(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111709349 B
(45)授权公告日 2023.12.01

(21)申请号 202010529446.9

(22)申请日 2020.06.11

(65)同一申请的已公布的文献号
申请公布号 CN 111709349 A

(43)申请公布日 2020.09.25

(73)专利权人 杭州尚尚签网络科技有限公司
地址 310012 浙江省杭州市西湖区万塘路317号华星世纪大楼2层202房

(72)发明人 程欢 吴青昀 徐俊杰

(74)专利代理机构 杭州求是专利事务所有限公司 33200
专利代理师 陈升华

(51)Int.Cl.
G06V 30/412(2022.01)
G06V 30/14(2022.01)
G06V 30/19(2022.01)
G06V 10/82(2022.01)
G06N 3/0464(2023.01)
G06N 3/0895(2023.01)

(56)对比文件
CN 110309746 A,2019.10.08
JP H11282957 A,1999.10.15
CN 109933756 A,2019.06.25
CN 110765739 A,2020.02.07
CN 111062187 A,2020.04.24
CN 109993112 A,2019.07.09
CN 111209831 A,2020.05.29
丁明宇 等.基于深度学习的图片中商品参数识别方法.软件学报.2017,第29卷(第4期),第1039-1048页.

审查员 熊洁

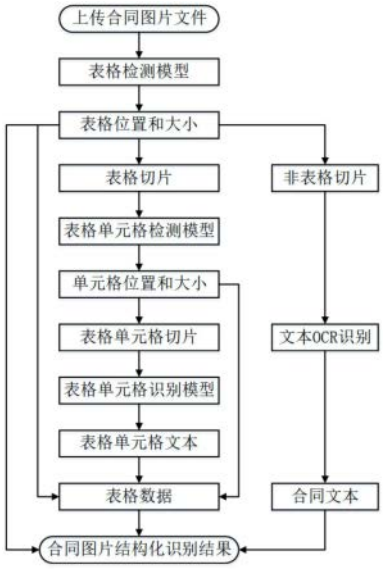
权利要求书2页 说明书7页 附图5页

(54)发明名称

一种针对带表格合同的OCR识别方法

(57)摘要

本发明公开了一种针对带表格合同的OCR识别方法,涉及OCR及深度学习领域。该方法具体包括:对输入的带表格合同图片使用基于YOLOv4的表格检测模型进行检测,利用检测结果对合同图片进行切片处理,得到表格图片;对表格图片使用基于YOLOv4的表格单元格检测模型进行检测,利用检测结果对表格图片进行切片处理,得到表格单元格切片;对表格单元格切片使用基于CRNN+CTC的文字识别模型进行识别,得到单元格内容;结合上述步骤输出信息得到整张合同图片结构化输出。该方法通过将带表格合同识别分成表格检测,单元格检测,单元格识别三个步骤,分别针对性地优化每个步骤模型的性能,提高了表格识别的效果。



CN 111709349 B

1.一种针对带表格合同的OCR识别方法,其特征在于,包括以下步骤:

1)对输入的合同图片进行表格检测,如果合同图片中包含表格,将表格从合同图片中切片出来,得到表格切片,剩余部分作为非表格切片,并记录各切片在原合同图片中的位置信息,得到表格切片的位置信息和非表格切片的位置信息;如果合同图片中不含表格,整张合同图片就作为一个非表格切片;

所述的表格检测采用训练后的表格检测模型,表格检测模型的训练过程包括:

1.1)将模板合同转换成docx格式的word文档,在word文档中插入表格,插入表格后的word文档解压出document.xml文件,然后操作文档里面的document.xml文件,将表格外框修改成特定颜色,得到表格外框修改成特定颜色的word文档;

1.2)将表格外框修改成特定颜色的word文档转换成pdf文档,利用pdf转图片工具转换成合同图片,然后对合同图片进行矩形识别,得到表格在合同图片中的位置和大小,再将未修改的word文档转换成图片,得到标注数据;

1.3)在不同的模板合同中不同位置插入不同类型的表格,重复步骤1.1)和1.2)得到不同的标注数据;

1.4)利用步骤1.2)和1.3)得到的标注数据训练YOLOv4表格检测模型,得到训练后的表格检测模型;

2)对步骤1)得到的表格切片进行单元格检测,根据检测结果对表格切片中的单元格进行切片,得到单元格切片,每个单元格切片仅包含原表格样式中的一个单元格,并且记录单元格切片在表格切片中的位置信息,得到单元格位置信息;

