海蓝目录自动识别系统的设计
Design of Highland Catalog Auto-Recognition System

梁 莹，施善旦
Liang Ying, Shi Shandan
（广西计算中心，广西南宁 530022）
(Guangxi Computing Center, Nanning, Guangxi, 530022, China)

摘要：设计海蓝目录自动识别系统。该系统在获取已预处理过的目录图像之后，对图像进行版面分析，利用OCR技术识别文字，并自适应地获取目录的缩进量作为判断目录层次的依据，最后通过目录提取和人工校正得到统一的目录格式。该系统具有自动识别、提取书籍目录结构等功能，能有效地处理多种格式的书籍目录类型。

关键词：目录识别 OCR 版面分析 缩进量 目录提取 人工校正
中图法分类号：TP391

Abstract: Highland catalog auto-recognition system is proposed, which can handle various styles of catalog images, and its key technical characteristics are described. This system firstly analyzes the content of the catalog in the preprocessed images, then uses OCR technique to recognize the characters, then exploits the relative indent to get the hierarchical structure of the catalog, which is corrected manually to get the unified catalog format.

Key words: catalog recognition, OCR, layout analysis, indent, catalog extraction, manual correction

在数字化信息资源建设中，需要将现有的印刷品数字化并制作成能够用计算机阅读、理解、查询、检索的电子出版物。在文档资料和信息数字化过程中，低成本录入、有效保存、高效检索是图文数字化系统能否成功推广、应用的关键所在。

原有海蓝图文数字化系统的一个重要环节是采用人工录入的方式将加工书籍的目录输入数据库，导成海蓝目录索引文件，提供给海蓝文献浏览器，使其更直观地展现书籍的目录树型结构。原有海蓝图文数字化系统的处理流程如图1所示。由于这种工作模式工人劳动量大、加工效率低，使其成为了大规模生产的瓶颈。因此，本文设计海蓝目录自动识别系统（图2），利用先进的OCR技术进行文字识别，然后自动对识别文本进行分析，提取出书籍目录正文、层次、页码，并将这些信息按树型结构组织起来，录入数据库。

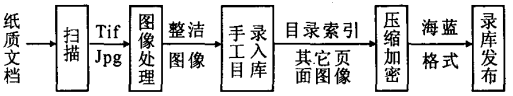


图1 原图文数字化系统处理流程

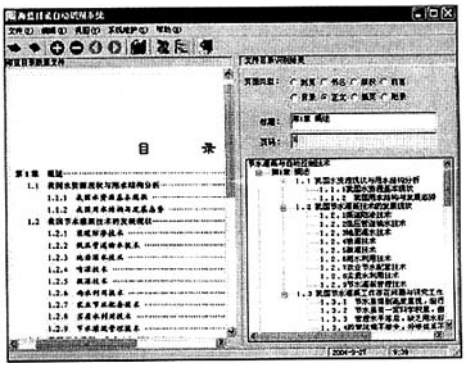


图2 海蓝目录自动识别系统

1 书籍目录类型分析

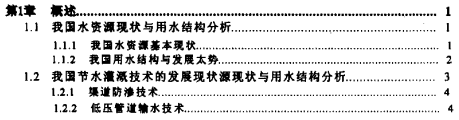
海蓝目录自动识别系统能够处理多种出版目录格式。现在市场上书籍的目录排版复杂、形式多样，目录排版格式大体可分为：单列有缩进型（图3），双列有缩进（图4），以及复杂目录类型（图5），其中，单列有缩进型、双列有缩进型是最常见的2种目录结构类型。

2 系统分析

海蓝目录自动识别系统首先获取经过预处理的目录图像，经过版面分析、文字识别之后，再通过目

2004-09-30 收稿。

录提取，将获得的目录结构录入数据库，结合人工校正，最终成功生成海蓝目录索引文件。其流程如图 6 所示。


目　　录

第1章　概述 …… 1
1.1　我国水资源现状与用水结构分析 …… 1
1.1.1　我国水资源基本现状 …… 1
1.1.2　我国用水结构与发展态势 …… 2
1.2　我国节水灌溉技术的发展现状 …… 3
1.2.1　渠道防渗技术 …… 4
1.2.2　低压管道输水技术 …… 4

