(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 112396047 B
(45)授权公告日 2022.03.08

(21)申请号 202011185686.8
(22)申请日 2020.10.30
(65)同一申请的已公布的文献号
申请公布号 CN 112396047 A
(43)申请公布日 2021.02.23
(73)专利权人 中电金信软件有限公司
地址 100192 北京市海淀区西小口路66号东升科技园C区4号楼401室
(72)发明人 周进洋 刘洋 刘渊 张科 梁扩战
(74)专利代理机构 北京华进京联知识产权代理有限公司 11606
代理人 魏朋
(51)Int.Cl.
G06V 10/22(2022.01)
G06V 10/774(2022.01)
G06V 30/10(2022.01)
G06K 9/62(2022.01)
(56)对比文件
CN 107491790 A,2017.12.19
CN 107491790 A,2017.12.19
CN 110728307 A,2020.01.24
CN 108446621 A,2018.08.24
CN 110033445 A,2019.07.19
CN 111582294 A,2020.08.25
CN 104899571 A,2015.09.09
审查员 秦涛

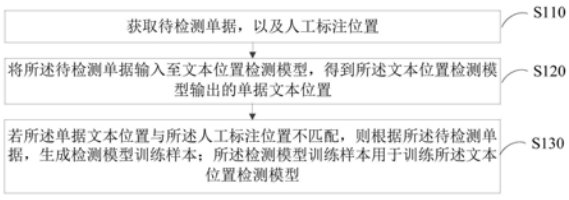
权利要求书2页 说明书11页 附图3页

(54)发明名称

训练样本生成方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种训练样本生成方法、装置、计算机设备和存储介质。所述方法包括：获取待检测单据，以及，获取人工标注位置；将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型。采用本方法能够提高训练样本生成效率。



CN 112396047 B

1.一种训练样本生成方法,其特征在于,所述方法包括:

获取待检测单据,以及人工标注位置;

将所述待检测单据输入至文本位置检测模型,得到所述文本位置检测模型输出的单据文本位置;

若所述单据文本位置与所述人工标注位置不匹配,则对所述待检测单据进行图像增强处理,生成检测模型训练样本;

根据所述检测模型训练样本对所述文本位置检测模型进行训练,得到目标文本位置检测模型;

将待识别单据输入至所述目标文本位置检测模型,得到所述目标文本位置检测模型输出的目标文本位置;根据所述目标文本位置,在所述待识别单据中截取目标文本图片;

将所述目标文本图片输入至文本识别模型,得到所述文本识别模型输出的文本内容,并统计所述文本识别模型的识别情况;

若所述识别情况符合预设的样本增强条件,通过对所述目标文本图片进行图像增强,生成识别模型训练样本;所述识别模型训练样本用于训练所述文本识别模型。

2.根据权利要求1所述的方法,其特征在于,所述对所述待检测单据进行图像增强处理,生成所述待检测单据的多个副本,将所述待检测单据和所述待检测单据的多个副本作为检测模型训练样本,包括:

获取所述文本位置检测模型的第一原始训练样本;

通过对所述待检测单据进行图像增强,得到第一增强训练样本;所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种;

根据所述第一原始训练样本和所述第一增强训练样本,得到所述检测模型训练样本。

3.根据权利要求2所述的方法,其特征在于,所述通过对所述待检测单据进行图像增强,得到第一增强训练样本,包括:

通过对所述待检测单据进行图像增强,得到待检测单据增强图像;

根据所述人工标注位置,得到所述待检测单据增强图像的文本位置;

根据所述待检测单据增强图像和所述文本位置,得到所述第一增强训练样本。

4.根据权利要求1所述的方法,其特征在于,所述方法还包括:

根据所述检测模型训练样本对所述文本位置检测模型进行训练,得到目标文本位置检测模型;

获取待识别单据;

将所述待识别单据输入至所述目标文本位置检测模型,得到所述目标文本位置检测模型输出的目标文本位置;

