(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108334805 B
(45)授权公告日 2020.04.03

(21)申请号 201710134711.1

(22)申请日 2017.03.08

(65)同一申请的已公布的文献号
申请公布号 CN 108334805 A

(43)申请公布日 2018.07.27

(73)专利权人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 朱传聪

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 何平 邓云鹏

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/20(2006.01)

(56)对比文件
CN 101866418 A,2010.10.20,

审查员 姚雅倩

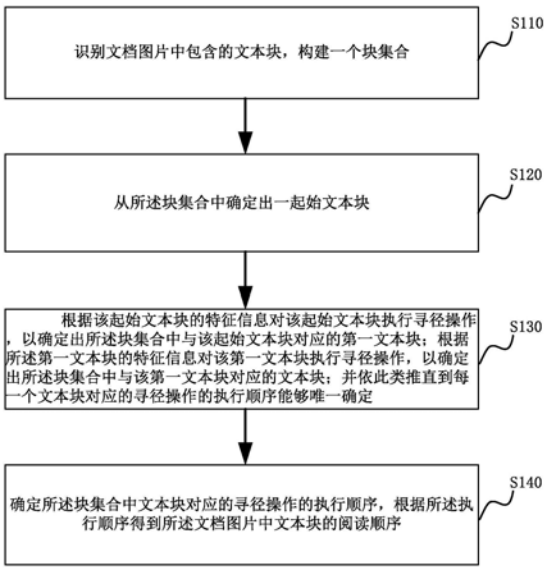
权利要求书4页 说明书13页 附图4页

(54)发明名称

检测文档阅读顺序的方法和装置

(57)摘要

本发明涉及检测文档阅读顺序的方法和装置。所述方法包括:识别文档图片中包含的文本块,构建一个块集合;从所述块集合中确定出一起始文本块;根据该起始文本块的特征信息对该起始文本块执行寻径操作,以确定出所述块集合中与该起始文本块对应的第一文本块;文本块的特征信息包括该文本块在文档图片中的位置信息以及该文本块的版面布局信息;依此类推直到所述块集合中每一个文本块对应的寻径操作的执行顺序能够唯一确定;确定所述块集合中文本块对应的寻径操作的执行顺序,根据所述执行顺序得到所述文档图片中文本块的阅读顺序。本发明能够准确识别各类文档图片的文档阅读顺序。



CN 108334805 B

1.一种检测文档阅读顺序的方法,其特征在于,包括:

识别文档图片中包含的文本块,构建一个块集合;

从所述块集合中确定出一起始文本块;

根据该起始文本块的特征信息对该起始文本块执行寻径操作,以确定出所述块集合中与该起始文本块对应的第一文本块;文本块的特征信息包括该文本块在文档图片中的位置信息以及该文本块的版面布局信息;

根据所述第一文本块的特征信息对该第一文本块执行寻径操作,以确定出所述块集合中与该第一文本块对应的文本块;并依此类推直到所述块集合中每一个文本块对应的寻径操作的执行顺序能够唯一确定;及

确定所述块集合中文本块对应的寻径操作的执行顺序,根据所述执行顺序得到所述文档图片中文本块的阅读顺序;

其中,所述寻径操作包括:

通过预先训练好的机器学习模型对所述文本块的特征信息进行学习,得出与该文本块对应的文本块的特征预测信息;

计算所述块集合中未执行寻径操作的各文本块的特征信息与所述特征预测信息的相关度;及

根据上述计算出的相关度确定出所述文本块对应的文本块。

2.根据权利要求1所述的检测文档阅读顺序的方法,其特征在于,所述从所述块集合中确定出一起始文本块,包括:

从所述块集合中选择出中心点坐标位于所述文档图片的一个顶点的文本块,并将该文本块确定为所述起始文本块。

3.根据权利要求1所述的检测文档阅读顺序的方法,其特征在于,从所述块集合中确定出一起始文本块,包括:

以所述文档图片的一个顶点为原点建立XOY坐标系,该XOY坐标系的x轴正方向指向所述文档图片的宽度方向,y轴正方向指向所述文档图片的长度方向;

从所述块集合中获取中心点的x坐标最小的一个文本块,作为文本块A;

获取中心点的y坐标小于所述文本块A的文本块,构建一个文本块集合G′;并依次将该集合G′中的每一个文本块B与所述文本块A进行对比;

若所述文本块B与该文本块A在x轴方向的投影不存在交集,则将所述文本块B从集合G′中删除;若所述文本块B与该文本块A在x轴方向的投影存在交集,则更新所述文本块A为所述文本块B,并将所述文本块B从集合G′中删除;

在每次文本块对比之后检测集合G′是否为空;若是,则将当前的文本块A确定为起始文本块;若否,则在所述文本块A发生更新时更新集合G′,并将更新后的集合G′中的每一个文本块与当前的文本块A进行上述对比;依次类推直到集合G′为空。

4.根据权利要求1所述的检测文档阅读顺序的方法,其特征在于,还包括:

预先训练机器学习模型,使得训练之后的机器学习模型输出的特征预测信息与对应的样本信息的欧式距离满足设定条件。

5.根据权利要求4所述的检测文档阅读顺序的方法,其特征在于,预先训练机器学习模型,包括:

建立样本库，所述样本库中的信息包含：样本块的集合，该集合中每个样本块在先后各次训练中的顺序状态，以及训练需确定的状态变化序列；若所述样本块的集合中样本块的总数为n，则训练需确定的状态变化序列为n-2个，且每个状态变化序列中的信息包括：当前参与训练的样本块，所述样本块的集合中每个样本块的当前顺序状态，以及所述样本块的集合中每个样本块的下一顺序状态；

依次采用各个状态变化序列对机器学习模型进行训练；当n-2个状态变化序列均参与训练之后，保存所述机器学习模型中的参数。

6.根据权利要求5所述的检测文档阅读顺序的方法，其特征在于，采用第k个状态变化序列对机器学习模型进行训练，包括：

将所述样本块的集合中第k个样本块R_k的特征信息输入机器学习模型，获取机器学习模型输出的所述样本块R_k对应的文本块的特征预测信息O_k，$k\in[1,n-2]$；

根据所述样本块的集合中每个样本块在所述样本块R_k参与训练时的顺序状态，获取其中阅读顺序未确定的样本块，得到集合G^*；

将所述集合G^*中各样本块的特征信息分别与O_k进行点积运算，得到集合V^*；

获取所述集合G^*中各样本块在第k+1个样本块参与训练时的顺序状态，得到集合V^{π}；

对集合V^*进行归一化处理得到集合V^{**}，对集合V^{π}进行归一化处理得到集合$V^{\pi\pi}$；根据集合V^{**}和集合$V^{\pi\pi}$构建所述样本块R_k参与训练时对应的损失函数，基于该损失函数通过BP算法更新所述机器学习模型中的参数。

7.根据权利要求1所述的检测文档阅读顺序的方法，其特征在于，

文本块在文档图片中的位置信息包括：文本块的中心点在文档图片中的x坐标，文本块的中心点在文档图片中的y坐标；

文本块的版面布局信息包括：文本块的宽度，文本块的高度，文本块中所有连通区域的尺度均值以及文本块的密度信息；

所述机器学习模型为6维输入且6维输出的神经网络模型。

8.根据权利要求7所述的检测文档阅读顺序的方法，其特征在于，所述神经网络模型包括6维输入层、6维输出层、第一隐层以及第二隐层，所述第一隐层、第二隐层分别为12维和20维的隐层。

9.根据权利要求1至8任一所述的检测文档阅读顺序的方法，其特征在于，识别文档图片中包含的文本块，包括：

对所述文档图片进行二值化处理和方向校正处理；

对经过二值化处理及方向校正处理的文档图片进行版面分析，得到文档图片中包括的文本块。

10.根据权利要求1至8任一所述的检测文档阅读顺序的方法，其特征在于，还包括：

对各个文本块进行文本识别，并按照所述确定出的阅读顺序得到所述文档图片的文本信息。

11.一种检测文档阅读顺序的装置，其特征在于，包括：

块识别模块，用于识别文档图片中包含的文本块，构建一个块集合；

起始块选择模块，用于从所述块集合中确定出一起始文本块；

自动寻径模块，用于根据该起始文本块的特征信息对该起始文本块执行寻径操作，以

确定出所述块集合中与该起始文本块对应的第一文本块;文本块的特征信息包括该文本块在文档图片中的位置信息以及该文本块的版面布局信息;根据所述第一文本块的特征信息对该第一文本块执行寻径操作,以确定出所述块集合中与该第一文本块对应的文本块;并依此类推直到所述块集合中每一个文本块对应的寻径操作的执行顺序能够唯一确定;及

顺序确定模块,用于确定所述块集合中文本块对应的寻径操作的执行顺序,根据所述执行顺序得到所述文档图片中文本块的阅读顺序;

其中,所述自动寻径模块在对文本块执行寻径操作时,通过预先训练好的机器学习模型对所述文本块的特征信息进行学习,得出与该文本块对应的文本块的特征预测信息;计算所述块集合中未执行寻径操作的各文本块的特征信息与所述特征预测信息的相关度;以及根据上述计算出的相关度确定出所述文本块对应的文本块。