图 3　单列有缩进型

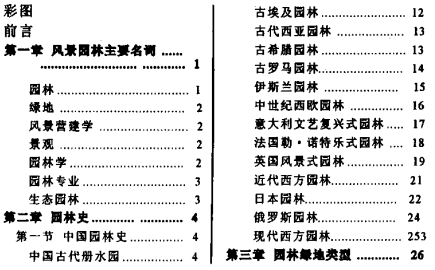

彩图
前言
第一章　风景园林主要名词 …… 1
园林 …… 1
绿地 …… 2
风景营建学 …… 2
景观 …… 2
园林学 …… 2
园林专业 …… 3
生态园林 …… 3
第二章　园林史 …… 4
第一节　中国园林史 …… 4
中国古代册水园 …… 4
古埃及园林 …… 12
古代西亚园林 …… 13
古希腊园林 …… 13
古罗马园林 …… 14
伊斯兰园林 …… 15
中世纪西欧园林 …… 16
意大利文艺复兴式园林 …… 17
法国勒·诺特乐式园林 …… 18
英国风景式园林 …… 19
近代西方园林 …… 21
日本园林 …… 22
俄罗斯园林 …… 24
现代西方园林 …… 253
第三章　园林绿地类型 …… 26

图 4　双列有缩进型

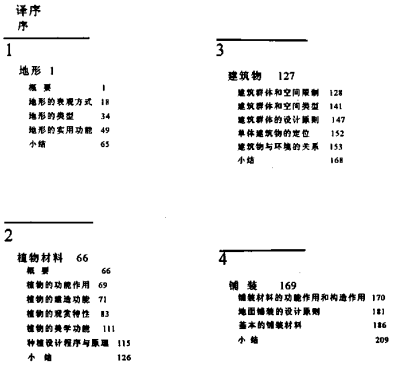

译序
序
1
地形　1
概　要　1
地形的表现方式　18
地形的类型　34
地形的实用功能　49
小结　65
3
建筑物　127
建筑群体和空间限制　128
建筑群体和空间类型　141
建筑群体的设计原则　147
单体建筑物的定位　152
建筑物与环境的关系　153
小结　168
2
植物材料　66
概　要　66
植物的功能作用　69
植物的基地功能　71
植物的观赏特性　83
植物的美学功能　111
种植设计程序与原理　115
小　结　126
4
铺　装　169
铺装材料的功能作用和构造作用　170
地面铺装的设计原则　181
基本的铺装材料　186
小　结　209

图 5　复杂目录页面

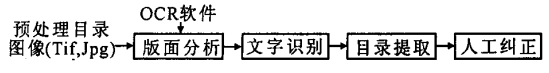


图 6　目录识别流程

2.1　图像预处理

利用高速扫描仪以提高文档的扫描效率。扫描前要选择适当的分辨率，太低达不到识别要求，太高使图像文件太大，影响识别速度，一般用 300dpi 光学分辨率就能达到较好的识别效果[1]。获取图像后，需要对图像进行自动倾斜纠正、污点消除等预处理操作[2]，以提高图像质量，进而提高 OCR 的识别准确率。本系统能兼容各式各样的扫描仪。

2.2　版面分析

书籍目录页面的类型多样，但它们的共同点是，每一个目录页面都是由多个单列有缩进型的子目录构成的。通过版面分析的方法，可以得到这些相对容易处理的子目录，避免复杂目录页面的处理。

版面分析是一种文字识别的前处理技术[3]，它是将扫描得到的图像，按页面内容的分布聚集程度、空白间隔等信息划分成多个矩形块。对于各个不同的区域块，遵循常用从上至下、从左至右的排版原则，标明不同区域块之间的顺序，以便系统进行识别处理。版面分析有自动、手动两种方式。自动分析适合应用在结构清晰、简单的目录上，减少人为干预，提高生产效率。当遇到结构复杂的目录可以辅以手动的方式，人为地划分目录板块，根据区域块划分的先后，排列目录内容的顺序。分析过程如图 7 所示。

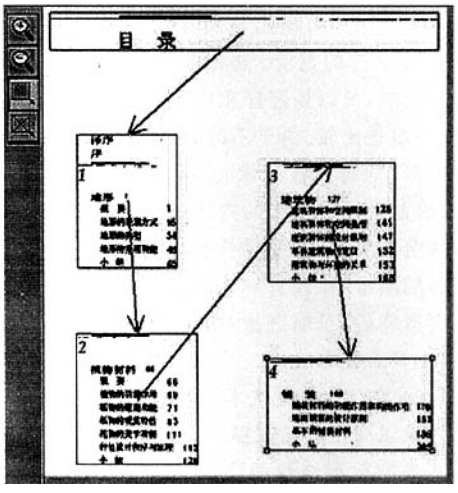


图 7　手动版面划分

2.3　文字识别与目录提取

2.3.1　文字识别

OCR 是图文数字化的核心技术之一[1]，它通过扫描和摄像等光学输入方式获取纸张上的文字图像信息，利用各种模式识别算法分析文字形态特征，断出汉字的标准编码，并按通用的格式存储在文本文件。海蓝目录自动识别系统中具有多引擎 OCR 识别功能，能够集成市场上比较成熟的各种 OCR 系统，如汉王 OCR、清华 TH-OCR 系统。

