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

2011,27(4):317~319
Vol. 27,No. 4 November 2011

基于多引擎的印刷体汉字识别系统的设计
Development of Multi-engine Printed Chinese Character Recognition System

梁 莹,肖 健,李 玥
LIANG Ying,XIAO Jian,LI Yue

(广西计算中心,广西南宁 530022)
(Guangxi Computing Center,Nanning,Guangxi,530022,China)

摘要:设计一种基于多引擎的印刷体汉字识别系统,优先采用汉王光学字符识别(OCR)引擎的版面分析结果,在汉王、清华OCR引擎分别完成字符识别之后,根据字符的图像坐标,整合两者的识别结果,并用彩色突出两OCR引擎的冲突字符、置信度低的字符及WiseCheck语义校对引擎提示的错误字符。该系统改善了现有大规模数字化加工生产线中人工比照图像时对识别文本逐字、全文遍历式校对的工作模式,能减轻劳动强度,提高工作效率,降低处理成本。

关键词:汉字识别 光学字符识别 语义校对 多引擎
中图法分类号:TP391.1 **文献标识码**:A **文章编号**:1002-7378(2011)04-0317-03

Abstract:A printed Chinese characters recognition system based on multi-engine has been constructed. Basing on the HW-OCR engine's layout analysis, the HW-OCR and TH-OCR engines accomplished character recognition respectively. According to the coordinate of the character image, the system will integrate the two OCR engine's recognition results using different colors to highlight their conflict character and low confidence character,and the other wrong words which are checked by the "WiseCheck" (a semantic collation engine). This system has improved the text verbatim identification by artificial contrast image and full-text search proofreading work mode in the existing mass digitization processing production line, which further can reduce labor intensity, improve work efficiency and reduce the cost of processing.

Key words:Chinese character recognition,OCR,semantic collation,multi-engine

原有海蓝大规模图文资料数字化生产线的高速扫描、图像处理、压缩加密、识别、入库发布等工序都已具有较高的自动化程度。但在识别工序的校对环节,仍然是全部采用人工比照图像对识别文本逐字、全文遍历式校对的工作模式,严重影响了海蓝大规模数字化加工生产线的效率。而且,在实际的市场应用中,有部分的政府机关、事业单位,将自己的馆藏资料面向公众发布并提供利用时,也提出要求识别文本应达到国家印刷出版的标准:误差率在万分之五以下。目前市场上的光学字符识别(OCR)软件均有一定局限性,看起来单机运行效果不错的也只有90%多的识别率,距离要求的5‰以下误差率仍存在很大的差距。本文设计了一种集成多种成熟印刷体汉字识别与语义校对的汉字识别系统(以下简称,多引擎系统)。它可以改善现有数字化加工生产线中的人工校对工作模式,提高生产线效率,减轻人工劳动强度,降低处理成本。

1 系统设计技术方案

1.1 工艺技术路线

(1)研究、收集市面上成熟的印刷体汉字识别软件,分析研究印刷体汉字识别的原理[1],建立印刷体汉字识别准确率评估体系,统计分析各识别软件的实际识别准确率与系统特点。

(2)研究、收集市面上成熟的汉语语义校对软

收稿日期:2011-09-16
作者简介:梁 莹(1978-),女,工程师,主要从事信息技术研究。

件，分析研究上下文关联的汉语语义校对原理与系统特点。

(3)分析研究各识别引擎(软件二次开发包)的调用方法，设计程序进行试验，并总结利用各系统的优点。

(4)分析研究各汉语语义校对引擎的调用方法，设计程序进行试验。

(5)研究多个OCR、语义校对引擎的集成整合技术并开发出新的软件系统。

(6)采集样本进行原型试验，验证研究成果。

通过对上述工艺技术路线的分析可以看出，多引擎系统实现的难点在于：多个OCR、语义校对引擎的集成整合。市面上成熟的汉语语义校对软件并不多见，由北京语言文化大学宋柔教授主持的国家"863"计划、国家自然科学基金项目——"现代汉语通用分词系统"，通过了专家鉴定委员会的鉴定，居国际领先水平[2]。通过对宋柔教授提供的汉语语义校对引擎WiseCheck的试用，明确其具有准确和高效的分词、灵活的二次开发接口、可定制词库等特性，比较适合大规模图文资料数字化加工所需面对的各种不同专业领域、不同华语地区的分词、语义校对需求。