12.根据权利要求11所述的检测文档阅读顺序的装置,其特征在于,所述起始块选择模块,用于从所述块集合中选择出中心点坐标位于所述文档图片的一个顶点的文本块,并将该文本块确定为所述起始文本块。

13.根据权利要求11所述的检测文档阅读顺序的装置,其特征在于,所述起始块选择模块,用于

以所述文档图片的一个顶点为原点建立XOY坐标系,该XOY坐标系的x轴正方向指向所述文档图片的宽度方向,y轴正方向指向所述文档图片的长度方向;

从所述块集合中获取中心点的x坐标最小的一个文本块,作为文本块A;

获取中心点的y坐标小于所述文本块A的文本块,构建一个文本块集合G′;并依次将该集合G′中的每一个文本块B与所述文本块A进行对比;

若所述文本块B与该文本块A在x轴方向的投影不存在交集,则将所述文本块B从集合G′中删除;若所述文本块B与该文本块A在x轴方向的投影存在交集,则更新所述文本块A为所述文本块B,并将所述文本块B从集合G′中删除;

在每次文本块对比之后检测集合G′是否为空;若是,则将当前的文本块A确定为起始文本块;若否,则在所述文本块A发生更新时更新集合G′,并将更新后的集合G′中的每一个文本块与当前的文本块A进行上述对比;依次类推直到集合G′为空。

14.根据权利要求11所述的检测文档阅读顺序的装置,其特征在于,还包括:

训练模块,用于预先训练机器学习模型,使得训练之后的机器学习模型输出的特征预测信息与对应的样本信息的欧式距离满足设定条件。

15.根据权利要求14所述的检测文档阅读顺序的装置,其特征在于,所述训练模块,包括:

样本库构建子模块,用于建立样本库,所述样本库中的信息包含:样本块的集合,该集合中每个样本块在先后各次训练中的顺序状态,以及训练需确定的状态变化序列;若所述样本块的集合中样本块的总数为n,则训练需确定的状态变化序列为n-2个,且每个状态变化序列中的信息包括:当前参与训练的样本块,所述样本块的集合中每个样本块的当前顺序状态,以及所述样本块的集合中每个样本块的下一顺序状态;

以及,训练子模块,用于依次采用各个状态变化序列对机器学习模型进行训练;当n-2个状态变化序列均参与训练之后,保存所述机器学习模型中的参数。

16.根据权利要求15所述的检测文档阅读顺序的装置,其特征在于,

所述训练子模块在采用第k个状态变化序列对机器学习模型进行训练时，将所述样本块的集合中第k个样本块R_k的特征信息输入机器学习模型，获取机器学习模型输出的所述样本块R_k对应的文本块的特征预测信息O_k，$k\in[1,n-2]$；

根据所述样本块的集合中每个样本块在所述样本块R_k参与训练时的顺序状态，获取其中阅读顺序未确定的样本块，得到集合G^*；

将所述集合G^*中各样本块的特征信息分别与O_k进行点积运算，得到集合V^*；

获取所述集合G^*中各样本块在第k+1个样本块参与训练时的顺序状态，得到集合V^{π}；

对集合V^*进行归一化处理得到集合V^{**}，对集合V^{π}进行归一化处理得到集合$V^{\pi\pi}$；根据集合V^{**}和集合$V^{\pi\pi}$构建所述样本块R_k参与训练时对应的损失函数，基于该损失函数通过BP算法更新所述机器学习模型中的参数。

17.根据权利要求11所述的检测文档阅读顺序的装置，其特征在于，

所述块识别模块，还用于获取各文本块的特征信息，包括：文本块的中心点在文档图片中的x坐标，文本块的中心点在文档图片中的y坐标，文本块的宽度，文本块的高度，文本块中所有连通区域的尺度均值以及文本块的密度信息；

所述机器学习模型为6维输入且6维输出的神经网络模型。

18.根据权利要求11至17任一所述的检测文档阅读顺序的装置，其特征在于，所述块识别模块，包括：

预处理子模块，用于对所述文档图片进行二值化处理和方向校正处理；

以及，版面识别子模块，用于对经过二值化处理及方向校正处理的文档图片进行版面分析，得到文档图片中包含的文本块。

19.根据权利要求11至17任一所述的检测文档阅读顺序的装置，其特征在于，还包括：

文本识别模块，用于对各个文本块进行文本识别，并按照所述确定出的阅读顺序得到所述文档图片的文本信息。

20.一种存储介质，其上存储有计算机程序，其特征在于，所述程序被处理器执行时可实现如权利要求1至10中任一项所述的检测文档阅读顺序的方法。

21.一种终端设备，包括存储介质，处理器及存储在存储介质上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如权利要求1至10中任一项所述的检测文档阅读顺序的方法。

检测文档阅读顺序的方法和装置

技术领域

[0001] 本发明涉及计算机技术领域,特别是涉及检测文档阅读顺序的方法和装置。

背景技术

[0002] OCR(Optical Character Recognition光学字符识别),是描述文档图片识别 的一类算法,其是针对印刷体字符,采用光学的方式将纸质文档中的文字转换 成为黑白点阵的图像文件,并通过识别软件将图像中的文字转换成文本格式, 供文字处理软件进一步编辑加工的技术。

[0003] 在OCR技术中,普遍采用基于有向图、固定规则、语义分析等方法来识别 文档的阅读顺序,然而这些方法在复杂环境下或者对于复杂文档图片来说,其 阅读顺序的识别错误率较高,存在识别性能不稳定的问题。

发明内容

[0004] 本发明实施例提供了检测文档阅读顺序的方法和装置,能够准确识别各类 文档图片的文档阅读顺序。

[0005] 本发明一方面提供检测文档阅读顺序的方法,包括:

[0006] 识别文档图片中包含的文本块,构建一个块集合;

[0007] 从所述块集合中确定出一起始文本块;

[0008] 根据该起始文本块的特征信息对该起始文本块执行寻径操作,以确定出所 述块集合中与该起始文本块对应的第一文本块;文本块的特征信息包括该文本 块在文档图片中的位置信息以及该文本块的版面布局信息;

[0009] 根据所述第一文本块的特征信息对该第一文本块执行寻径操作,以确定出 所述块集合中与该第一文本块对应的文本块;并依此类推直到所述块集合中每 一个文本块对应的寻径操作的执行顺序能够唯一确定;及

[0010] 确定所述块集合中文本块对应的寻径操作的执行顺序,根据所述执行顺序 得到所述文档图片中文本块的阅读顺序。

[0011] 本发明另一方面提供一种检测文档阅读顺序的装置,包括:

[0012] 块识别模块,用于识别文档图片中包含的文本块,构建一个块集合;

[0013] 起始块选择模块,用于从所述块集合中确定出一起始文本块;

[0014] 自动寻径模块,用于根据该起始文本块的特征信息对该起始文本块执行寻 径操作,以确定出所述块集合中与该起始文本块对应的第一文本块;文本块的 特征信息包括该文本块在文档图片中的位置信息以及该文本块的版面布局信 息;根据所述第一文本块的特征信息对该第一文本块执行寻径操作,以确定出 所述块集合中与该第一文本块对应的文本块;并依此类推直到所述块集合中每 一个文本块对应的寻径操作的执行顺序能够唯一确定;及

[0015] 顺序确定模块,用于确定所述块集合中文本块对应的寻径操作的执行顺序, 根据

所述执行顺序得到所述文档图片中文本块的阅读顺序。

[0016] 基于上述实施例提供的检测文档阅读顺序的方法和装置，首先识别文档图 片中包含的文本块，构建一个块集合；从块集合中确定出一起始文本块；从起 始文本块开始寻径，根据文本块的位置信息以及版面布局信息决定下一步应该 走到哪一个文本块，依次类推得出文档图片包含的全部文本块的阅读顺序。该 方案能够兼容多种场景，对文档图片的尺寸、噪声、样式具有更好的鲁棒性，因此能够准确识别各类文档图片对应的文档阅读顺序。

附图说明

[0017] 图1为一个实施例中的本发明方案的工作环境示意图；

[0018] 图2为一实施例的检测文档阅读顺序的方法的示意性流程图；

[0019] 图3为一实施例的文档图片包含的文本块示意图；

[0020] 图4为一实施例的神经网络模型的示意图；

[0021] 图5为一实施例的根据训练样本训练神经网络模型的示意流程图；

[0022] 图6为一实施例的检测文档阅读顺序的装置的示意性结构图；

[0023] 图7为另一实施例的检测文档阅读顺序的装置的示意性结构图。

具体实施方式

[0024] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实 施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅 仅用以解释本发明，并不用于限定本发明。