3)对步骤2)产生的单元格切片进行文本识别,得到单元格的文本信息,并对步骤1)产生的非表格切片进行文本识别,得到非表格切片的文本信息;

对步骤2)产生的单元格切片进行文本识别采用训练后的表格单元格识别模型,表格单元格识别模型的训练过程具体包括:

3.1)根据常用于表格中的字符及组合生成文字图片,得到标注数据;

3.2)采用标注数据训练表格单元格识别模型,表格单元格识别模型采用CRNN+CTC模型,得到训练后的表格单元格识别模型;

4)结合步骤3)得到的单元格的文本信息以及步骤2)得到的单元格位置信息,得到结构化的表格信息,再结合步骤3)得到的非表格切片的文本信息和步骤1)得到的表格切片的位置信息和非表格切片的位置信息,得到整张合同图片的识别结果。

2.根据权利要求1所述的针对带表格合同的OCR识别方法,其特征在于,步骤1.2)中,对图片进行矩形识别采用基于opencv库的矩形识别方法。

3.根据权利要求1所述的针对带表格合同的OCR识别方法,其特征在于,步骤2)中,所述的单元格检测采用训练后的表格单元格检测模型,表格单元格检测模型的训练过程包括:

2.1)在空白的docx格式word文档中插入表格,插入表格后的word文档解压出document.xml文件,然后操作文档里面的document.xml文件,将表格线框修改成特定颜色,得到表格线框修改成特定颜色的word文档;

2.2)将表格线框修改成特定颜色的word文档转换成pdf文档,利用pdf转图片工具转换成合同图片,然后对合同图片进行矩形识别,得到表格在合同图片中的位置和大小,再将未修改的word文档转换成图片,得到标注数据;

2.3)利用步骤2.2)的标注数据训练表格单元格检测模型,得到训练后的表格单元格检测模型。

4.根据权利要求3所述的针对带表格合同的OCR识别方法,其特征在于,步骤2.2)中,对图片进行矩形识别采用基于opencv库的矩形识别方法。

5.根据权利要求3所述的针对带表格合同的OCR识别方法,其特征在于,步骤2.3)中,所述的表格单元格检测模型为YOLOv4。

一种针对带表格合同的OCR识别方法

技术领域

[0001] 本发明涉及OCR及深度学习技术领域,具体涉及一种针对带表格合同的OCR识别方法。

背景技术

[0002] 随着越来越多的公司采用电子签约的形式完成合同的签署,线下的纸质合同进行线上化处理的需求就变得很迫切。线下合同通过扫描成图片的形式上传,合同中存在大量的表格,而且一般而言,表格中的信息对于整个合同有着非常重要的意义,所以OCR(Optical Character Recognition,光学字符识别)系统对表格识别的支持显得非常重要。目前基于OCR技术对包含表格文件的识别存在表格定位不准、表格本身的结构干扰识别结果、不同类型表格的差异化支持以及表格中存在的多行文字识别等问题。

[0003] 公开号CN107133621A(申请号为CN201710334784.5)的中国发明专利了一种基于OCR的格式化传真的分类和信息提取方法,包括:对传真的图像进行自适应阈值的二值化;对图像进行校正;找到校正后的图像中表格的最大包围框的轮廓,从图像中表格的最大包围框的上部区域截取图像的表头区域;筛选表头区域中的字体轮廓并对字体轮廓进行融合;检测表头区域合并后的字段的数量,对图像进行分类;提取分类成功的图像,对图像中待识别区域进行定位;根据OCR识别技术对表格中的待识别的区域的字段进行识别;优化已识别的字段。该技术方案基于传统的特征分析的方法处理表格,要求表格具备特定的明显的特征,对无表头,无边框,边框不清晰,表外包含直线等异型表格不能很好地进行识别。

[0004] 因此,特别需要一种基于深度学习的识别方法,能很好地适应特征不明显的各种异型表格。

发明内容

[0005] 针对上述问题,本发明提出了一种针对带表格合同的OCR识别方法,可以支持类型众多的表格识别,同时提高了表格信息识别的准确率。

[0006] 一种针对带表格合同的OCR识别方法,包括以下步骤:

[0007] 1)对输入的合同图片进行表格检测,如果合同图片中包含表格,将表格从合同图片中切片出来,得到表格切片,剩余部分作为非表格切片,并记录各切片在原合同图片中的位置信息,得到表格切片的位置信息和非表格切片的位置信息;如果合同图片中不含表格,整张合同图片就作为一个非表格切片;