2.3.2　目录格式处理

用空格符对页面的空白处进行填充，然后将识别结果以回车换行符为结尾的方式写入文本文件，以确保识别结果能原文原貌地反映目录页面结构。

(1)文字定位。每个字符都能查找到它们在页面图片中的位置(即坐标)、高度(厚度)、宽度。这些定位信息为目录层次的判定提供依据。

(2)目录行定位。结合识别文本和文字定位信息，可以得到目录页面图片上任一文字行的起始坐标、高度(厚度)。

(3)目录行调整。有时候 OCR 引擎会把一个目录标题行识别为多个文本行，或者由于目录行本身

过长被分两行排版，这两种情况都需要把多目录拼接为一个。

2.3.3 缩进的自适应判断

缩进的自适应判断包括缩进量的自适应确定和比较两个方面。

(1)缩进量自适应的确定。在不同的排版、印刷方式下，目录文字的缩进量也不同，因此无法用统一的缩进量标准来判断不同格式的目录页的目录标题的缩进量。为此，本文采用自适应的方法，根据页面上的字符宽度，判断出该目录页的缩进量大概是1.5～2个字符的宽度。这种自适应的方法没有排版格式的限制，可以处理任意格式的目录图像。

(2)缩进比较。由于不同的排版方法会导致页边距的差异，同时扫描图像时也可能产生图像的相对移动，因此采用自适应的方法：在页面中要按不同的版面块分别识别，以板块中最靠左边的字符坐标作为缩进的基准，比较每一行的起始坐标，就可判断该行是否有缩进，且缩进量多大，进而确定它的层次。

2.3.4 目录提取

解决了以上关键技术后，读取OCR识别文本，并将目录以树型的结构提取出来。本文采取以编码的方式反映树型结构的层次，并将这种结构信息写入数据库。目录树型如图8所示。

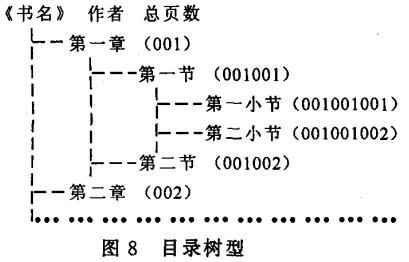


图8 目录树型

2.4 人工纠正

由于OCR文字识别不可能达到100%的正确率，而且在目录页中，目录标题缩进的比较是相对于同一版面块中而言，有时会出现同一版面块由同一层次、且恰好都是较深层次的目录标题构成的情况，但是由于没有比较，程序往往会把它们作为第一层目录标题，破坏了书籍目录的层次结构。因此，必须要有人工纠正模块，可以修改错字，调整目录结构。该模块采用数据库系统，能将修改好的书籍目录信息存储到数据库，以利于今后的进一步利用。例如：网站发布，除了能直接在网上泛泛了解到题名、作者等基本信息外，能更具体、详细了解到资料内部的内容和框架。

在通过上述处理流程后，前面举的例子全部能得到统一的目录格式。

3 系统功能及开发工具

3.1 系统功能

海蓝目录自动识别系统具有目录自动识别、入库，人工校正和备份、恢复数据库三大功能。

(1)目录自动识别、入库功能。可快速、准确地从目录页面图像中自动识别、提取出书籍目录正文、层次、页码，并将这些信息按树型结构组织起来，写入数据库；

(2)人工校正功能。可对照目录图像，对识别、提取结果进行校正，确保各种目录页面最终都能正确、统一再现目录树型层次；

(3)备份、恢复数据库功能。只需鼠标，即可实现远程备份、恢复数据库，不需用户掌握SQL SERVER的各种复杂的编程命令，并有效防止数据丢失。

3.2 开发工具

由于该系统主要由目录识别和数据入库两部分组成。目录识别的效率对目录提取的高效性至关重要，因此这部分采用Visual C++ 6.0开发。另一方面，由于本系统主要在局域网环境中应用，网络传输速度高达10M～100M，因而对数据入库的速度、性能没有过高的要求，而Visual Basic 6.0开发周期短，还拥有强大、易用的数据操作对象ADO，所以数据入库部分采用VB6.0开发。考虑其数据量并不大和开发环境Windows、开发工具的兼容性等情况，本系统数据库采用Micosoft的SQL Server 2000。

4 结束语

本文介绍海蓝目录自动识别系统的设计方法，该软件的出现打破了以前完全依赖人工的工作模式，能自动提取出书籍目录结构，改善加工生产线的生产效率，提高了海蓝图文数字化系统的生产自动化程度。

参考文献：

1 汉王OCR、TH-OCR网站服务与支持. http://www.wintone.com.cn，http://www.hw99.com.2002.

2 Kenneth R Castleman. 字图像处理，北京：清华大学出版社，1998.

3 张永慧. 版面分析与理解. 语言文字应用，1997，(2)：92～98.

(责任编辑：黎贞崇)