[0025] 图1为一个实施例中的本发明方案的工作环境示意图；实现本发明实施例 的检测文档阅读顺序的方法的工作环境为设置有OCR系统的智能终端，并且所 述智能终端至少还包括通过系统总线连接的处理器、显示模组、电源接口和存 储介质，所述智能终端通过OCR系统将文档图片中包含的文本信息识别并显示 出来。其中，显示模组可对OCR系统识别出的文本信息进行显示；电源接口用 于与外部电源连接，外部电源通过该电源接口向智能终端电池供电；所述存储 介质中至少存储有操作系统、OCR系统、数据库以及一种检测文档阅读顺序的 装置，该装置可用于实现本发明实施例的检测文档阅读顺序的方法。所述智能终端可以为手机、平板电脑等，也可以是其他具有上述结构的设备。

[0026] 结合图1及上述对工作环境的说明，以下对检测文档阅读顺序的方法的实 施例进行说明。

[0027] 图2为一实施例的检测文档阅读顺序的方法的示意性流程图；如图2所示， 本实施例中的检测文档阅读顺序的方法包括步骤：

[0028] S110，识别文档图片中包含的文本块，构建一个块集合；

[0029] 本实施例中，可先对文档图片进行二值化处理，得到二值化文档图片，在 二值化文档图片中，各个像素点的值均用0或者1表示。然后基于二值化文档 图片进行尺度分析和版面分析，得出文档包含的全部文本块。其中的尺度分析 是指寻找二值化文档图片中每个字符的尺度信息，尺度以像素为单位，其值为 字符所占用的矩形区域面积的平方根。版面分析是指在OCR中，将文档图片中 的内容按照段落、分页等信息划分为多个不重叠的区域

的算法。由此可得出文 档中包含的全部文本块，例如图3所示或者图5所示。

[0030] 在另一优选实施例中，对文档图片进行预处理的过程中，还包括对校正文 档图片的步骤。即若待检测的文档图片的初始状态相对于预设的标准状态存在 偏差时，校正所述文档图片使其符合所述标准状态。例如：若检测到文档图片 的初始状态下存在倾斜、上下颠倒等情况，则需先对所述文档图片的方向进行 校正。

[0031] S120，从全部文本块中(即所述块集合中)确定出一起始文本块。

[0032] 通常情况下，人们在阅读文档时会从文档的一顶点(例如左上角)开始进 行阅读，基于此，在一优选实施例中，可从所述块集合中选择出中心点坐标位 于所述文档图片的一个顶点的文本块，并将该文本块确定为所述起始文本块。例如：将位于文档图片的左侧且最上方的一文本块确定为起始文本块，如图3 中所示的文本块R_1，或者图5中所示的文本块R_1。

[0033] 可以理解的，在其他实施例中，对于不同的文档和实际的阅读习惯(例如 从右到左排版的文档)，也可将其他文本块确定为起始文本块。

[0034] S130，从起始文本块开始寻径；根据该起始文本块的特征信息对该起始文 本块执行寻径操作，以确定出所述块集合中与该起始文本块对应的第一文本块；根据所述第一文本块的特征信息对该第一文本块执行寻径操作，以确定出所述 块集合中与该第一文本块对应的文本块；并依此类推直到所述块集合中每一个 文本块对应的寻径操作的执行顺序能够唯一确定。

[0035] 其中，文本块的特征信息包括该文本块在文档图片中的位置信息以及该文 本块的版面布局信息。

[0036] 对文本块进行寻径操作实际上是基于该文本块的特征信息得出其对应的下 一文本块的特征预测信息。在一实施例中，对文本块的寻径操作包括：通过预 先训练好的机器学习模型对所述文本块的特征信息进行学习，得出与该文本块 对应的文本块的特征预测信息；计算所述块集合中未执行寻径操作的各文本块 的特征信息与所述特征预测信息的相关度；然后根据上述计算出的相关度确定 出所述文本块对应的文本块。

[0037] 本实施例中，步骤S130即是一个自起始文本块起，对文档包含的文本块进 行自动寻径的过程，每次寻径只需确定当前文本块对应的下一文本块。例如图3 所示的文档图片，当前文本块为R_1，通过本次寻径可确定文本块R_1的下一文本块 为R_2；然后将R_2作为当前文本再次进行寻径，得到R_2的下一文本块为R_4；以此 类推，直到对R_6执行完寻径操作，并确定出R_6对应的下一文本块为R_7，虽然此 时R_7和R_8未执行寻径操作，但由于已经确定出R_6对应的下一文本块为R_7，因此R_7和R_8对应的寻径操作的执行顺序已经能够唯一确定(即先R_7后R_8)。通过上述自 动寻径方式，对文档图片的尺寸、样式具有更好的鲁棒性。并且自动寻径的依据是基于文本块之间位置以及版面布局信息的相关性，因此能够较好的克服图 片噪声或者识别环境对检测结果的影响，有利于保证检测结果的准确性。

[0038] 本实施例中，预先通过合适的训练样本对所述机器学习模型进行训练，可 使得所述机器学习模型输出较为准确的预测结果，然后基于相关性可确定出准 确的下一文本块，适用于各种混合文档类型的文档阅读顺序检测。其中，所述 机器学习模型可以为神经网络模型，也可以为其他非神经网络的概率模型。

[0039] S140，确定所述块集合中文本块对应的寻径操作的执行顺序，根据所述执 行顺序

得到所述文档图片中文本块的阅读顺序。

[0040] 通过步骤S130的自动寻径，可得到每一个文本块及其对应的下一文本块，当自动寻径结束时，根据所有文本块以及各文本块对应的下一文本块，便可得 到全部文本块的阅读顺序。例如在自动寻径结束后，可得到图3所示的文档图 片中文本块的阅读顺序为$R_1 \rightarrow R_2 \rightarrow R_4 \rightarrow R_5 \rightarrow R_3 \rightarrow R_6 \rightarrow R_7 \rightarrow R_8$。

[0041] 基于上述实施例的检测文档阅读顺序的方法，首先识别文档图片中包含的 全部文本块；从全部文本块中确定出一起始文本块，从起始文本块开始寻径，根据文本块在文档图片中的位置信息以及该文本块的版面布局信息决定下一步 应该走到哪个文本块区域，直到得出全部文本块的阅读顺序。由此能够兼容多 种场景，对文档图片的尺寸、噪声、样式具有更好的鲁棒性，因此能够准确识 别各类文档图片对应的文档阅读顺序。

[0042] 在一优选实施例中，所述机器学习模块中包含多个参数，在所述检测文档 阅读顺序的方法中，还包括对所述机器学习模型进行训练的步骤，以使得训练 之后的机器学习模型输出的特征预测信息与对应的样本信息的欧式距离满足设 定条件。欧式距离指的是欧几里得度量，表示两个相同维度向量的空间距离。

[0043] 在一优选实施例中，对机器学习模块进行训练的方式可包括如下过程：

[0044] 首先，获取训练样本。样本是指在机器学习过程中，已经标定好了的数据，包括输入数据和输出数据。本实施例中训练样本即参与机器学习模块训练的若 干样本块，且所述若干样本块的阅读顺序为已知的。

[0045] 然后，基于训练样本建立对应的样本库$M=\{G,S,T\}$。其中G表示样本块的集 合，S表示样本块在先后各次训练中的顺序状态的集合，T表示训练过程中需确 定的状态变化序列。若G中样本块的总数为n，则有，

[0046] $S=\{s_i;i\in[1,n],s_i\in[0,n]\}$；

[0047] $T=\{\{R_1,S_1,S_2\},\{R_2,S_2,S_3\},...\{R_{n-2},S_{n-2},S_{n-1}\}\}$；

[0048] 若$s_i=0$表示样本块R_i的阅读顺序未确定(即执行寻径操作的顺序未确定)，若$s_i>0$表示样本块R_i的阅读顺序已确定(即执行寻径操作的顺序已确定)，且 阅读顺序为s_i的值，表示为$S(R_i)=s_i$。上述T中的每一个序列中的各项分别表 示当前参与训练的样本块、G中每个样本块当前的顺序状态的集合和需预测出的 G中每个样本块的下一顺序状态的集合。具体的，以$\{R_2,S_2,S_3\}$序列为例，R_2表示 当前参与训练的样本块为R_2，S_2表示R_2参与训练时G中各个样本块对应的顺序状 态，S_3表示采用R_2参与训练时需预测出的G中每个样本块的下一个顺序状态。其 中，由于剩余的最后两个样本块可采用排除法直接确定出来，因此其不需要训 练，故在T中只需包括n-2个序列。

[0049] 然后，基于上述的样本库$M=\{G,S,T\}$，依次采用T中的各个状态变化序列对 机器学习模型进行训练；当T中的所有状态变化序列均参与训练之后，保存所述 机器学习模型中的参数。

[0050] 在一优选实施例中，根据T中的第k个序列$\{R_k,S_k,S_{k+1}\}$对机器学习模型中的 参数进行训练的具体实施方式可包括如下步骤1～步骤5：

[0051] 步骤1，将样本块R_k的特征信息输入机器学习模型，获取机器学习模型输出 的R_k的下一文本块的特征预测信息O_k，$k\in[1,n-2]$；

[0052] 步骤2，获取S_k中顺序状态为0的样本块R_i，得到集合G^*：

[0053] $G^{*}=\{R_i;S_k(R_i)=0\};i\in[1,n]$;

[0054] 集合G^{*}的维度为n-k;.

