《中文印刷体文档识别技术》

图书基本信息

书名：《中文印刷体文档识别技术》

13位ISBN编号：9787030287601

10位ISBN编号：7030287606

出版时间：2010-8-1

出版社：科学出版社

作者：王科俊,冯伟兴

页数：203

《中文印刷体文档识别技术》

前言

随着科技的发展，人类社会正经历从工业化社会向信息化社会的转变，信息化程度越来越高。近年来，伴随互联网的迅速普及，通过互联网这一方式进行信息传播和交换已成为人们日常工作生活的首选信息交流方式。为了促进信息交流效率，中文印刷体文档识别技术日益受到众多学者的关注。目前，具有汉字和符号识别功能的印刷体识别软件（OCR）已在实际中得到广泛应用，但是一个中文文档中不仅含有汉字和符号，还含有特殊字符以及各种各样的公式和图表.而现阶段的中文文档识别软件尚不能对公式等这些文档内容进行识别和处理，迫切需要一种既能识别汉字又能识别和处理公式等其他文档内容的较为全面的中文文档识别系统。针对这一现状，我们开展了以公式为主的中文印刷体文档识别研究，本书就是我们近几年来在这一领域研究成果的总结。本书作为国内第一部关于中文印刷体文档识别技术的著作，系统地分析了中文印刷体文档识别技术的各个方面，包括文档图像的预处翠、版面分析、文字和符号识别、公式定位和提取、公式结构分析与表示、表格识别和文档中的图形图像处理等内容。结合作者多年来在公式识别方面取得的研究成果重点给出了公式的定位与提取和公式的结构分析的理论与方法。

《中文印刷体文档识别技术》

内容概要

《中文印刷体文档识别技术(附光盘1张)》全面阐述了中文印刷体文档识别的原理、方法和系统组成，依据中文印刷体文档的特点，分别介绍了文档图像预处理、版面分析、汉字识别、公式的定位与提取、公式字符分割与识别、公式结构分析与表示、图表处理等内容的基本原理和技术实现方法，并提供了一个中文印刷体文档识别系统实例。

《中文印刷体文档识别技术》

书籍目录

《智能科学技术著作丛书》序前言第1章 绪论 1.1 中文印刷体文档识别基本原理 1.2 中文印刷体文档识别研究现状 1.2.1 印刷体文档的汉字识别 1.2.2 印刷体文档的公式识别 1.2.3 印刷体文档的表格识别 1.3 中文印刷体文档识别中的难点第2章 中文印刷体文档图像预处理 2.1 中文印刷体文档图像采集 2.1.1 文档图像采集 2.1.2 文档图像显示 2.1.3 文档图像格式 2.2 中文印刷体文档图像特点 2.3 二值化处理 2.3.1 图像灰度化 2.3.2 图像二值化 2.4 平滑去噪 2.4.1 邻域平均法 2.4.2 中值平均法 2.4.3 噪声直接去除法 2.5 倾斜校正 2.5.1 图像倾斜检测 2.5.2 图像倾斜校正第3章 版面分析 3.1 版面结构 3.2 版面分析方法 3.2.1 基于连通域的版面分析方法 3.2.2 二分法 3.2.3 基于组合特征的版面分析方法 3.2.4 基于神经网络的版面分析方法 3.2.5 基于最近邻连接强度和行列可信度的版面分析方法 3.3 版面理解 3.3.1 文字区域 3.3.2 图片区域 3.3.3 表格区域 3.3.4 版面结构表示与存储 3.4 版面重构第4章 印刷体汉字识别 4.1 文本区域预处理 4.1.1 文本增强 4.1.2 字符分割 4.1.3 字符细化 4.1.4 字符归一化 4.1.5 文本区域处理效果图 4.2 印刷体汉字的特征提取 4.2.1 印刷体汉字的统计特性 4.2.2 印刷体汉字的常用特征 4.3 印刷体汉字识别的实现方式第5章 公式的定位与提取 5.1 印刷体文档公式的特点 5.2 基于投影的公式定位和提取 5.2.1 独立行公式的定位 5.2.2 内嵌公式的定位 5.3 基于Parzen窗的独立行公式定位和提取 5.3.1 待分类文本行的特征数据提取 5.3.2 Parzen窗方法 5.3.3 公式定位与提取效果 5.4 基于字符宽度中心矩的公式定位和提取 5.4.1 文本区域基本数据获取 5.4.2 含公式的文本行提取 5.4.3 文本行中公式判别 5.4.4 独立行公式的定位 5.4.5 内嵌公式的定位 5.4.6 公式定位与提取效果 5.5 基于汉字拒识的内嵌公式定位和提取 5.5.1 内嵌公式的定位 5.5.2 公式定位与提取效果第6章 公式字符分割与识别 6.1 公式字符的特点 6.2 公式字符的分割 6.2.1 基于轮廓跟踪的字符分割 6.2.2 基于连通域的字符分割 6.3 公式字符的识别 6.3.1 公式字符图像预处理 6.3.2 基于模板匹配的公式字符识别 6.3.3 基于特征的公式字符识别 6.3.4 印刷体公式字符识别的实现 6.3.5 公式字符识别方法第7章 公式结构分析与表示 7.1 公式结构分析的难点 7.1.1 数学运算符的模糊性 7.1.2 符号的上下文敏感性 7.1.3 表示习惯的差异性 7.1.4 公式的复杂性 7.1.5 公式的多行结构 7.2 公式结构分析前的字符预处理 7.3 公式结构分析方法 7.4 公式结构表示方法 7.4.1 公式的典型表示方法 7.4.2 实验结果第8章 图表处理 8.1 文档中图形图像的表示与处理 8.1.1 游程压缩 8.1.2 霍夫曼编码压缩 8.1.3 算术压缩方法 8.1.4 Rice压缩方法 8.1.5 LZW压缩方法 8.2 文档中表格的分析与识别 8.2.1 表格预处理 8.2.2 表格直线提取 8.2.3 表格结构分析 8.2.4 表格字符提取与识别第9章 中文印刷体文档识别软件HEUOCR的设计与实现 9.1 应用程序框架的构建 9.1.1 框架风格 9.1.2 数字图像处理类 9.2 文档图像预处理 9.2.1 图像灰度化 9.2.2 图像平滑滤波 9.2.3 图像阈值分割 9.3 文档图像版面分析 9.3.1 基本连通域提取 9.3.2 基本连通域分析 9.4 文本汉字识别 9.4.1 字符分割 9.4.2 字符识别 9.5 公式识别 9.5.1 公式定位 9.5.2 公式字符分割 9.5.3 公式字符特征提取 9.5.4 公式字符识别 9.5.5 公式结构分析参考文献