[0008] 2)对步骤1)得到的表格切片进行单元格检测,根据检测结果对表格切片中的单元格进行切片,得到单元格切片,每个单元格切片仅包含原表格样式中的一个单元格,并且记录单元格切片在表格切片中的位置信息,得到单元格位置信息;

[0009] 3)对步骤2)产生的单元格切片进行文本识别,得到单元格的文本信息,并对步骤1)产生的非表格切片进行文本识别,得到非表格切片的文本信息;

[0010] 4)结合步骤3)得到的单元格的文本信息以及步骤2)得到的单元格位置信息,得到

结构化的表格信息,再结合步骤3)得到的非表格切片的文本信息和步骤1)得到的表格切片的位置信息和非表格切片的位置信息,得到整张合同图片的识别结果。

[0011] 本发明的方法为一种基于深度学习的识别方法,能很好地适应特征不明显的各种异型表格。

[0012] 以下作为本发明的优选技术方案:

[0013] 步骤1)中,所述的表格检测采用训练后的表格检测模型,表格检测模型的训练过程包括:

[0014] 1.1)将模板合同转换成docx格式的word文档,在word文档中插入表格,插入表格后的word文档解压出document.xml文件,然后操作文档里面的document.xml文件,将表格外框修改成特定颜色,得到表格外框修改成特定颜色的word文档;

[0015] 特定颜色可以自己设定,与文字和表格的颜色能够区分即可,如可选红色。

[0016] 1.2)将表格外框修改成特定颜色的word文档转换成pdf文档,利用pdf转图片工具转换成合同图片,然后对合同图片进行矩形识别,得到表格在合同图片中的位置和大小,再将未修改的word文档转换成图片,得到标注数据;

[0017] 步骤1.2)中,对图片进行矩形识别可采用基于opencv库的矩形识别方法。

[0018] 1.3)在不同的模板合同中不同位置插入不同类型的表格,重复步骤1.1)和1.2)得到不同的标注数据;

[0019] 1.4)利用步骤1.2)和1.3)得到的标注数据训练表格检测模型,得到训练后的表格检测模型。

[0020] 步骤1.4)中,所述的表格检测模型为YOLOv4,属于目标检测模型。表格检测模型是基于YOLOv4目标检测模型使用数据集通过微调训练而得。

[0021] 步骤2)中,所述的单元格检测采用训练后的表格单元格检测模型,表格单元格检测模型的训练过程包括:

[0022] 2.1)在空白的docx格式word文档中插入表格,插入表格后的word文档解压出document.xml文件,然后操作文档里面的document.xml文件,将表格线框修改成特定颜色,得到表格线框修改成特定颜色的word文档;

[0023] 特定颜色可以自己设定,与文字和表格的颜色能够区分即可,如可选红色。

[0024] 2.2)将表格线框修改成特定颜色的word文档转换成pdf文档,利用pdf转图片工具转换成合同图片,然后对合同图片进行矩形识别,得到表格在合同图片中的位置和大小,再将未修改的word文档转换成图片,得到标注数据;

[0025] 步骤2.2)中,对图片进行矩形识别可采用基于opencv库的矩形识别方法。

[0026] 2.3)利用步骤2.2)的标注数据训练表格单元格检测模型,得到训练后的表格单元格检测模型。

[0027] 步骤2.3)中,所述的表格单元格检测模型为YOLOv4,属于目标检测模型。表格单元格检测模型是基于YOLOv4目标检测模型使用数据集通过微调训练而得。

[0028] 步骤3)中,对步骤2)产生的单元格切片进行文本识别采用训练后的表格单元格识别模型,表格单元格识别模型的训练过程具体包括:

[0029] 3.1)根据常用于表格中的字符及组合生成文字图片,得到标注数据;

[0030] 3.2)采用标注数据训练表格单元格识别模型,得到训练后的表格单元格识别模

型。

[0031] 步骤3.2)中,表格单元格识别模型采用CRNN+CTC模型。表格单元格识别模型是基于CRNN+CTC构建的文字识别模型,使用针对表格单元格的数据集训练而得。

[0032] 与现有技术相比,本发明具有如下优点:

[0033] 本发明方法具体包括:对输入的带表格合同图片使用基于YOLOv4的表格检测模型进行检测,利用检测结果对合同图片进行切片处理,得到表格图片;对表格图片使用基于YOLOv4的表格单元格检测模型进行检测,利用检测结果对表格图片进行切片处理,得到表格单元格切片;对表格单元格切片使用基于CRNN+CTC的文字识别模型进行识别,得到单元格内容;结合上述步骤输出信息得到整张合同图片结构化输出。该方法还包括使用弱监督方式为三种模型生成大量高质量标注数据,用来训练模型,提高模型性能。该方法通过将带表格合同识别分成表格检测,单元格检测,单元格识别三个步骤,分别针对性地优化每个步骤模型的性能,提高了表格识别的效果。