[0055] 步骤3,将G^{*}中各项分别与O_k进行点积运算,得到集合$V^{*}=\{v_i=R_i\cdot O_k\}$;

[0056] 步骤4,获取G^{*}中各样本块R_i在S_{k+1}中对应的顺序状态,得到集合 $V^{\pi}=\{v_i'=S_{k+1}(R_i)\}$;集合$V^{\pi}$的维度与集合$G^{*}$的维度相等.

[0057] 步骤5,对V^{*}进行归一化处理可得到$V^{**}=\{\hat{v}_i=v_i/sum(V^{*})\}$,对$V^{\pi}$进行归一 化处理得到集合$V^{\pi\pi}=\{v_i''=v_i'/sum(V^{\pi})\}$;根据$V^{**}$和$V^{\pi\pi}$构建所述样本块$R_k$参与训 练时对应的损失函数loss,基于该损失函数通过BP算法更新所述机器学习模型 中的参数。其中所述损失函数loss为:

[0058] $$loss=\left|V^{**}-V^{\pi\pi}\right|=\sum_i(\hat{v}_i-v_i'')^2/2。$$

[0059] 本实施例中,损失函数是指在机器学习过程中,通过机器学习计算所得到 的误差,误差可以使用多种函数进行度量,且该函数一般为凸函数。即根据V^{**}和 $V^{\pi\pi}$的欧式距离构建所述样本块R_k参与训练时对应的损失函数。欧式距离即欧几 里得度量,表示两个多为维向量的空间距离。通过每次学习过程中得到的损失 函数,使用BP算法对机器学习模型的参数进行调整,当损失函数收敛到一定程 度时,机器学习模型的输出准确度也会提高到某个程度。其中BP算法即误差反 向传播算法(Error Back Propagation),尤其适用于多层前馈网络模型的训练, 是指在训练过程中误差会累积到输出层,然后通过输出层将误差反向传递到每 一个前馈网络层,从而达到调节各前馈网络层参数的目的。

[0060] 在一优选实施例中,为了准确的对各个文本块的特征信息进行学习,对识 别出的文本块采用文本框进行标记,并将每个文本块的特征信息用特征向量的 形式表示为:

[0061] R={x,y,w,h,s,d};

[0062] R表示文本块的特征向量,包含6个特征信息;x表示文本块的中心点的x 坐标;y表示文本块的中心点的y坐标;w表示文本块的宽度;h表示文本块的 高度;s表示文本块中所有连通区域的尺度均值;d表示文本块的密度信息。所 述连通区域是指在二值化图像中,能够通过像素之间的连接形成的区域;像素 之间的连接有4邻域和8邻域算法,例如8邻域连通算法,即在(x,y)位置的像素 点,如果与其相邻的8个点中的某一个与(x,y)的像素值相同,则两者是8邻域连 通的,递归查找所有连通的点,这些点的集合即为一个连通区域。

[0063] 其中,

[0064] $$s=\sum_i^K\sqrt{(W(r_i))^2+(H(r_i))^2}/K;$$

[0065] $$d=\sum_i^w\sum_j^h p_{wi,hj}/(wh);$$

[0066] W、H分别表示取长度和取宽度的函数,r_i为连通区域i,K表示文本块中包 含的连通区域的总量;p表示像素点的像素值。

[0067] 在一优选实施例中,在识别文档图片中包含的文本块之后,还包括获取各 文本块的特征向量R={x,y,w,h,s,d}的步骤。为了让机器学习的模型对尺度信息 不敏感,进一步将文本块的对应特征信息做归一化处理,例如约定:

[0068] w＝1.0;h＝1.0;max(p)＝1.0。

[0069] 在一优选实施例中,从全部文本块中确定出一起始文本块的方式可包括:

[0070] 以文档图片左上角顶点为原点建立XOY坐标系(参考图3、图5所示),并 且该XOY坐标系的x轴正方向指向文档图片的宽度方向,y轴正方向指向文档 图片的长度方向。首先,从所述块集合中获取中心点的x坐标最小的一个文本 块,作为文本块A。然后,获取中心点的y坐标小于所述文本块A的文本块, 构建一个文本块集合G′;并依次将该集合G′中的每一个文本块B与所述文本块A 进行对比;若所述文本块B与该文本块A在x轴方向的投影不存在交集,则将 所述文本块B从集合G′中删除;若所述文本块B与该文本块A在x轴方向的投 影存在交集,则更新所述文本块A为所述文本块B,并将所述文本块B从集合G′ 中删除。在每次文本块对比之后检测集合G′是否为空;若是,则将当前的文本 块A确定为起始文本块;若否,则在所述文本块A发生更新时更新集合G′,并 将更新后的集合G′中的每一个文本块与当前的文本块A进行上述对比;依次类 推直到集合G′为空。本实施例的起始文本块的确定方法,适用于各类复杂的文 档,并能准确识别出起始文本块。

[0071] 在一优选实施例中,假设将每个文本块的特征向量表示为 $R=\{r_1,r_2,r_3,r_4,r_5,r_6\}=\{x,y,w,h,s,d\}$,简记为$R=\{r_j;j\in[0,6)\}$,$r_j$为样本块的特征信息 j。所述机器学习模型选为神经网络模型。对应的,如图4所示,所述神经网络 模型可包括6维输入层、6维输出层、第一隐层以及第二隐层。在神经网络模型 中,输入层负责接收输入及分发到隐层(因为用户看不见这些层,所以见做隐 层),隐层负责所需的计算及输出结果给输出层,而用户则可以看到最终结果。

[0072] 优先的,所述第一隐层、第二隐层分别为12维和20维的隐层。将所述 $R=\{r_j;j\in[0,6)\}$输入所述神经网络模型,则所述第一隐层的输出为K_1:

[0073] $$K_1=\{k_{1i}=\sum_i a_{1i}r_j+b_{1i};i\in[0,12),j\in[0,6)\}$$;

[0074] 所述第二隐层的输出为K_2:

[0075] $$K_2=\{k_{2m}=\sum_i a_{2m}k_{1i}+b_{2m};m\in[0,20),i\in[0,12)\}$$;

[0076] 所述6维输出层的输出为O:

[0077] $O=\{o_n=\mathrm{sigmoid}\sum a_{on}k_{2m}+b_{on};n\in[0,6),m\in[0,20)\}$;

[0078] 其中a_{1i}、b_{1i}为第一隐层对应的参数,k_{1i}为第一隐层的第i维输出;a_{2m}、b_{2m}为第二隐层对应的参数,k_{2m}为第二隐层的第m维输出;a_{on}、b_{on}为6维输出层 对应的参数,o_n为第n维输出,Sigmoid表示S型的非线性函数。

[0079] 对于上述的神经网络模型的训练,以图5中的文本块为例,将图5中的文 本块作为样本块进行所述神经网络模型的训练,样本块包括R_1,R_2,R_3,R_4以及R_5, 可分别表示为:

[0080] $R_1=\{x_1,y_1,w_1,h_1,s_1,d_1\}$

[0081] $R_2=\{x_2,y_2,w_2,h_2,s_2,d_2\}$;

[0082] $R_3=\{x_3,y_3,w_3,h_3,s_3,d_3\}$;

[0083] $R_4=\{x_4,y_4,w_4,h_4,s_4,d_4\}$;

[0084] $R_5=\{x_5,y_5,w_5,h_5,s_5,d_5\}$;

[0085] 且已知R_1,R_2,R_3,R_4,R_5的正确阅读顺序为$R_1\rightarrow R_3\rightarrow R_2\rightarrow R_4\rightarrow R_5$。

[0086] 根据所述训练样本，设定每个样本块的当前顺序状态的集合为 $S=\{s_i;i\in[1,5],s_i\in[0,5]\}$，其中当$s_i=0$时表示对应的文本块$R_i$还未确定执行 寻径操作的顺序(即$R_i$的阅读顺序未确定)，$s_i>0$表示对应的文本块$R_i$已确定 执行寻径操作的顺序(即$R_i$的阅读顺序已确定)，且确定执行寻径操作的顺序为 s_i的值，表示为$S(R_i)=s_i$。因此所述训练样本在训练过程中对应的阅读状态可 包括：

[0087] $S_0=(0,0,0,0,0)$；

[0088] $S_1=(1,0,0,0,0)$；

[0089] $S_2=(1,0,2,0,0)$；

[0090] $S_3=(1,3,2,0,0)$；

[0091] $S_4=(1,3,2,4,0)$；

[0092] $S_5=(1,3,2,4,5)$；

[0093] 进一步的，所述训练样本R_1,R_2,R_3,R_4,R_5还可描述为以下状态序列：

[0094] $\{R_1,S_1,S_2\},\{R_3,S_2,S_3\},\{R_2,S_3,S_4\},\{R_4,S_4,S_5\}$；

[0095] 其中由于$\{R_4,S_4,S_5\}$序列可以直接确定出来，因此其不需要训练，因此在样 本库中，$T=\{\{R_1,S_1,S_2\},\{R_3,S_2,S_3\},\{R_2,S_3,S_4\}\}$。基于所述样本库，首先采用 $\{R_1,S_1,S_2\}$序列进行所述神经网络模型的训练，过程如下：

