广西科学院学报
Journal of Guangxi Academy of Sciences

2007,23(4):275~276
Vol. 23,No. 4 November 2007

大规模图文资料数字化的实现方法*
The Implementation on Largescale Document Imaging

龙 波[1],杨丽芳[2],肖 健[2],梁 莹[2]
LONG Bo[1],YANG Li-fang[2],XIAO Jian[2],LIANG Ying[2]

(1.南宁海蓝数据有限公司,广西南宁 530022;2.广西计算中心,广西南宁 530022)
(1. Highland Digital Technology INC. ,Nanning,Guangxi,530022,China;2. Guangxi Computing Center,Nanning,Guangxi,530022,China)

摘要:将大规模纸质图文资料通过高速扫描仪快速、大批量地输入计算机系统,并采用独特的高比率图像压缩技术以紧凑的结构形式原文原貌地存储于光盘等存储介质之中,可支持网络查询、调阅、打印等,也可以进行印刷体汉字识别,为传统纸质媒体大规模转换为电子媒体并得以有效利用提供了一个实际有效的方法。

关键词:图文资料 数字化 二值图像无损压缩 字符识别
中图法分类号:TP311.52 文献标识码:A 文章编号:1002-7378(2007)04-0275-02

Abstract: Document imaging is that paper documents and graphic files are put into the computer system by high-speed scanners and stored in the medium with high-rate image condensing technology such as disk. It can not only support inquiry, reading and printing on the internet but can recognize printed Chinese characters. It is an effective way to convert traditional paper media into electronic media.

Key words: document and graphic file, imaging, image condensing algorithm, character recognition

图文资料数字化的目标是使纸质媒体上的信息进入电子媒体,而且应尽可能保持原文原貌,并能方便地检索、阅读和引用。人们能够轻易实现少量纸质媒体的数字化,这在技术上并不是什么难题,然而要使千百年来积累下来的浩如烟海的纸质媒体实现数字化,例如实现图书馆和档案馆的数字自动化等,时间、准确性和成本费用才是最根本的问题。本文探讨大规模图文资料数字化生产的实现方法,旨在解决大量纸质图文资料进入计算机网络的存储、管理和发布等问题,以期为资料管理和使用提供更有效的支持。

1 软硬件平台及生产流程

1.1 软硬件平台

大规模图文资料数字化的实现包括硬件设备和软件平台,硬件包括高速扫描录入设备、高性能计算机和大容量存储设备以及图文资料的分解装订等辅助设备。

软件平台包括图像扫描输入、图像处理、管理和发布4部分。

(1)图像扫描输入:高速扫描图文资料并自动输入计算机;

(2)图像处理:图像去污与纠偏、图像压缩、图像加密、自动入库、全自动汉字识别与校对、目录自动提取录入数据库;

(3)管理:图像数据库设计与维护、图像编目标引、图像检索查询、图像数据的安全备份;

