(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106529521 A
(43)申请公布日 2017.03.22

(21)申请号 201610964844.7

(22)申请日 2016.10.31

(71)申请人 江苏文心古籍数字产业有限公司
地址 212000 江苏省镇江市京口区解放路17号

(72)发明人 王良君 褚正东 徐朝龙 王旭

(51)Int.Cl.
G06K 9/20(2006.01)
G06K 9/34(2006.01)

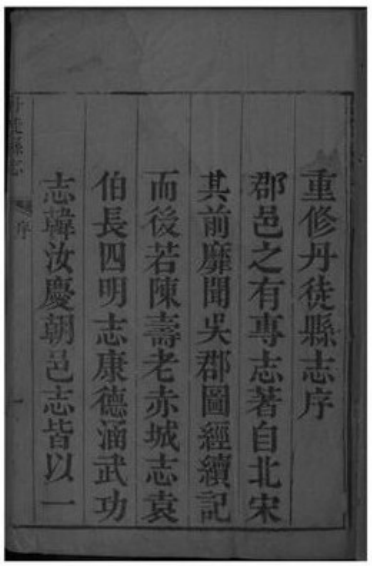
权利要求书1页 说明书3页 附图1页

(54)发明名称

一种古籍文字数字化录入方法

(57)摘要

本发明公开了一种古籍文字数字化录入方法，包括对古籍进行全文扫描，将扫描图像碎片化，利用古籍字体数据库对所述文字单元格进行自动识别，采用众包模式针对未自动识别成功的文字单元格在录入平台进行录入，并对录入结果进行比较，校检，完善古籍字体数据库，根据录入结果和坐标，还原位置，生成PDF文件。本发明可以提高图像数字化信息安全度和效率。



CN 106529521 A

1.一种古籍文字数字化录入方法,其特征在于,步骤包括:

步骤1、对古籍进行全文扫描,将扫描图像碎片化,先用OCR定位文字区域,再识别出文字区域的行数和列数,根据文字区域和行列数,确定每个文字的单元格;利用OCR对图像进行切割,根据灰度值阈值,依次判定文字单元格内像素点是否为有效像素点,并统计文字单元格内有效像素点数量,再根据文字像素点阈值,判定文字单元格内是否为文字;所述文字单元格指文字所在的矩形块;

步骤2、利用古籍字体数据库对所述文字单元格进行自动识别,当识别成功时则执行步骤5,否则执行步骤3;

步骤3、采用众包模式针对未自动识别成功的文字单元格在录入平台进行录入,并对录入结果进行比较,校检:由两名作业人员对同一文字单元格进行录入,比较两名作业人员的录入结果,当录入结果一致时执行步骤4,否则由第三名作业人员进行校验再执行步骤4;

步骤4、完善古籍字体数据库:根据古籍字体进行分类,将所述文字单元格及其对应的录入结果存入古籍数据库中,执行步骤5;

步骤5、根据录入结果和坐标,还原位置,生成PDF文件:

根据坐标,确定PDF上的单个文字矩形区域,将网上作业人员录入的文字存入PDF相应的位置;根据网上作业人员录入的文字数量,将图像坐标对应的矩形区域,切割成同等数量的区域,并将文字放入对应的位置。

2.根据权利要求1所述的一种古籍文字数字化录入方法,其特征在于,所述的步骤1还包括以下内容:

步骤1-1、根据文献对比度,设定灰度值阈值,正文灰度值均值在0-150的文献,灰度值阈值设定在100-150,正文灰度值均值在150-255的文献,灰度值阈值设定在150-200;当文献的灰度值小于灰度值阈值时,判定为有效像素点;

步骤1-2、根据文字单元格大小,设定文字像素点阈值,设定公式为$(w*h)/4*n^2$,四舍五入取整,其中w为文字单元格宽度,h为文字单元格高度,n为笔画粗度均值;

步骤1-3、统计文字单元格内有效像素点数量,当数量大于文字像素点阈值时,判定为有效文字;

步骤1-4、对于判定为有效文字的矩形块进行切割,并记录文字坐标;使用了图片裁剪工具类imgscalr,调用imgscalr提供的crop方法,根据坐标裁剪矩形块;

步骤1-5、完成全文图像的碎片化。

3.根据权利要求1一种古籍文字数字化录入方法,其特征在于,所述的步骤3还包括以下内容:

步骤3-1、作业人员的选择:发布测试稿件,测试合格人员方可进行作业;