[0096] 将R_1输入到神经网络模型中，获取神经网络模型输出的下一阅读状态的预 测信息O_1。选取S_1中值为0所对应的样本块，可得到集合$G^*=\{R_2,R_3,R_4,R_5\}$。将 集合G^*中的各项分别与O_1进行点积，可得到$V^*=\{v_2,v_3,v_4,v_5\}$，归一化后得到 $V^{**}=\{\hat{v}_2,\hat{v}_3,\hat{v}_4,\hat{v}_5\}$。

[0097] 获取G^*中各项在S_2中对应的状态值，可得到集合V^{π}:

[0098] $V^{\pi}=\{v_2',v_3',v_4',v_5'\}=\{0,2,0,0\}$；

[0099] 归一化处理可得到$V^{\pi\pi}=\{v_2'',v_3'',v_4'',v_5''\}=\{0,1,0,0\}$。

[0100] 根据集合V^{**}和集合$V^{\pi\pi}$可构建样本块R_1参与训练时对应的损失函数：

[0101] $$loss=\sum_{i=2}^{5}(\hat{v}_i-v_i'')^2/2$$；

[0102] 通过BP算法可更新所述神经网络模型中的所有参数。

[0103] 按照上述步骤继续训练，即根据序列$\{R_3,S_2,S_3\},\{R_2,S_3,S_4\}$也按照上述步骤继 续训练，由此可完成所述神经网络模型的训练。本实施例中，通过选取适当的 训练样本，可得到性能稳定的神经网络模型；基于训练后的神经网络模型进行 文本块寻径，可准确得到当前文本块的下一文本块，有利于准确检测出各类型 文档图片中的文档阅读顺序。

[0104] 本发明上述实施例的检测文档阅读顺序的方法，可应用于OCR系统中自动 文档分析模块，所述自动文档分析模块在识别出文档图片包含的文本块之后，对识别出的文本块进行排序，然后将文本块的阅读顺序输出给文本识别模块，在文本识别模块中进行文本识别后，基于已经得到的阅读顺序，整理成最终的 可阅读文档，从而进行自动分析和存储。具体的，所述自动文档分析模块在对 文本块进行排序时，涉及信息处理过程包括：

[0105] 设定选择算法$A=A(R,S)$，该算法根据当前文本块R和当前的阅读顺序的状 态S，推导出下一个阅读顺序的状态S，可以表示为：

[0106] $$S_0\xrightarrow[a_0]{}S_1\xrightarrow[a_1]{}S_2\cdots\xrightarrow[a_{n-1}]{}S_n$$

[0107] 其中$S_0=\{s_i=0;i\in[1,n]\}$,$S_n=\{s_i=i;i\in[1,n]\}$,n表示文档图片包含的 文本块的总数。

[0108] 进一步的,所述算法A可分成三个部分:

[0109] 1)R_{start}选择器Ψ_1

[0110] Ψ_1用于对起始文本块进行选择,起始文本块用R_{start}标记。在所有的文本块R 中,选取中心点坐标位于文档图片最左边的一个R,标记为R_1,然后对剩余的R 相对于R_1进行计算,选取$y(R)<y(R_1)$的文本块构建集合G′,优先的,还可对G′中 的R按照y坐标降序排列,然后按照顺序将G′中的每一个R与R_1进行对比,如 果R与R_1在x轴方向的投影有交集,则将此R标记为R_1,将所述R从G′中删除; 否则,不更新R_1,直接将此R从G′中删除;重复上述动作,直到G′为空,可确 定$R_{start}=R_1$。

[0111] 在一优选实施例中,每次在将新的R标记为R_1,将所述R从G′中删除之后, 若检测到此时集合G′不为空,则更新集合G′(即获取所有中心点y坐标小于更新 后R_1中心点y坐标的文本块得到新的集合G′),通过更新集合G′,可进一步减少 选择起始文本块的时间。

[0112] 2)特征生成器Ψ_2

[0113] Ψ_2用于根据当前文本块R_i得出下一个阅读顺序状态的特征预测信息O_{i+1},可 以描述为:

[0114] $$R_i \xrightarrow{\Psi_2} O_{i+1}$$

[0115] 如上所述,各文本块可描述为$R=\{x,y,w,h,s,d\}$,对应的Ψ_2可选用一个 包括6维输入、6维输出和两个分别为12维和20维的隐层的全连神经网络,其 结构如图4所示,其中每个圆圈表示一个神经元。对于每个样本块,若表示为 $R=\{r_i;i\in[0,6)\}$,则第一个隐层的输出K_1为:

[0116] $$K_1=\{k_{1i}=\sum_i a_{1i}r_j+b_{1i};i\in[0,12),j\in[0,6)\}$$

[0117] 第二隐层的输出为:

[0118] $$K_2=\{k_{2i}=\sum_i a_{2i}k_{1j}+b_{2i};i\in[0,20),j\in[0,12)\}$$

[0119] 6维输出层的输出为:

[0120] $O=\{o_i=\mathrm{sigmoid}\sum a_{oi}k_{2j}+b_{oi};i\in[0,6),j\in[0,20)\}$

[0121] 其中a、b均为需要训练的参数。O即为Ψ_2的输出。

[0122] 3)特征合成器Ψ_3

[0123] 通过Ψ_2得到下一阅读顺序状态的特征预测信息之后,按照如下方式更新当前 的阅读顺序状态S,以得到的下一阅读顺序状态:

[0124] I)获取在当前阅读顺序状态S状态中为值0的文本块,构建集合G^*,

[0125] $G^*=\{R_i;S_k(R_i)=0\};i\in[1,n]$;

[0126] II)对于每一个$R_i\in G^*$,计算$v_i=R_i\cdot O$,得到集合V^*,$V^*=\{v_i=R_i\cdot O\}$;

[0127] III)找出V^*中的最大值,并找出该值对应的文本块,记为R^*;

[0128] IV)更新当前阅读顺序状态S,即更新S中的$S(R^*)$的值为$S(R^*)=\max(S)+1$; 由此可得到对应的下一阅读顺序状态,即得到对应的下一文本块。以此类推, 可到全部文本块的排序。

[0129] 结合上述实施例所述，下面以图5所示的文档图片为例，对本发明的检测 文档阅读顺序的方法进行举例说明。包括步骤一～步骤五，各步骤具体说明如下：

[0130] 步骤一，对原始的文档图片进行二值化处理和方向校正处理；再对经过二 值化处理及方向校正处理的文档图片进行版面分析，得到文档中包含的全部文 本块。如图5所示，得到文档中包含的文本块为R_1,R_2,R_3,R_4以及R_5。

[0131] 步骤二，确定起始文本块。

[0132] 由于在R_1,R_2,R_3,R_4以及R_5中，R_3的中心点x坐标位于最左侧，因此初始时将 R_{start}赋值为R_3。

[0133] 获取所有中心点y坐标小于R_3中心点y坐标的文本块，并按照y坐标增序 排列，可得到集合$G'=(R_2,R_1)$。

[0134] 循环更新R_{start}。检测到文本块R_2与R_3在x轴方向的投影没有交集，因此从 集合G′中删除R_2；检测到文本块R_1与R_3在x轴方向的投影有交集，因此将R_{start}更 新为R_1，并从集合G′中删除R_1，由于此时集合G′已经为空，因此无需更新集合G′ (即无需获取所有中心点y坐标小于R_1中心点y坐标的文本块以更新集合G′)， 循环结束。获取当前R_{start}对应的文本块为R_1，由此可确定出图5所示文档的起 始文本块为R_1。

[0135] 步骤三，从起始文本块R_1开始自动寻径。

[0136] 当前文本块为$R_1=\{x_1,y_1,w_1,h_1,s_1,d_1\}$，当前状态为$S_1=(1,0,0,0,0)$；将 $R_1=\{x_1,y_1,w_1,h_1,s_1,d_1\}$输入到训练好的神经网络模型，获取神经网络模型输出的 预测信息为$O=\{o_1,o_2,o_3,o_4,o_5,o_6\}$；

[0137] 基于当前状态为$S_1=(1,0,0,0,0)$，可得到集合$G^*=\{R_2,R_3,R_4,R_5\}$；

[0138] 进一步可得到：

[0139] $V^*=\{R_2\cdot O,R_3\cdot O,R_4\cdot O,R_5\cdot O,\}$；

[0140] $R_i\cdot O=x_i\times o_1+y_i\times o_2+w_i\times o_3+h_i\times o_4+d_i\times o_5$；

[0141] 选取V^*中的最大值所对应的文本块，本实施例中可得出$R_3\cdot O$的值最大，更新 当前阅读顺序状态$S_1=(1,0,0,0,0)$中文本块R_3对应的值为$s_3=1+1=2$，由此可 得出下一状态为$S_2=(1,0,2,0,0)$，确定出下一文本块为R_3。

