高性能中日韩文档识别理解重构系统

关 键 词：韩文 文档识别 电子文档 自动识别 文档重构 中文 日文 文档理解

所属年份：2004	成果类型：应用技术
所处阶段：	成果体现形式：
知识产权形式：	项目合作方式：

成果完成单位：清华大学电子工程系

成果摘要：

在全球信息化时代，在信息资源获取的全球战略竞争中，中日韩文等国文字和文档识别研究对于获取和利用以文档为载体的全球化信息资源具有重要战略意义。清华大学电子工程系研制成功的高性能中日韩文档识别理解重构系统为中日韩文档自动识别录入的信息获取、文档物理结构与逻辑结构理解、重构提供了统一全面的解决方案。经美国微软和SCANSOFT公司的测试，系统性能达到国际领先水平。高性能中日韩文档识别理解重构系统是清华大学电子工程系在国家863计划信息领域计算机软硬件主题“全方位智能化中文信息处理平台”(项目编号2001AAll4081)和国家自然科学基金项目“汉字识别研究中若干核心问题的新开拓”(项目编号69972024)支持下研制成功的。该系统为中日韩文档自动识别录入及文档物理、逻辑结构信息获取提供了统一全面的解决方案。2002年7月10日，教育部委托清华大学主持在清华大学召开了“高性能中日韩文档识别理解重构系统”技术鉴定会。鉴定意见如下：高性能中日韩文档识别理解重构系统实现了高识别率、高鲁棒性、超大字符集的统一印刷体汉、日、韩；文字识别核心；具有融合字型结构的识别信息、字符外形以及上下文语言知识的多层次信息融合迭代运算的汉、日、韩文字与英文混排文本的字符切分技术；具有能够适应中日韩文本图像变化的自动版面分析技术和版面物理结构重构技术；具有文档逻辑结构自动理解，从多页文档页面物理结构及文本内容信息进行页面及整体逻辑结构的理解，并以此实现自动提取多页文档逻辑结构及原数据，最终重构生成PDF和以XML为基础的OEB标准电子图书等标准格式电子文档。鉴定委员会一致认为：高性能中日韩文档识别理解重构系统首次在一个统一的框架下实现了印刷体汉、日、韩文字的识别和文档重构，其整体性能达到了国际领先水平。该系统为纸介质文档转化为保留全信息电子文档提供了有效的工具。希望进-步完善多页文档逻辑结构自动理解的功能，在信息化事业中推广应用。技术指标：有能抵御各种干扰变形和各种字体变化的高鲁棒性、超大字符集的印刷体东方文字(简繁汉日韩)识别核心；有东方文字(简繁汉日韩)与英文混排文本的字符切分技术，通过融合字型结构识别信息，字符外形以及上下文语言知识的多层次信息，进行迭代搜索最佳切分位置，成功解决了复杂类型字符混排、粘连和断裂情况下的字符切分问题；具有自动版面分析技术，对中日韩等文档图像进行版面物理结构和属性的分析；具有文档版面理解和重构技术，从多页文档页面物理结构及文本内容信息到页面及整体逻辑结构的理解，并一次实现自动提取多页图书文档逻辑结构及元数据，最终重构生成以PDF和以XML为基础的OEB标准电子图书等标准格式电子文档；统一的高性能中日韩文档识别理解重构系统为中日韩文档信息资源的建设提供了有力的工具，可广泛用于数字图书馆工程建设，中国信息基础建设上，具有显著的经济效益和社会效益。应用说明：此项成果已被美国微软公司和SCANSOFF公司以80万美金购买使用权，还将大量使用在数字图书馆工程和中国信息基础建设上，有广泛的经济效益和社会效益，有广泛的应用前景。合作方式：面议。

成果完成人：

完整信息

行业资讯

新疆综合信息服务平台
准噶尔盆地天然气勘探目标评价
维哈柯俄多文种操作系统FOR ...
社会保险信息管理系统
塔里木石油勘探开发指挥部广...
四合一多功能信息管理卡MISA...
数字键盘中文输入技术的研究
软开关高效无声计算机电源
邮政报刊发行订销业务计算机...
新疆主要农作物与牧草生长发...

成果交流

推荐成果

· 液压负载模拟器	04-23
· 新一代空中交通服务平台、关...	04-23
· Adhoc网络中的QoS保证(Wirel...	04-23
· 电信增值网业务创意的构思与开发	04-23
· 飞腾V基本图形库的研究与开发...	04-23
· ChinaNet国际(国内)互联的策...	04-23
· 电信企业客户关系管理(CRM)系...	04-23
· “易点通”餐饮管理系统YDT2003	04-23
· MEMS部件设计仿真库系统	04-23