1.2 OCR引擎差异试验

1.2.1 挑选样张

为了保障对汉字的识别覆盖，挑选了不同类型的文集，如现代汉语专业书籍、近代章回小说、台湾散文集等；同时，兼顾大规模图文资料数字化生产线上常见的各种典型样张如，简繁体、图文混排、表格等，进行识别试验，累计识别样张1000多页，简体近35万字、繁体近18万字。

1.2.2 记录各OCR软件识别情况

用ACCESS做识别情况记录，将不同OCR引擎的识别结果分成汉王简体、汉王繁体、清华简体、清华繁体4个数据库存储记录。每个数据库，包含3张表：tb_Book(表1)，tb_Page(表2)、tb_Error(表3)。

表1 书籍 tb_Book

字段名称	说明
ISBN	ISBN号，主键
BookName	书籍名称
CharType	字符类型：简体/繁体
OCREngine	OCR引擎：汉王/清华

1.2.3 试验结果分析

通过在ACCESS中建立关联查询，对试验结果进行统计、分析。在单字识别准确率方面，简体(准确率|错误率)：汉王99.73%|27‰，清华98.84%|116‰；繁体：汉王98.90%|110‰，清华98.86%|114‰；在版面分析准确率方面，简体：汉王100%，清华95.88%；繁体：汉王96.94%，清华87.22%。因此，汉王比清华的版面分析更准确，两者的单字识别准确率相当接近。

表2 图像页面 tb_Page

ID	自增，主键
ISBN	ISBN号
PageName	图像文件名
WordCount	字数
LayoutErr	版面分析错误：0/1

表3 识别错误 tb_Error

字段名称	说明
ID	自增，主键
ISBN	ISBN号
PageName	图像文件名
ErrChar	错误字原文(例如：瘦)
ErrResult	识别结果(例如：痕)
Confidence	置信度集(值越小识别结果可信度越高，如，清华：8 /251/ 254/ 255 /255；汉王：A /D /E)
Candidate	候选字集(例如：痕疸瘦瘟摸)
Hit	第几命中(例如：3，如无命中则为"—")
ErrType	错误类型(字符错误：0，标点错误：1)
ErrLine	错误字所在行
CharNO	错误字所在列(行中的位置)
Coordinate	图像中的矩形坐标

同时，还发现两个OCR软件在版面分析的大板块划分结果一致的情况下，字符图像位置的误差在半个字符宽、高的范围内，就可以认定是同一字符，每一个字符就可以通过坐标一一对应。这一发现，为找出两OCR的差异及整合奠定了基础。根据这条规律，设计程序将两OCR软件识别错误结合到一张数据表里。通过对汉王、清华识别结果错误一致的字符的查询，评估出两个OCR软件识别结果结合后的错误率(盲点)，简体：1.45‰，繁体2.22‰，达到国家印刷出版要求错误率低于5‰的标准。