[0142] 然后将R_3作为当前文本块，按照同样的方式，可得到R_3对应的下一状态为 $S_3=(1,3,2,0,0)$，即R_3对应的下一文本块为R_2；再将R_2作为当前文本块，按照同 样的方式，可得到R_2对应的下一状态为$S_4=(1,3,2,4,0)$，即R_2对应的下一文本块 为R_4；然后将R_4作为当前文本块，由于此时对应的集合G^*中只有一个文本块(即 R_5)，可直接将该文本块作为当前文本块的下一文本块并得到对应的下一状态为 $S_5=(1,3,2,4,5)$；自此自动寻径结束。

[0143] 步骤四，根据自动寻径的结果，可得到文档阅读顺序为 $R_1\rightarrow R_3\rightarrow R_2\rightarrow R_4\rightarrow R_5$。

[0144] 步骤五：按照$R_1\rightarrow R_3\rightarrow R_2\rightarrow R_4\rightarrow R_5$的顺序依次对文本块进行文本识别， 得到文档对应的可阅读文本信息，对可阅读文本信息进行保存以及输出显示。

[0145] 其中，对文本块的文本识别包括行分割和行识别等步骤，依次以行为单位 进行字符识别，由此可得到整个文本块的文本信息。

[0146] 通过上述实施例检测文档阅读顺序的方法，由于神经网络算法拥有大量的 参数，根据训练好的神经网络模型，能够兼容多种场景，对文档图片的尺寸、噪声、样式具有更好的鲁棒性。

[0147] 需要说明的是，对于前述的各方法实施例，为了简便描述，将其都表述为 一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动 作顺序的限制，因为依据本发明，某些步骤可以采用其它顺序或者同时进行。此外，还可对上述实施例进行任意组合，得到其他的实施例。

[0148] 基于与上述实施例中的检测文档阅读顺序的方法相同的思想，本发明还提供 检测文档阅读顺序的装置，该装置可用于执行上述检测文档阅读顺序的方法。为了便于说明，检测文档阅读顺序的装置实施例的结构示意图中，仅仅示出了 与本发明实施例相关的部分，本领域技术人员可以理解，图示结构并不构成对 装置的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不 同的部件布置。

[0149] 图6为本发明一实施例的检测文档阅读顺序的装置的示意性结构图；如图6 所示，本实施例的检测文档阅读顺序的装置包括：块识别模块610、起始块选择 模块620、自动寻径模块630以及顺序确定模块640，各模块详述如下：

[0150] 所述块识别模块610，用于识别文档图片中包含的文本块，构建一个块集合；

[0151] 在一优选实施例中，所述块识别模块610具体可包括：预处理子模块，用 于对所述文档图片进行二值化处理和方向校正处理；以及，版面识别子模块， 用于对经过二值化处理及方向校正处理的文档图片进行版面分析，得到文档中 包含的文本块。其中，版面分析是指在OCR中，将文档图片中的内容按照段落、分页等信息划分为多个不重叠的区域的算法。由此可得出文档中包含的全部文 本块，例如图3所示或者图5所示。

[0152] 所述起始块选择模块620，用于从所述块集合中确定出一起始文本块。

[0153] 通常情况下，人们在阅读文档时会从文档的一角开始进行阅读，基于此， 在一优选实施例中，所述起始块选择模块620可用于从所述块集合中选择出中 心点坐标位于所述文档图片的一个顶点的文本块，并将该文本块确定为所述起 始文本块。例如，所述起始块选择模块620可用于从全部文本块中选择出中心 点坐标位于文档图片的左侧且最上方的一文本块(即左上角的文本块)，将该文 本块确定为起始文本块。如图3中所示的文本块R_1，或者图5中所示的文本块R_1。

[0154] 可以理解的，在其他实施例中，对于不同的文档和实际的阅读习惯(例如 从右到左排版的文档)，所述起始块选择模块620也可将其他文本块确定为起始 文本块。

[0155] 所述自动寻径模块630，用于根据该起始文本块的特征信息对该起始文本块 执行寻径操作，以确定出所述块集合中与该起始文本块对应的第一文本块；文 本块的特征信息包括该文本块在文档图片中的位置信息以及该文本块的版面布 局信息；根据所述第一文本块的特征信息对该第一文本块执行寻径操作，以确 定出所述块集合中与该第一文本块对应的文本块；并依此类推直到所述块集合 中每一个文本块对应的寻径操作的执行顺序能够唯一确定。

[0156] 本实施例中，所述自动寻径模块630用于执行一个自起始文本块起，对文 档包含的文本块进行自动寻径的过程，且每次寻径只需确定当前文本块对应的 下一文本块。例如图3所示的文档图片，当前文本块为R_1，通过本次寻径可确 定文本块R_1的下一文本块为R_2；然后将R_2作为当前文本再次进行寻径，得到R_2的 下一文本块为R_4；以此类推，直到确定出R_6的下一文本块为R_7为止，每一个文 本块对应的寻径操作的执行顺序能够唯一确定。

[0157] 所述顺序确定模块640，用于确定所述块集合中文本块对应的寻径操作的执 行顺

序，根据所述执行顺序得到所述文档图片中文本块的阅读顺序。

[0158] 例如所述顺序确定模块640可得到图3所示的文档图片中文本块的阅读顺 序为R_1→R_2→R_4→R_5→R_3→R_6→R_7→R_8。

[0159] 在一优选实施例中，所述起始块选择模块620具体可用于以文档图片左上 角顶点为原点建立XOY坐标系，并且该XOY坐标系x轴正方向指向文档图片 的宽度方向，y轴正方向指向文档图片的长度方向；从所述块集合中获取中心点 的x坐标最小的一个文本块，作为文本块A；

[0160] 获取中心点的y坐标小于所述文本块A的文本块，构建一个文本块集合G′； 并依次将该集合G′中的每一个文本块B与所述文本块A进行对比；

[0161] 若所述文本块B与该文本块A在x轴方向的投影不存在交集，则将所述文 本块B从集合G′中删除；若所述文本块B与该文本块A在x轴方向的投影存在 交集，则更新所述文本块A为所述文本块B，并将所述文本块B从集合G′中删 除；在每次文本块对比之后检测集合G′是否为空；若是，则将当前的文本块A 确定为起始文本块；若否，则在所述文本块A发生更新时更新集合G′，并将更 新后的集合G′中的每一个文本块与当前的文本块A进行上述对比；依次类推直 到集合G′为空。

[0162] 在一优选实施例中，每次在用新的文本块B更新所述文本块A，将所述文 本块B从G′中删除之后，若检测到此时集合G′不为空，则更新集合G′(即获取 所有中心点y坐标小于更新后的文本块A中心点y坐标的文本块得到新的集合 G′)，通过更新集合G′，可进一步减少选择起始文本块的时间。

[0163] 在一优选实施例中，如图7所示，所述检测文档阅读顺序的装置还包括：训练模块650，用于预先训练机器学习模型，使得训练之后的机器学习模型输出 的特征预测信息与对应的样本信息的欧式距离满足设定条件。

[0164] 在一优选实施例中，所述训练模块650可包括样本库构建子模块和训练子 模块。其中，样本库构建子模块，用于获取训练样本，建立样本库M＝{G,S,T}， 其中G表示样本块的集合，S表示样本块在先后各次训练中的顺序状态的集合， T表示训练过程中需确定的状态变化序列；若G中样本块的总数为n，则有，

[0165] $S=\{s_i;i\in[1,n],s_i\in[0,n]\}$；

[0166] $T=\{\{R_1,S_1,S_2\},\{R_2,S_2,S_3\},...\{R_{n-2},S_{n-2},S_{n-1}\}\}$；

[0167] $s_i=0$表示样本块R_i的阅读顺序未确定(即执行寻径操作的顺序未确定)， 若$s_i>0$表示样本块R_i的阅读顺序已确定(即执行寻径操作的顺序已确定)，且 阅读顺序为s_i的值，表示为$S(R_i)=s_i$；T中的每一个序列中的各项分别表示当 前参与训练的样本块、当前所有样本块的顺序状态的集合和需预测出的所有样 本块的下一顺序状态的集合。

[0168] 其中，训练子模块，用于依次采用T中的各个序列对机器学习模型中的参数 进行训练；当T中的所有序列均参与训练之后，保存所述机器学习模型中的参数。

[0169] 在一优选实施例中，所述训练子模块在根据T中的第k个序列$\{R_k,S_k,S_{k+1}\}$对 机器学习模型中的参数进行训练时，用于实现以下过程：

[0170] 将样本块R_k的特征信息输入机器学习模型，获取机器学习模型输出的R_k的 下一文本块的特征预测信息O_k，$k\in[1,n-2]$；

[0171] 获取S_k中顺序状态为0的样本块R_i，得到集合G*，

[0172] $G^*=\{R_i;S_k(R_i)=0\};i\in[1,n]$；

[0173] 将集合G^*中各项分别与O_k进行点积运算，得到集合$V^*=\{v_i=R_i\cdot O_k\}$；

[0174] 获取集合G^*中各项在S_{k+1}中对应的顺序状态，得到集合$V^{\pi}=\{v_i'=S_{k+1}(R_i)\}$；

[0175] 对集合V^*进行归一化处理得到集合V^{**}，对集合V^{π}进行归一化处理得到集合 $V^{\pi\pi}$；根据集合V^{**}和集合$V^{\pi\pi}$构建样本块R_k参与训练时对应的损失函数，基于该损 失函数通过BP算法更新所述机器学习模型中的参数，其中所述损失函数为：

