计算机工程与应用 2010 46 (12): 5-8 ISSN: 1002-8331 CN: 11-2127/TP

博士论坛

综合文字和非文字区域特征的文档图像检索

张 田

山东大学 信息科学与工程学院，济南 250100

收稿日期 2010-1-27 修回日期 2010-3-12 网络版发布日期 2010-4-21 接受日期

摘要 提出一种改进的自适应文字区域提取算法，将文档图像分割成文字区域和非文字区域。对文字区域提取连通字符间空白、连通字符高度和宽度等局部特征，以及书写样式、段落特征等全局特征；对非文字区域，提取关键块特征。然后利用检索算法将文字区域特征和非文字区域特征结合起来，提高检索的准确性。同时，在检索算法中引入多维数据检索结构，有效地提高检索速度。通过对大规模文档数据库（包含12 024个文档）的检索，表明该算法具有较高的效率，优于现有的一般文档图像检索算法。

关键词 文档图像检索 文字区域提取 段落特征 多维数据检索结构

分类号 TP391

Document image retrieval method using combination of text and non-text features

ZHANG Tian

School of Information Science and Engineering，Shandong University，Jinan 250100，China

Abstract

An improved self-adaptive method for text area extraction is proposed.With it，the document image is segmented into text area and non-text area firstly.And then，for text area，local features and global features are extracted.The local features include gaps between connected characters，height and width of connected characters，and the global features contain writing style and paragraph features.For non-text area，the key block feature is extracted.After that，the retrieval method combines all the features to improve the accuracy.Meantime，multi-dimensional retrieval structure is introduced to improve the speed.The experiments performed on a large-scale document image database （including 12，024 images） reveal that the method is more efficient than existing ones.

Key words document image retrieval text area extraction paragraph feature multi-dimensional retrieval structure

DOI: 10.3778/j.issn.1002-8331.2010.12.002

通讯作者 张 田