步骤3-2、作业人员作业质量的控制:作业过程中会随机抽检作业人员的作业稿件,当抽检样正确率低于95%时,取消作业人员作业资格;作业完成后,系统会分析作业人员的正确率,低于95%时,取消作业人员作业资格。

一种古籍文字数字化录入方法

技术领域

[0001] 本发明涉及图像数字化领域，特别是一种古籍文字数字化录入方法。

背景技术

[0002] 传统的文字图像数字化，以古籍为例，先将古籍扫描成电子图像，然后由录入人员依照电子图像内容进行文字录入，最后对照原图进行排版，整理生成数字化文献，比如PDF文件，XML文件等。传统的图像数字化，存在一些弊端，录入人员可以看到整张古籍图像，信息安全度不高。驻厂人员进行文字录入，成本过高。对照原图手工排版，效率低下。

发明内容

[0003] 针对现有技术中存在的问题，本发明提供了一种可以提高图像数字化信息安全度和效率的古籍文字数字化录入方法，本发明结合互联网技术解决传统数字化面临的难题。

[0004] 本发明的目的通过以下技术方案实现。

[0005] 一种古籍文字数字化录入方法，步骤包括：

步骤1、对古籍进行全文扫描，将扫描图像碎片化，先用OCR定位文字区域，再识别出文字区域的行数和列数，根据文字区域和行列数，确定每个文字的单元格；利用OCR对图像进行切割，根据灰度值阈值，依次判定文字单元格内像素点是否为有效像素点，并统计文字单元格内有效像素点数量，再根据文字像素点阈值，判定文字单元格内是否为文字；所述文字单元格指文字所在的矩形块；

步骤2、利用古籍字体数据库对所述文字单元格进行自动识别，当识别成功时则执行步骤5，否则执行步骤3；

步骤3、采用众包模式针对未自动识别成功的文字单元格在录入平台进行录入，并对录入结果进行比较，校检：由两名作业人员对同一文字单元格进行录入，比较两名作业人员的录入结果，当录入结果一致时执行步骤4，否则由第三名作业人员进行校验再执行步骤4；

步骤4、完善古籍字体数据库：根据古籍字体进行分类，将所述文字单元格及其对应的录入结果存入古籍数据库中，执行步骤5；

步骤5、根据录入结果和坐标，还原位置，生成PDF文件：

根据坐标，确定PDF上的单个文字矩形区域，将网上作业人员录入的文字存入PDF相应的位置；根据网上作业人员录入的文字数量，将图像坐标对应的矩形区域，切割成同等数量的区域，并将文字放入对应的位置。

[0006] 进一步的，所述的步骤1还包括以下内容：

步骤1-1、根据文献对比度，设定灰度值阈值，正文灰度值均值在0-150的文献，灰度值阈值设定在100-150，正文灰度值均值在150-255的文献，灰度值阈值设定在150-200；当文献的灰度值小于灰度值阈值时，判定为有效像素点；

步骤1-2、根据文字单元格大小，设定文字像素点阈值，设定公式为$(w*h)/4*n^2$，四舍五入取整，其中w为文字单元格宽度，h为文字单元格高度，n为笔画粗度均值；

步骤1-3、统计文字单元格内有效像素点数量，当数量大于文字像素点阈值时，判定为有效文字；

步骤1-4、对于判定为有效文字的矩形块进行切割，并记录文字坐标；使用了图片裁剪工具类imgscalr，调用imgscalr提供的crop方法，根据坐标裁剪矩形块；

步骤1-5、完成全文图像的碎片化。

[0007] 进一步的，所述的步骤3还包括以下内容：

步骤3-1、作业人员的选择：发布测试稿件，测试合格人员方可进行作业；

步骤3-2、作业人员作业质量的控制：作业过程中会随机抽检作业人员的作业稿件，当抽检样正确率低于95%时，取消作业人员作业资格；作业完成后，系统会分析作业人员的正确率，低于95%时，取消作业人员作业资格。

[0008] 相比于现有技术，本发明的优点在于：本发明提高了图像数字化信息安全度和效率，结合互联网技术解决传统数字化面临的难题。将整张文献图片切割成一个个碎片，因为每个作业人员只能看到图像中的一个碎片块，对于提高信息安全度重要性不言而喻，尤其是一些重要资料的录入，如名片，银行票据等，对信息安全度要求较高。根据古籍字体数据库进行自动识别，避免了重复劳动，使得录入过程更加智能化，根据坐标自动还原位置，生成PDF，效率较高，位置也比较精确，解决了手工排版效率低下的难题。切割成单字后，大大降低了作业人员的技能要求，又采用众包模式，利用广大网民在互联网上进行生产作业，大大节省了生产成本的开支(人员、房租、交通，招聘、培训，解聘等)。采用众包模式，数以万计的网民同时在线作业，可以实现大规模的数字化生产。