[0034] 本发明方法可以支持类型众多的表格识别,同时提高了表格信息识别的准确率。本发明方法不仅可以支持企业的线下合同线上化需求,还可以支持企业年报、员工报销发票中的表格信息提取等。

附图说明

[0035] 图1为本发明的流程图。

[0036] 图2为本发明中使用的YOLOv4模型网络架构图,其中,CSPDarknet53是一种神经网络;SPP是空间金字塔池化;PAN是路径汇聚网络。

[0037] 图3为本发明中使用的CRNN+CTC模型网络架构图。

[0038] 图4为本发明中操作word文档增加边框处理的示意图。

[0039] 图5为未经过边框处理的word文档中表格的样式示意图。

[0040] 图6为经过边框处理后的word文档中表格的样式示意图。

具体实施方式

[0041] 下面结合附图对本发明的技术方案做进一步的讲解:

[0042] 如图1所示,一种针对带表格合同的OCR识别方法是指,利用基于图像的深度学习技术对包含表格的合同文件图片进行处理,处理过程主要分为四个步骤:

[0043] 第一步,对输入的图片进行表格检测,如果图片中包含表格,将表格从原文件中切片出来,剩余部分作为非表格切片,并记录各切片在原图中的位置信息;如果图片中不含表格,整张图片就作为一个非表格切片。

[0044] 第二步,对切片出来的表格图片进行单元格检测,根据检测结果对表格图片进行单元格切片,每个单元格切片仅包含原表格样式中的一个单元格,并且记录单元格切片在原表格图片中的位置信息。

[0045] 第三步,对第二步产生的单元格切片和第一步产生的非表格切片进行文本识别,得到文本信息。

[0046] 第四步,结合单元格的文本信息和位置信息,得到结构化的表格信息,再结合其他非表格切片的文本信息和位置信息,得到整张图片的识别结果。

[0047] 具体来讲：

[0048] 以合同文件图片作为输入；

[0049] 经过表格检测模型，得到表格的大小和位置信息，将表格从原图中切片出来，将剩余部分切成非表格切片。具体即，沿着表格上边缘切出表格上部，沿表格下边缘切出表格下部，剩下的部分(包含表格)再沿表格左边缘切出表格左部，沿表格右边缘切出表格右部，最后剩余的部分即表格图片；

[0050] 对表格图片，利用单元格检测模型进行处理，得到各个单元格的大小和位置信息，将单元格从表格图片中切片出来；

[0051] 对以上产生的表格单元格切片，利用单元格识别模型，识别出文本信息；

[0052] 对以上产生的非表格切片(表格上，下，左，右切片)利用OCR技术识别出其文本信息；

[0053] 利用单元格切片位置，水平位置相近的作为同一行，垂直位置相近的作为同一列，结合切片文本信息，组装成结构化的表格信息；

[0054] 利用非表格切片和表格图片的位置信息，结合非表格切片识别文本和上述结构化表格信息，组装成最终识别结果。

[0055] 基于表格识别的信息提取方法，包含表格检测模型，表格单元格检测模型，单元格识别模型三种神经网络模型，其中，

[0056] 表格检测模型和表格单元格检测模型是基于YOLOv4目标检测模型分别使用不同的数据集通过微调训练而得。YOLOv4模型网络架构图如图2所示，图2中，CSPDarknet53是一种神经网络；SPP是空间金字塔池化；PAN是路径汇聚网络。

[0057] 表格单元格识别模型是基于CRNN+CTC构建的文字识别模型，使用针对表格单元格的数据集训练而得。CRNN+CTC模型网络架构图如图3所示。

[0058] 基于表格识别的信息提取方法，还包括表格检测模型，表格单元格检测模型，单元格识别模型三种神经网络模型的训练方法：

[0059] 针对表格检测模型，需要使用带表格的合同文件图片进行训练。训练数据集采用以下方法生成：

[0060] 处理word格式的合同文件，通过在合同中插入表格生成带表格的合同文件。通过解析和操作word文件，在表格外围加上特定颜色的框，然后将word文件转换成pdf文件，通过pdf转图片工具，生成带表格的合同文件图片，对该图片进行特定颜色框识别，即可以得到表格在合同文件中的位置和大小，生成相应的标注数据。通过在不同类型合同中插入不同类型的表格，按照上述方法可以生成大量高质量的标注数据。