2 系统的主要流程

根据试验的结论，多引擎系统(图1)优先采用汉王的版面分析结果，在两OCR引擎完成识别之后，根据字符图像坐标，整合识别结果，并用彩色突

出两个 OCR 引擎识别结果的冲突字符、置信度低的字符,及语义校对提示的错误字词,从而使校对人员仅需检查彩色字符,就能把系统识别结果错误率控制在万分之五以下。

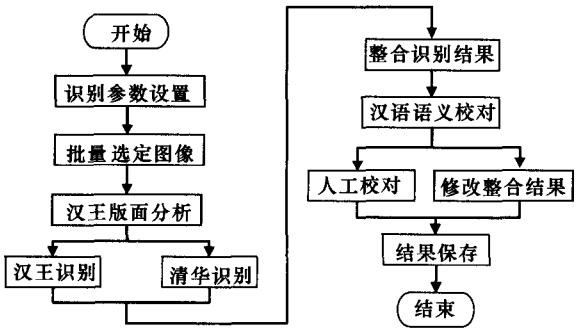

图 1　系统主要流程

3　系统的功能模块

系统的功能模块如图 2 所示。

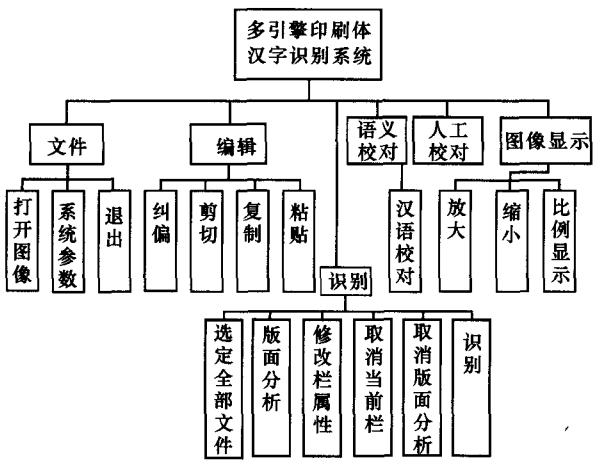

图 2　功能模块

3.1　识别模块

3.1.1　版面分析

对图像进行版面分析,在图像中显示分析结果:用彩色边框表示不同属性的板块(横文——红色边框、竖文——褐色边框、表格——蓝色边框、图像——绿色边框),各板块左上角的数字为识别顺序。同时,还能对版面分析结果(板块数量、顺序、属性)进行手动的修改。

3.1.2　识别

对版面分析结果进行单字符识别。为了便于人工校对,多引擎系统在识别结果文本区域用彩色突出两个 OCR 引擎识别结果中冲突的字符、置信度低的字符。同时,为了便于今后对系统做进一步的研究、改进,用不同的颜色做区分。蓝色——汉王、清华 OCR 识别冲突字符,浅蓝色——汉王、清华 OCR 识别冲突但有重合集的字符,粉红色——置信度较低(D 等级)的字符,红色——置信度最低(E 等级)的字符。

3.2　语义校对模块

对识别结果文本进行汉语语义校对,即运用上下文关联的计算机语义校对模型对识别结果进行校对,标示出语义有问题的字符。例如,红色——易错词;绿色——其它不太重要的错误,如接续错误、重复错误、格式错误、数字标记、敏感词、人名、地名、日期和时间、企业和机构等。

3.3　人工校对模块

随着校对人员控制光标在识别结果文本间的移动,系统会自动同步在图像显示区域用红色小方块自动标示出对应字符的图像区域;同时,候选字栏将显示对应字符的汉王、清华、语义校对候选字词集,点击候选字可自动替换文中被怀疑的字符。

3.4　图像显示模块

校对人员可根据查看的需要控制图像显示的大小,可以按"显示"菜单中设定的比例逐步放大、缩小图像;图像显示比例可控制在 100%～12.5%。

4　结束语

基于多引擎的印刷体汉字识别系统的出现,打破了以前完全依赖人工比照图像对识别文本逐字、全文遍历式校对的工作模式,使校对人员仅需检查彩色字符,就能把系统识别结果错误率控制在 5‰以下。该系统提高生产线效率,减轻人工劳动强度,降低处理成本。

参考文献:

[1] 汉王科技股份有限公司. 汉王 OCR、TH-OCR 网站服务与支持[EB/OL]. http://www.hw99.com, http://winton.com.cn, 2007.

[2] 罗智勇,宋柔. 现代汉语通用分词系统中歧义切分的实用技术[J]. 计算机研究与发展, 2006, 43(6): 1122-1128.

(责任编辑:尹　闯)