[0176] $loss=|V^{**}-V^{\pi\pi}|$。

[0177] 在一优选实施例中，所述块识别模块610还用于获取各文本块的特征向量 $R=\{x,y,w,h,s,d\}$；其中x表示文本块的中心点的x坐标，y表示文本块的中心 点的y坐标，w表示文本块的宽度，h表示文本块的高度，s表示文本块中所有 连通区域的尺度均值，d表示文本块的密度信息。

[0178] 对应的，所述机器学习模型为6维输入且6维输出的神经网络模型。例如：所述神经网络模型包括6维输入层、6维输出层、第一隐层以及第二隐层，所述 第一隐层、第二隐层分别为12维和20维的隐层；

[0179] 若每个文本块的特征信息表示为$R=\{r_j;j\in[0,6)\}$，r_j表示样本块的特征信息 j，则所述第一隐层的输出K_1和第二隐层的输出K_2分别为：

[0180] $$K_1=\{k_{1i}=\sum_i a_{1i}r_j+b_{1i};i\in[0,12),j\in[0,6)\}$$；

[0181] $$K_2=\{k_{2m}=\sum_i a_{2m}k_{1i}+b_{2m};m\in[0,20),i\in[0,12)\}$$；

[0182] 所述6维输出层的输出为O：

[0183] $O=\{o_n=\mathrm{sigmoid}\sum a_{on}k_{2m}+b_{on};n\in[0,6),m\in[0,20)\}$；

[0184] 其中a_{1i}、b_{1i}为第一隐层对应的参数，k_{1i}为第一隐层的第i维输出；a_{2m}、b_{2m}为第二隐层对应的参数，k_{2m}为第二隐层的第m维输出；a_{on}、b_{on}为6维输出层 对应的参数，o_n为第n维输出，Sigmoid表示S型的非线性函数。

[0185] 在一优选实施例中，所述的检测文档阅读顺序的装置还包括：文本识别模 块660，用于对各个文本块进行文本识别，并按照确定出的阅读顺序得到所述文 档图片的文本信息。

[0186] 基于上述实施例提供的检测文档阅读顺序的装置，可识别文档图片中包含 的全部文本块，并从全部文本块中确定出一起始文本块；接下来从起始文本块 开始寻径，根据预先训练好的机器学习模型决定下一步应该走到哪个文本块区 域，直到得出全部文本块的阅读顺序。根据文本块在文档图片中的位置信息以 及该文本块的版面布局信息执行寻径能够兼容多种场景，对文档图片的尺寸、噪声、样式具有更好的鲁棒性，能够准确识别各类文档图片对应的文档阅读顺 序。

[0187] 需要说明的是，上述示例的检测文档阅读顺序的装置的实施方式中，各模 块之间的信息交互、执行过程等内容，由于与本发明前述方法实施例基于同一 构思，其带来的技术效果与本发明前述方法实施例相同，具体内容可参见本发 明方法实施例中的叙述，此处不再赘述。

[0188] 此外，上述示例的检测文档阅读顺序的装置的实施方式中，各功能模块的 逻辑划

分仅是举例说明,实际应用中可以根据需要,例如出于相应硬件的配置 要求或者软件的实现的便利考虑,将上述功能分配由不同的功能模块完成,即 将所述检测文档阅读顺序的装置的内部结构划分成不同的功能模块,以完成以 上描述的全部或者部分功能。其中各功能模既可以采用硬件的形式实现,也可 以采用软件功能模块的形式实现。

[0189] 本领域普通技术人员可以理解,实现上述实施例方法中的全部或部分流程, 是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一计算 机可读取存储介质中,作为独立的产品销售或使用。所述程序在执行时,可执 行如上述各方法的实施例的全部或部分步骤。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0190] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详 述的部分,可以参见其它实施例的相关描述。

[0191] 以上所述实施例仅表达了本发明的几种实施方式,不能理解为对本发明专 利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本 发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范 围。因此,本发明专利的保护范围应以所附权利要求为准。

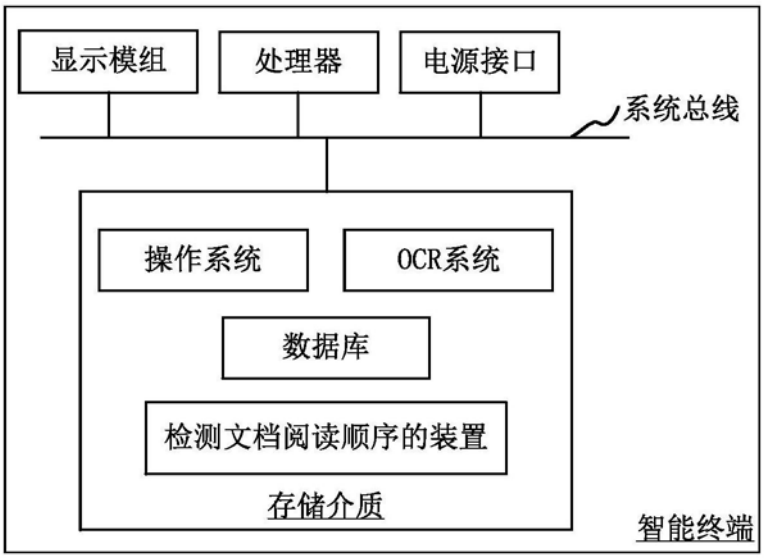


图1

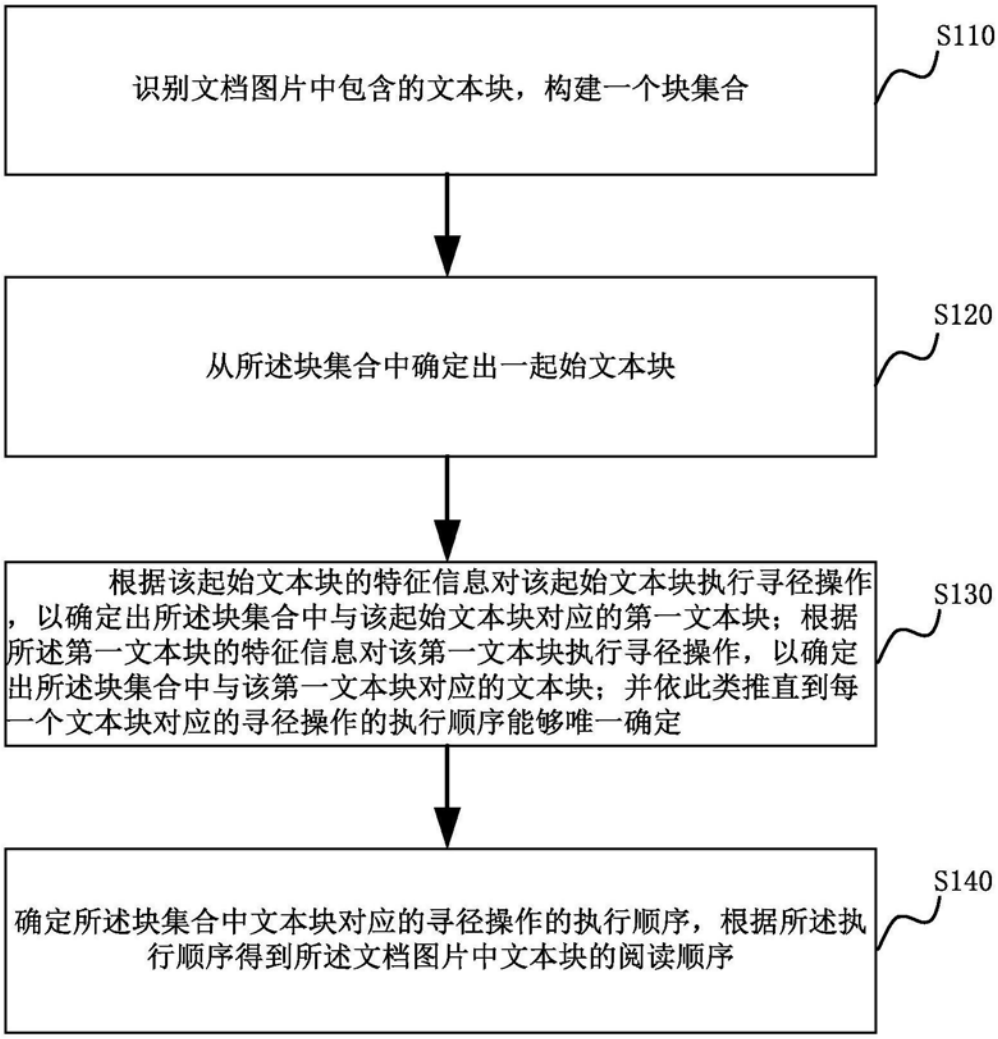

识别文档图片中包含的文本块，构建一个块集合
S110
从所述块集合中确定出一起始文本块
S120
根据该起始文本块的特征信息对该起始文本块执行寻径操作，以确定出所述块集合中与该起始文本块对应的第一文本块；根据所述第一文本块的特征信息对该第一文本块执行寻径操作，以确定出所述块集合中与该第一文本块对应的文本块；并依此类推直到每一个文本块对应的寻径操作的执行顺序能够唯一确定
S130
确定所述块集合中文本块对应的寻径操作的执行顺序，根据所述执行顺序得到所述文档图片中文本块的阅读顺序
S140