[0061] 利用标注数据基于YOLOv4模型，通过调优神经网络超参数进行训练，得到表格检测模型。

[0062] 针对表格单元格检测模型，需要使用标注好单元格的表格图片进行训练，训练数据集采用下列方法生成：

[0063] 处理word格式的表格，通过解析和操作word文件，在表格外围加上特定颜色的框，并且在单元格上添加不同的颜色，然后将word文件转换成pdf文件，通过pdf转图片工具，生成表格图片，对该图片利用表格外框颜色切分出表格图片，再对表格单元格利用颜色区分进行识别，即可得到表格中单元格的位置和大小，生成相应的标注数据。通过操作和编辑不

同类型的表格,按照上述方法可以生成大量高质量的标注数据。

[0064] 利用标注数据基于YOLOv4模型,通过调优神经网络超参数进行训练,得到表格单元格检测模型。

[0065] 针对表格单元格识别模型,使用常用于表格中的字符组合生成文字图片作为标注数据,构建基于CRNN+CTC的神经网络模型,通过调优神经网络超参数进行训练,得到表格单元格识别模型。

[0066] 表格检测模型和表格单元格检测模型通过计算CIoU(Complete Intersection over Union)来作为模型评估指标。表格单元格识别模型通过比对文本识别准确率作为模型评估指标。

[0067] 本发明中操作word文档增加边框处理的示意图如图4所示,特定颜色可以自己设定,与文字和表格的颜色能够区分即可,如可选红色。未经过边框处理的word文档中表格的样式示意图如图5所示。经过边框处理后的word文档中表格的样式示意图如图6所示,图6中外框加粗实际上为红色。

[0068] 具体实施时,作为输入的合同图片文件,可以是扫描设备输出的合同图片文件,也可以是PDF转图片获取到的合同图片文件,还可以是其他电子文档转图片方式获得的合同图片文件。合同图片文件可以是png,jpeg或者jpg格式。

[0069] 合同图片文件经表格检测模型处理后,输出检测到的表格在原输入图片中的位置及大小。

[0070] 将表格部分从原图片中切片出来,得到表格切片,剩余部分作为非表格切片。具体来讲,非表格切片就是合同中除去表格部分的合同内容。

[0071] 将表格切片输入到表格单元格检测模型进行处理,输出为各单元格在表格图片中的位置及大小。

[0072] 然后对表格切片做进一步的切片处理,得到表格单元格切片。

[0073] 将表格单元格切片输入到表格单元格识别模型,输出识别到的表格单元格信息。

[0074] 根据单元格间的相对位置信息,在水平方向上相近的单元格,作为表格的同一行,在垂直方向上相近的单元格,作为表格的同一列,即得到结构化的表格信息。

[0075] 将非表格切片经OCR处理,得到合同文本。OCR处理具体可以是支持图片的OCR识别软件,也可以是OCR服务商提供的SaaS服务。

[0076] 结合表格切片在原合同图片中的位置信息,即得到包含合同文本和表格信息的识别结果。

[0077] 在标注数据准备阶段,需要为不同模型分别准备大量高质量标注数据,因人工标注成本高昂,而且真实的合同数据一般为客户隐私数据,难以获取,具体实施时,采用程序生成大量标注数据。各模型数据生成方式如下:

[0078] 针对表格检测模型,将模板合同转换成docx格式的word文档,在word文档中插入表格,然后操作文档里面的document.xml文件,将表格外框修改成特定颜色。再将word文档转换成pdf文档,利用pdf转图片工具转换成合同图片,然后基于opencv库对图片进行矩形识别,可得到表格在合同图片中的位置和大小,再将未修改的word文档转换成图片,即得到标注数据。利用不同word文件,在不同位置插入不同类型的表格就可以产生大量标注数据。

[0079] 针对表格单元格检测模型,采用在空白的docx格式word文档中插入表格,然后操

作word文档里面的document.xml文件,将表格线修改成特定颜色,再将word文档转换成pdf文档,利用pdf转图片工具转换成表格图片,然后基于opencv库对表格图片进行矩形识别,得到表格单元格在表格图片中的位置和大小,再将未修改的word文档转换成图片,即得到标注数据。利用在不同形式的表格中填充不同长度不同内容的文字就可以产生大量标注数据。

[0080] 针对表格单元格识别模型,根据常用于表格中的字符及组合生成文字图片,可得到大量标注数据。

[0081] 在模型训练阶段,表格检测模型和表格单元格检测模型,都选用基于Darknet实现的YOLOv4作为基础模型,修改检测类型数为1,再分别利用各自标注数据进行微调训练,训练模型至收敛。通过计算模型检测框和标注框的CIoU值对模型进行评估。