(4)发布:网页模板库的建立、网上资料阅读系统的自动生成、光盘资料阅读系统的自动生成等。

1.2 生产流程

经过多年的实践经验,我们对传统数字化生产流程进行优化,总结出了一套适合于大规模图文资料数字化的生产流程,具体如图1所示。

收稿日期:2007-09-28

作者简介:龙 波(1975-),女,助理工程师,主要从事图文资料数字化技术研究。

*应用基础研究专项项目(桂科基0342016)资助。

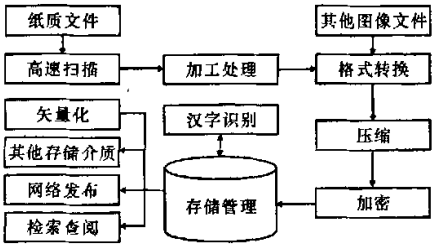

图1 大规模图文资料数字化生产流程

2 关键技术及指标效果

2.1 关键技术

实现大规模图文资料数字化的一个重要环节是如何实现数字化的图像资料进行高效无损的压缩。没有图像压缩技术的图文资料大规模数字化的产物是存储在电子媒介上的庞大繁复的数据，难以管理，无法共享，图文资料虽被数字化却无法信息化，大规模图文资料的数字化不能进入实用[1]。研究高效高质的图像压缩技术是要使含有一定信息量的数字化图文资源占用尽可能少的存贮空间和传输带宽[2]。现存的图书馆藏书及档案资料等资源主要是以黑白形式存在，同时，因为光学字符识别(OCR)，技术的要求，这些图文资料数字化后主要以二值文件形式存贮，针对二值图像的特点，我们对黑白图像的无损高效的压缩算法进行深入研究[3,4]。新算法采用的是自适应算术编码方法，研究方法如下。

(1)采用自适应算术编码方法，研究建立一种自适应的统计模型，该模型有快速的适应力，能以较高精度表示被压缩的数据。

(2)对算术编码方法进行研究，针对二值数据简化算术编码方法，重点解决算术编码的速度问题，最大限度减少或消除乘法运算，提高编解码速度。

(3)对大面积着色部分研究相应算法，进一步提高编解码速度和压缩率。

同时根据新算法生成的图像格式开发了包括具有放大、缩小、旋转、剪贴、打印等功能的该图像格式浏览器，以及与其他通用图像格式转换的工具软件，使大规模图文资料数字化技术上得以应用与实现。

2.2 指标效果

使用本文研究的大规模图文资料数字化生产流程和高效率二值图像无损压缩技术，可以达到如下的生产指标，满足了图文资料的大规模数字化生产要求。

(1)由于采用了自动纠偏、自动去污等多种自动处理功能，使系统的人工操作变得非常简单，扫描输入的速度与资料的复杂程度无关，只与扫描工作站的性能和人工熟练程度有关，资料数字化的速度基本上是随着扫描仪的性能和数量而直线上升，单机日产量不低于5000页。

(2)采用先进的压缩算法，彩色图像的压缩比达到1∶200，黑白二值图像的压缩比高于CCITTGroup4标准60%以上。

(3)每页资料的还原时间不高于0.5s。

(4)浏览器可同时阅读扫描图象文件、照片及直接浏览INTENET上的HTML语言。并支持远程查询。内含OCR功能，可随时对图书进行局部OCR，识别后的汉字文本可以任意进行修改、打印输出。

(5)5.2G光盘片可存储20万页A4幅面的文件资料。

3 结束语

目前全球数字内容产业年增长率达33%，正在成为全球经济最具活力、发展最快的新的增长点[5]。数字媒体成为一种新兴产业越来越得到政府、投资人、企业的重视，被誉为“经济发展的新引擎”[6]。本文通过对高效二值图像无损压缩算法的研究，在实践中对数字化生产流程进行优化，总结研究出一套适用于大规模图文资料数字化的信息加工技术，该技术解决了大量纸质图文资料进入计算机网络的存储、管理和发布问题，使大规模图文资料数字化得以实际意义上的实现。作为数字内容产业的核心技术，该技术在数字媒体数字化生产、数字内容管理等数字内容产业的应用领域具有十分广阔的应用前景。

参考文献：

[1] 陈武凡. 小波分析及其在图像处理中的应用[M]. 北京：科学出版社，2002.

[2] 耿则勋. 小波理论及在遥感影像压缩中的应用[M]. 北京：测绘出版社，2002.

[3] 傅德胜，寿亦禾. 图形图像处理学[M]. 福州：东南大学出版社，2002.

[4] 杨波，汪同庆，彭健，等. 基于模型基础的二值图像压缩方法[J]. 计算机应用，2002，22(5)：12-13.

[5] 曾红颖. 数字内容产业链未完全成形，产业融合化趋势明显[N]. 计算机世界，2007-07-30(133).

[6] 金元浦. 从文化产业到数字内容产业[EB/OL]. (2005-2-26). http://www.china.com.cn/chinese/zhuanti/whbg04-05/795874.htm.

(责任编辑：韦廷宗)