图2

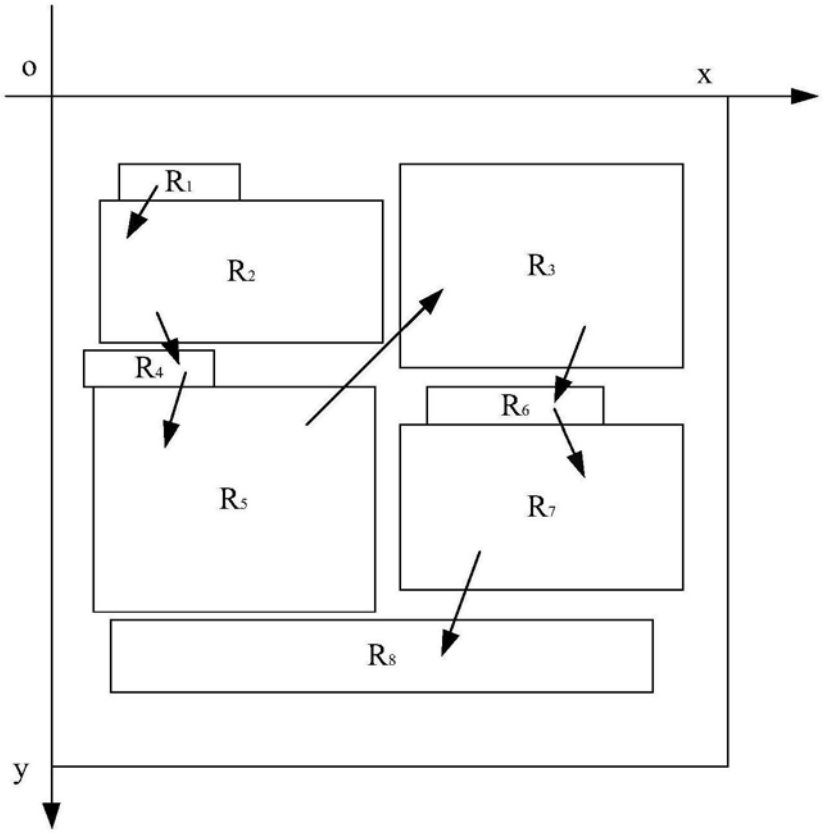


图3

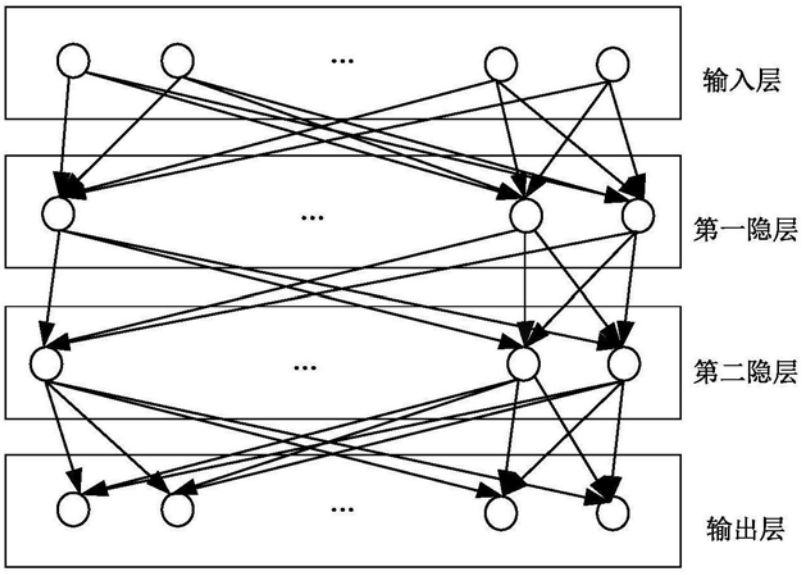


图4

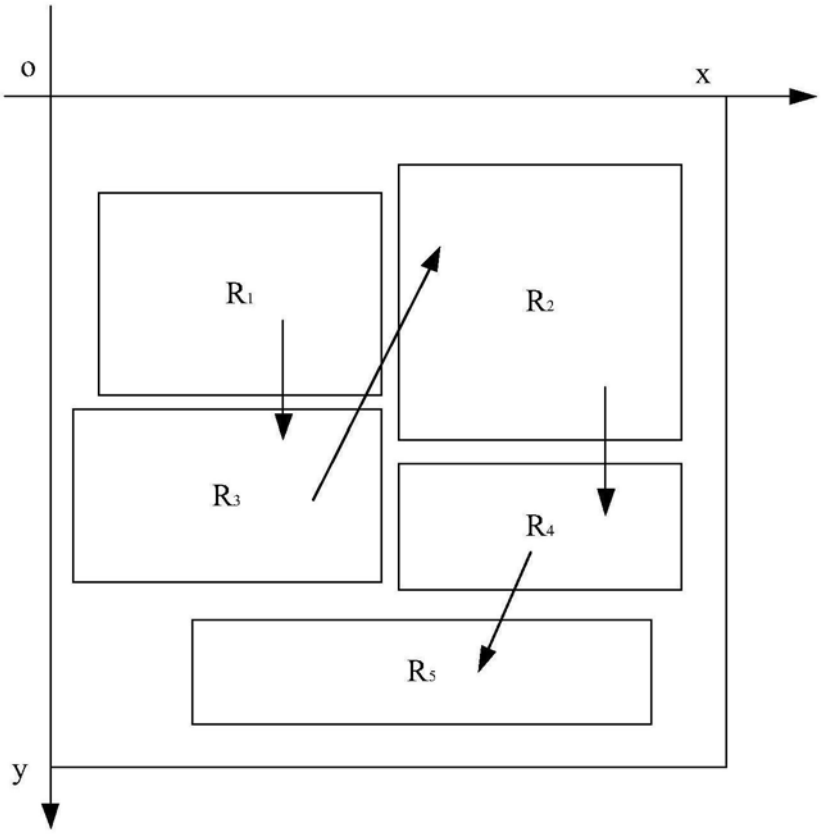

o
x
y
R1
R2
R3
R4
R5

图5

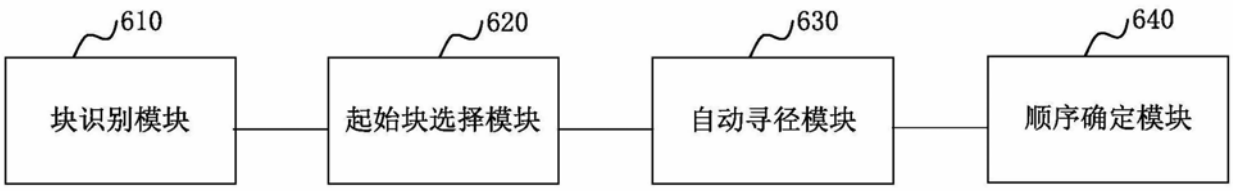

610
块识别模块
620
起始块选择模块
630
自动寻径模块
640
顺序确定模块

图6

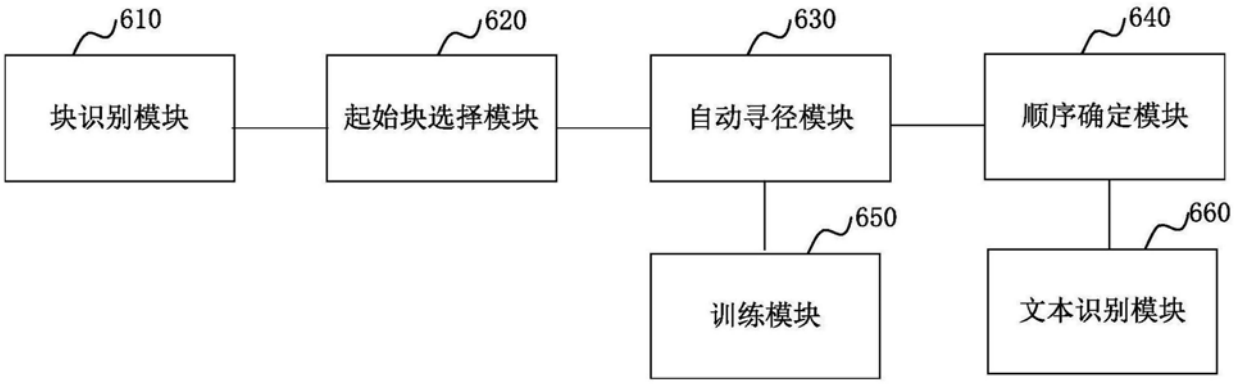

610
块识别模块
620
起始块选择模块
630
自动寻径模块
640
顺序确定模块
650
训练模块
660
文本识别模块

图7