[0082] 其中CIoU采用以下公式计算:

[0083] $$CIoU = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon$$

[0084] $$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon}$$

[0085] $$\upsilon = \frac{4}{\pi^2}(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h})^2$$

[0086] IoU为预测框和实际框的交并比,$\rho^2(b,b^{gt})$为预测框和实际框中心点的欧氏距离,b为预测框中心点坐标,b^{gt}为实际框中心点坐标,c为预测框和实际框外接矩形的对角线距离。αv计算预测框和实际框宽高比例的偏差,其中,π是圆周率,ω^{gt},h^{gt}分别是实际框的宽度和高度,ω,h分别是预测框的宽度和高度。CIoU不仅考虑了检测框和实际框的IoU,还考虑了框中心位置及框的宽高的偏差,能更准确地评估模型效果。

[0087] YOLOv4是YOLO(You Only Look Once)的第四版,属于一种one-stage目标检测模型。合同图片文件输入模型后,首先通过CSPDarknet-53卷积神经网络进行特征提取,然后对网络输出进行上采样,并将上采样结果拼接CSPDarket-53中间层输出,经SPP(Spatial Pyramid Pooling)和PAN(Path Aggregation Network)网络进行特征融合,最后采用原YOLOv3头部网络分别在三个尺寸上对目标进行预测,得到19×19×18,38×38×18,76×76×18三种预测结果。模型训练即是在每种尺寸的每个位置上回归目标边框及类别,使用的损失函数为:

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}[(b_x - \hat{b}_x)^2 + (b_y - \hat{b}_y)^2 + (b_w - \hat{b}_w)^2 + (b_h - \hat{b}_h)^2]$$

[0088] $$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{obj}[-log(p_c) + \sum_{i=1}^{n}BCE(\hat{c}_i, c_i)]$$

$$+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{i,j}^{noobj}[-log(1 - p_c)]$$

[0089] 其中,

[0090] S:网格数,S^2即是19×19,38×38,76×76;

[0091] B:预测目标边框;

[0092] $1_{i,j}^{obj}$:如果预测框中包含目标,则值为1,否则为0;

[0093] BCE(二值交叉熵):$BCE(\hat{c}_i, c_i) = -\hat{c}_i * log(c_i) - (1-\hat{c}_i) * log(1-c_i)$

[0094] $1_{i,j}^{noobj}$:如果预测框中包含目标,则值为0,否则值为1;

[0095] YOLOv4在训练时还会采用Mosaic和SAT(Self-Adversarial Training)做数据增强。

[0096] 表格单元格识别模型,选用基于PyTorch实现的CRNN+CTC模型,在标注数据上进行训练微调至模型收敛。通过计算文本识别的准确率对模型进行评估。

[0097] 本发明将包含表格的合同图片识别过程分成表格检测,单元格检测,单元格识别三个阶段,每个阶段分别使用专门训练的计算机深度学习模型,保证模型的准确性和泛化能力,不仅能处理格式特征明显的各种表格,还能处理无表头,无边框,表格线残缺等格式特征不明显的各种表格,提高表格信息识别的准确率。

