E-mail: lcs@ocrserv.ee.tsinghua.edu.cn

含大量含有文字信息的字符图像,有着非常好的规律性和结构特征。通常一幅图像中的这些字符大小相近,按行或按列排列,同一字符经常在图像的不同位置重复出现多次。在字符层上,文本图像有很大的冗余度。利用这种字符层的冗余度,可以对文本图像进行有效的压缩。这种对文本图像的压缩方法被称为基于"模式匹配"的压缩。这种思想最先由 Ascher 和 Nagy 提出[2],之后又有了很大的发展[3,4]。对每一个字符,人们认为是一种"模式",可以利用一个"字典"记录这个字符,不同的字符在字典中利用不同的索引进行标记。当以后的文本图像再次出现类似的字符图像的时候,人们只需要在字典中找到相应的索引项,并对位置进行编码即可,而不需要重复对字符图像进行编码。

以往的 JBIG2 文本图像压缩算法,通常可以分为无损和有损两种。无损压缩的代表是 Howard 在 1996 年提出的 SPM (soft pattern matching) 方法[3]。无损压缩可以实现对原始图像无失真的重建;但是无损压缩通常可以达到的压缩比低于有损压缩。有损压缩的代表是 PM&S (pattern matching and substitution) 算法[2]。有损压缩通常可以达到高于无损压缩的压缩比;但是,在有损压缩时,当两个不同的字符比较相近的时候,容易引起"替换误差"。如图 1 所示,由于 h 和 b 的相似性,在有损压缩时错误地利用字符"b"替换了字符"h"。而由于人眼对字符的敏感性,这样的误差是很严重的。



(a) 原图 (b) 解压后图像

图 1 有损压缩引起的"替换误差"

2 基于 OCR 信息的 JBIG2 编码算法

以往的文本图像压缩方法多数没有用到 OCR 信息,利用 OCR 速度会比较慢,而且消耗的内存资源比较多。但是现在的硬件性能已经有了很大的提高,因此 OCR 带来的负担是可以承受的。此外,OCR 可以识别的文种是有限的,本研究主要处理的图像是中文和拉丁字母文本图像,均是可以给出识别结果的图像。

由于以往无损和有损 JBIG2 文本压缩算法存在的问题,本研究引入了 OCR 信息对二值文本图像进行有损压缩。对文本图像进行 OCR 处理可以得到很多辅助信息,比如文字的识别结果,识别置信度等。由 OCR 结果可以对文本中的相似字符进行字模重建,重建的结果就可以作为代表字符图像。许多代表字符图像可以组成字符字典,然后根据 JBIG2 标准生成压缩码流。对于文本区域的处理流程见图 2,其中字模重建是本研究算法的核心。

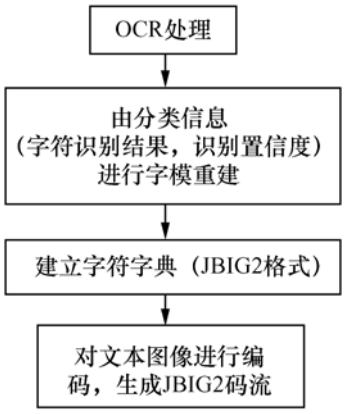


图 2 文本图像的处理流程

在处理一幅文本图像时,可能会遇到很多在一定标准下相似的字符。如果可以用同一个字符图像代表所有相似的字符,压缩比无疑会得到很大的提高。如何确定字符之间的相似度,以及如何根据相似的字符生成一个字模的过程,就是字模重建的问题。文本图像经过 TH-OCR (清华 OCR) 系统处理,可以得到文本区域每一个字的识别结果,同时也可以给出该识别结果的置信度[5]。本研究利用了这两种 OCR 结果进行字模重建。首先,利用 OCR 识别结果和字符的大小对字符图像进行聚类,所有相同大小且识别结果相同的字符图像聚为一类。然后,判断字符识别的置信度,将所有可信的字符进行平均,得到该类的代表字模。对于不可信的字符,比较它们与已重建字模的匹配程度,当匹配误差低于一定阈值时,就认为该字符可以用代表字符替换;当匹配误差高于一定阈值时,就认为该字符不能用代表字符表示。经过上述处理后,就可以实现符合 JBIG2 码流的有损图像压缩。该算法流程图如图 3 所示。

可见本文算法的性能与 OCR 识别结果的准确性有很大的关系。采用的是 TH-OCR 系统,而在置信度估计的时候,只有置信度很高的识别结果才被认为是可信的。

在比较新的字符图像与已有重建字模图像的差异时,可以有不同的选择匹配的方案[6]。本研究采用的方法是 WXOR (weighted exclusive-OR) 方法[7]。