附图说明

[0009] 图1为本发明的古籍文字碎片化示意图。

具体实施方式

[0010] 下面结合说明书附图和具体的实施例，对本发明作详细描述。

[0011] 一种古籍文字数字化录入方法，包括以下内容，

步骤1、将图像碎片化，利用OCR对图像进行切割，并记录碎片坐标：

古籍字符间距较窄，文字内容生僻，市面上流行的OCR软件对古籍的识别度普遍较低。本发明采用的OCR算法，是在传统OCR的基础上结合空间阈值算法，只进行切割，不进行识别；先用OCR定位文字区域，再识别出文字区域的行数和列数，根据文字区域和行列数，确定每个文字的单元格；根据灰度值阈值，依次判定文字单元格内像素点是否为有效像素点，并统计文字单元格内有效像素点数量，再根据文字像素点阈值，判定文字单元格内是否为文字；文字单元格指文字所在的矩形块。

[0012] 步骤1-1、根据文献对比度，设定灰度值阈值，有些文献在扫描时，存在反面文字透过来的情形，设定灰度阈值，就要是在保存正文的同时，尽可能的过滤掉这些躁点。一般正文颜色较深的文献(灰度值均值在0-150)，灰度值阈值设定比较低，设定在100-150，正文颜色较浅的文献(灰度值均值在150-255)，设定在150-200；如图1所示，像素点的灰度值阈值设置为150，当文献的灰度值小于灰度值阈值时，判定为有效像素点。

[0013] 步骤1-2、根据文字单元格大小，设定文字像素点阈值，设定公式为$(w*h)/4*n^2$，四

舍五入取整，其中w为文字单元格宽度，h为文字单元格高度，n为笔画粗度均值。例如文字单元格宽度为80px，高度为60px，笔画粗度均值为2px，则根据公式计算，设定文字像素点阈值为70。图1所示，文字像素点阈值设定为50。

[0014] 步骤1-3、统计文字单元格内有效像素点数量，当数量大于文字像素点阈值时，判定为有效文字。

[0015] 步骤1-4、对于判定为有效文字的矩形块进行切割，并记录文字坐标(文字所在矩形块左上角横坐标、纵坐标，矩形框高度，宽度)；这里使用了第三方图片裁剪工具类imgscalr，调用imgscalr提供的crop方法，根据坐标裁剪矩形块。

[0016] 步骤1-5、如图1所示，这样文字图像被切割成一张张文字图像碎片。

[0017] 步骤2、利用古籍字体数据库对所述文字单元格进行自动识别，当识别成功时则执行步骤5，否则执行步骤3；

步骤3、采用众包模式针对未自动识别成功的文字单元格在录入平台进行录入，并对录入结果进行比较，校检：由两名作业人员对同一文字单元格进行录入，比较两名作业人员的录入结果，当录入结果一致时执行步骤4，否则由第三名作业人员进行校验再执行步骤4；

步骤3-1、作业人员的选择，发布测试稿件，测试合格人员方可进行作业。

[0018] 步骤3-2、作业人员作业质量的控制，作业过程中会随机抽检作业人员的作业稿件，当抽检样正确率低于95%时，取消作业人员作业资格。作业完成后，系统会分析作业人员的正确率，低于95%时，取消作业人员作业资格。

[0019] 步骤4、完善古籍字体数据库：根据古籍字体进行分类，将所述文字单元格及其对应的录入结果存入古籍数据库中，执行步骤5；

步骤5、根据录入结果和坐标，还原位置，生成PDF文件：根据坐标，确定PDF上的单个文字矩形区域，将网上作业人员录入的文字存入PDF相应的位置；根据网上作业人员录入的文字数量，将图像坐标对应的矩形区域，切割成同等数量的区域，并将文字放入对应的位置。

[0020] 以上所述仅为本发明的优选实施例而已，并不限制于本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的权利要求范围之内。


重修丹徒縣志序
郡邑之有專志著自北宋
其前靡聞吳郡圖經續記
而後若陳壽老赤城志袁
伯長四明志康德涵武功
志攝汝慶朝邑志皆以一

郡

慶

图1