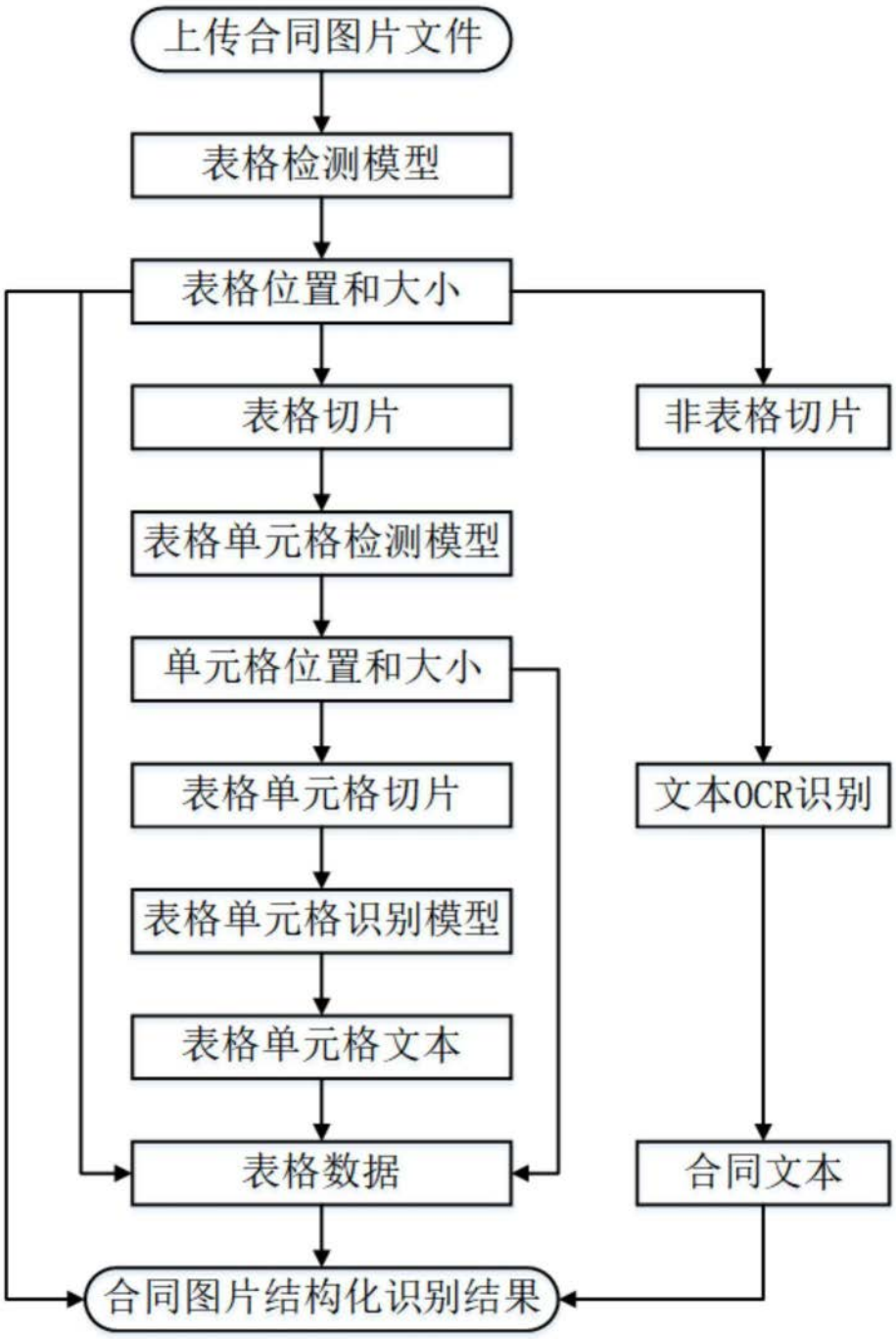

上传合同图片文件
表格检测模型
表格位置和大小
表格切片
非表格切片
表格单元格检测模型
单元格位置和大小
表格单元格切片
文本OCR识别
表格单元格识别模型
表格单元格文本
表格数据
合同文本
合同图片结构化识别结果