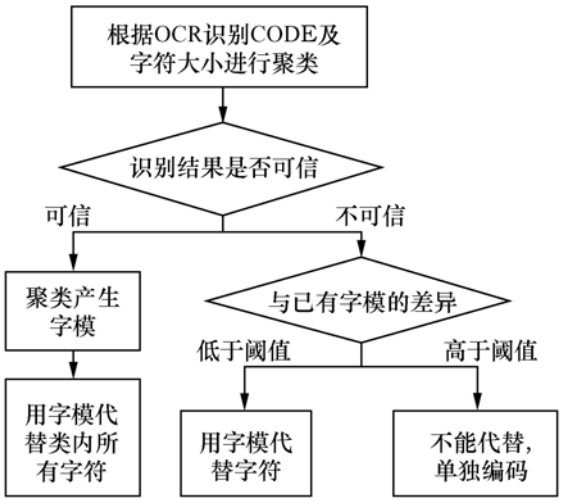


图 3　基于 OCR 信息的字模重建流程

3　实验结果

实验用二值文本图像，F01-200 与 F04-200 是 CCITT（Consultative Committee on International Telephone and Telegraphy）标准传真图像。sampleJT1 是中文报纸图像，sampleJT2 是中文杂志图像。HW 0004 和 SCBM 1 是中文书本图像。TOTAL 表示前 6 幅图像的总和。压缩比为

$$C_R = \frac{\text{原始图像数据大小}}{\text{压缩后图像数据大小}}.$$

其中原始图像数据大小为图像在没有经过压缩处理的位图格式的数据大小。

表 1 是 G4、基于 SPM 无损 JBIG2 压缩[4]和基于 OCR 的有损 JBIG2 压缩结果的比较。

表 1　无损压缩与有损 OCR 压缩比

图像名称	G4	无损 SPM	有损 OCR
F01-200	30.13	45.24	73.97
F04-200	7.87	14.58	39.67
HW 0004	23.27	32.00	34.43
SCBM 1	10.81	14.62	14.50
sampleJT1	7.36	9.89	10.74
sampleJT2	7.58	9.89	12.56
TOTAL	14.51	20.38	23.76

在实验图像上，本文算法比基于 SPM 算法的压缩比提高 14.3% 。

由表 1 看出，对于实验图像，本算法的压缩比都高于 G4。对于拉丁字母文本图像，本算法的压缩比高于基于 SPM 的无损 JBIG2 压缩；但是对于中文图像，提高并不明显。对图像 HW 0004、sampleJT1、sampleJT2，有损压缩比相对无损压缩有所提高；而对于 SCBM 1，压缩比有所下降。拉丁字母数量远少于汉字的数量，从而字母重复出现的次数远高于汉字重复的次数。因此，在有损压缩时，中文图像可提高的压缩比低于拉丁字母图像。可见文本图像压缩比的高低与图像本身关系很大，同一幅图像中重复字符出现次数越多，压缩效果越好。JBIG2 标准支持多页图像压缩，对于中文图像，多页图像压缩是一种提高压缩比的方法。当汉字增多时，字符重复出现的概率会增高，从而压缩比也会得到提高。

和以前的有损压缩方法相比，利用 OCR 也有其优势。选取的参考有损压缩方法是 PM&S 方法[2]，这种方法没有利用 OCR 信息。表 2 是基于 PM&S 有损压缩和基于 OCR 的有损压缩结果的比较。

表 2　无 OCR 与利用 OCR 有损压缩比

图像名称	PM&S 有损压缩比	利用 OCR 信息有损压缩比
F01-200	81.12	73.97
F04-200	36.09	39.67
HW 0004	34.69	34.43
SCBM 1	19.06	14.50
sampleJT1	12.83	10.74
sampleJT2	12.50	12.56
TOTAL	25.85	23.76

由表 2 看出，PM&S 有损压缩比在有些图像上高于基于 OCR 的有损压缩比。但是，这是有代价的。图 4 是对 F04-200 进行有损压缩的比较，令匹配阈值为 0.15。PM&S 出现了替换误差，而基于 OCR 的压缩则消除了这个问题。图 5 是对 F01-200 进行有损压缩的比较，令匹配阈值为 0.25，替换误差更加明显。PM&S 结果出现了大量的替换误差，而基于 OCR 的压缩消除了该问题。

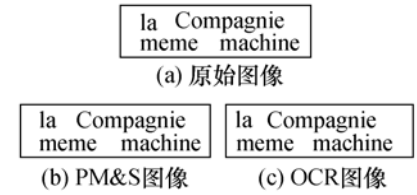


图 4　对 F04-200 进行有损压缩的结果

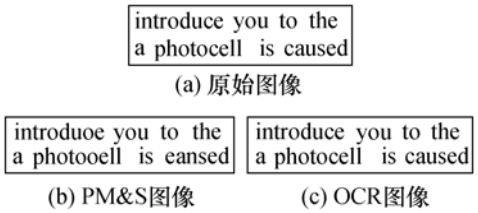


图 5　对 F01-200 进行有损压缩的结果

（下转第 1253 页）