《中文印刷体文档识别技术》

章节摘录

插图：信息化理念已经被很多人所熟悉，人们越来越追求一种有力的、简洁的、准确无误的信息交流手段。由于人们日常生活中接收到的绝大多数信息是以图像的形式进行传递的，尤其是依托互联网的数字图书馆和远程教育的兴起，使得图像信息自动识别技术有着广泛的应用前景和重要的研究价值。中文印刷体文档识别技术就是一个典型的针对含有中文字符图像的信息自动识别技术。1.1 中文印刷体文档识别基本原理现有的文字识别技术一般采用光学的方式将文字图像信息采集到计算机中，因此，该类技术常被称为光学字符识别（optiCal character recognition，OCR）技术。经过近一个世纪的发展，OCR已经成为当今模式识别领域中最活跃的研究内容之一。它综合了数字图像处理、计算机图形学和人工智能等多方面的知识，并在计算机及其相关领域中得到了广泛应用。按照识别方法，OCR识别方法可以分为如下三类：统计特征字符识别技术、结构特征字符识别技术和基于人工神经网络的字符识别技术。作为OCR技术的一个重要研究方向，印刷体文档识别主要针对比较正式、规范的书籍、报刊和杂志的图像信息进行采集和识别。与一般文档图像相比，印刷体文档图像存在前景信息与背景信息色差显著，文字信息形式规范等特点，这都为印刷体文档的信息处理和识别创造了便利条件。然而，各类印刷体文档中除了包含文字信息以外，还常有公式、表格以及各种各样的图形等信息，因此，若将印刷体文档中包含的所有信息都完整地识别出来，也不是一件易事。

《中文印刷体文档识别技术》

精彩短评

1、里面光盘断裂了。
2、讲得很系统，从原理上深度刨析了ocr的各个技术要点，但是最好配合一本讲算法的书，一起学习，这样效果更佳
3、送货比较快，可惜书上有价值的内容不多，已经束之高阁
4、光盘内容太乐色了，根本没法用，可见这本书基本乐色。
5、内容还不错，就是主要光盘质量。
6、不够深入.适合略做了解
7、正版.比书店便宜
8、不错，国内一本不错中文识别书
9、大家不要买这本书，就是**

《中文印刷体文档识别技术》

章节试读

1、《中文印刷体文档识别技术》的笔记-第68页

根据汉字笔画密度先做一个粗分类——恩，多层特征提取匹配。。。

2、《中文印刷体文档识别技术》的笔记-第138页

Canny算子具有最好的边缘检测性能。。。
Shit，都10年过去了，居然还是Canny算子，fuck

3、《中文印刷体文档识别技术》的笔记-第59页

印刷体汉字的特征提取：
1、像素分布
2、外围
3、像素段
4、笔画
5、网格 --这种做法实在是太垃圾了！
6、封闭区域
7、方向线
8、投影
9、结构点

4、《中文印刷体文档识别技术》的笔记-第16页

最大墒图象阈值分割：意思就是让分割出来的背景更像背景。。。

哦，我明白了

5、《中文印刷体文档识别技术》的笔记-第21页

倾斜校正：在y轴上投影，每行文字应该对应一个明显的波峰。

6、《中文印刷体文档识别技术》的笔记-第14页

H(D) = - d/dD A(D)

图象直方图是其面积的负导数

7、《中文印刷体文档识别技术》的笔记-第77页

基于Parzen窗的独立行公式定位与提取。。。

《中文印刷体文档识别技术》

版权说明