图1

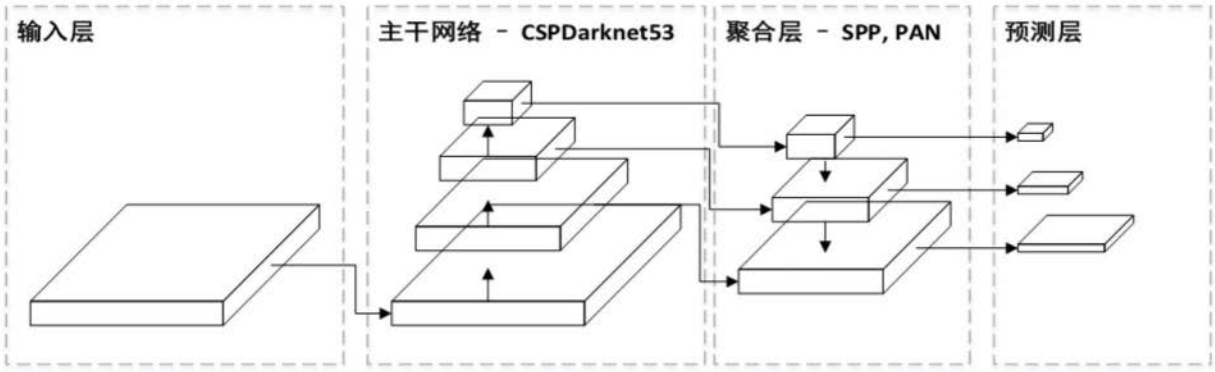


图2

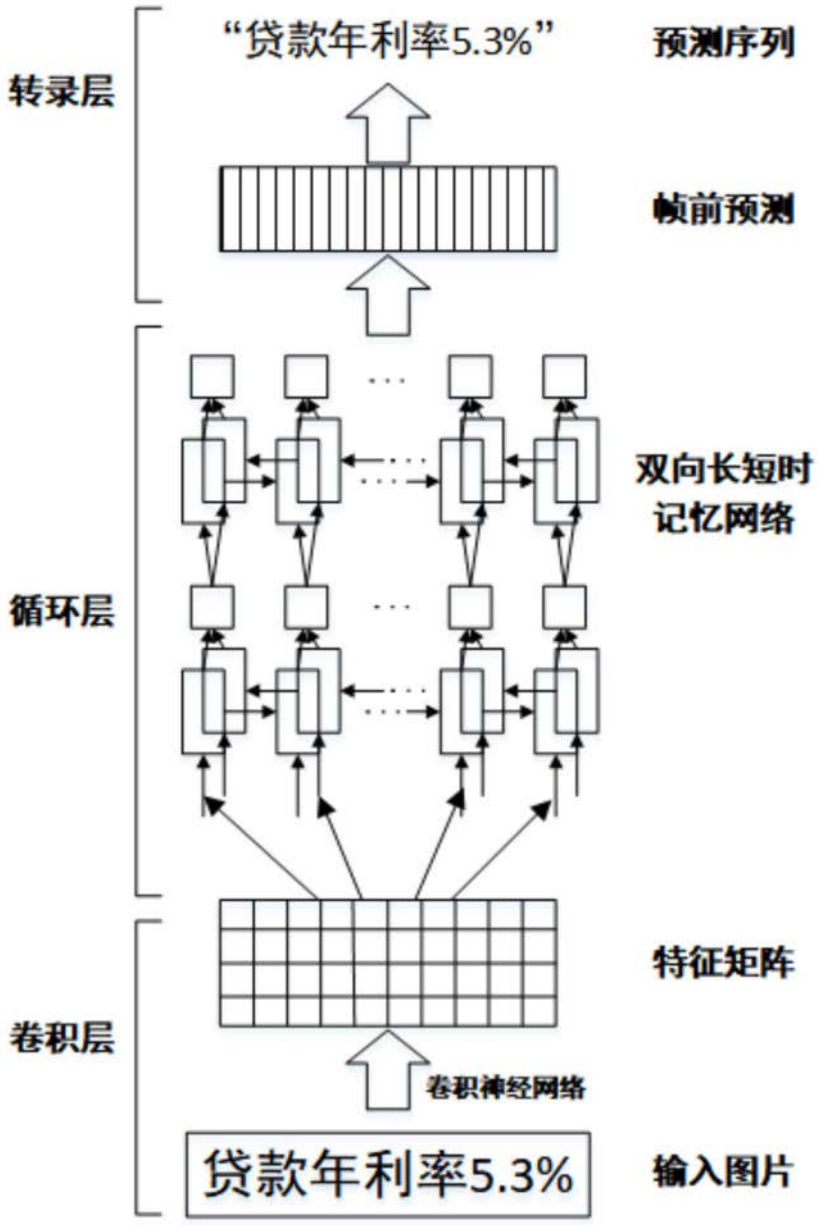

转录层
“贷款年利率5.3%”
预测序列
帧前预测
循环层
双向长短时
记忆网络
卷积层
特征矩阵
卷积神经网络
贷款年利率5.3%
输入图片

图3

图4

第一条 贷款

1.1 贷款相关信息

贷款年利率	6.37%	万	千	百	十	元	角	分
贷款本金	人民币（大写）	玖	捌	肆	陆	零	零	零
贷款用途	购车							

1.2 乙方指定下述账户为贷款接收账户：

账户名称：王某某

账　　号：6227002470130278190

开 户 行：中国建设银行

1.3 甲方在批准乙方贷款请求后的两个工作日内将贷款金额汇入本协议 1.2 款所示的乙方指定贷款接收账户。

图5

第一条 贷款

1.1 贷款相关信息

贷款年利率	6.37%	万	千	百	十	元	角	分
贷款本金	人民币（大写）	玖	捌	肆	陆	零	零	零
贷款用途	购车							

1.2 乙方指定下述账户为贷款接收账户：

账户名称：王某某

账 号： 6227002470130278190

开 户 行：中国建设银行

1.3 甲方在批准乙方贷款请求后的两个工作日内将贷款金额汇入本协议 1.2 款所示的乙方指定贷款接收账户。

图6