根据所述目标文本位置,在所述待识别单据中截取目标文本图片;

将所述目标文本图片输入至文本识别模型,得到所述文本识别模型输出的文本内容;

根据所述文本内容,统计所述文本识别模型的识别情况;

根据所述识别情况,生成识别模型训练样本;所述识别模型训练样本用于训练所述文本识别模型。

5.根据权利要求4所述的方法,其特征在于,所述根据所述识别情况,生成识别模型训练样本,包括:

若所述识别情况符合预设的样本增强条件，则获取所述文本识别模型的第二原始训练样本；

通过对所述目标文本图片进行图像增强，得到第二增强训练样本；所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；

根据所述第二原始训练样本和所述第二增强训练样本，得到所述识别模型训练样本。

6.根据权利要求4所述的方法，其特征在于，所述识别情况包括准确率等级；所述根据所述文本内容，统计所述文本识别模型的识别情况，包括：

将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；

统计所述文本内容与所述交叉验证文本内容之间的匹配情况；所述匹配情况包括完全匹配、部分匹配和不匹配中的至少一种；

根据所述匹配情况，得到所述准确率等级。

7.根据权利要求4所述的方法，其特征在于，所述识别情况还包括识别置信度；所述根据所述文本内容，统计所述文本识别模型的识别情况，还包括：

将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；

统计所述文本内容的第一置信度和所述交叉验证文本内容的第二置信度；

根据预设的第一置信度权重和第二置信度权重，对所述第一置信度和所述第二置信度进行加权，得到所述识别置信度。

8.一种训练样本生成装置，其特征在于，所述装置包括：

获取模块，用于获取待检测单据，以及人工标注位置；

文本位置检测模块，用于将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

训练样本生成模块，用于若所述单据文本位置与所述人工标注位置不匹配，则对所述待检测单据进行图像增强处理，生成检测模型训练样本；

所述训练样本生成装置，还用于根据所述检测模型训练样本对所述文本位置检测模型进行训练，得到目标文本位置检测模型；将待识别单据输入至所述目标文本位置检测模型，得到所述目标文本位置检测模型输出的目标文本位置；根据所述目标文本位置，在所述待识别单据中截取目标文本图片；将所述目标文本图片输入至文本识别模型，得到所述文本识别模型输出的文本内容，并统计所述文本识别模型的识别情况；若所述识别情况符合预设的样本增强条件，通过对所述目标文本图片进行图像增强，生成识别模型训练样本；所述识别模型训练样本用于训练所述文本识别模型。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述的方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的方法的步骤。

训练样本生成方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及字符识别技术领域，特别是涉及一种训练样本生成方法、装置、计算机设备和存储介质。

背景技术

[0002] 随着字符识别技术的发展，出现了字符识别模型，在使用字符识别模型进行文字识别时，通常采用逐行扫描的方式，对纸上打印的字符或照片中的字符依次进行识别。

[0003] 在对单据进行识别时，由于字符通常位于单据上的特定位置，采用传统的逐行扫描方式效率较低。为提高识别效率和降低识别难度，可以先通过位置检测模型在单据上检测字符位置，根据字符位置将单据拆分为多个包含字符内容的图片，再使用字符识别模型对图片进行识别。

[0004] 然而，目前在获取位置检测模型的训练样本时，通常是在采集单据后，通过人工方式对单据上的字符位置进行识别和标注，将带有标注的单据作为训练样本，当需要提高位置模型检测准确率时，需要人工对大量的单据进行识别和标注，训练样本的生成效率较低。

[0005] 因此，目前用于字符识别的训练样本生成技术存在效率较低的问题。

发明内容

[0006] 基于此，有必要针对上述技术问题，提供一种能够提高效率的训练样本生成方法、装置、计算机设备和存储介质。

[0007] 本发明实施例提供一种训练样本生成方法，所述方法包括：

[0008] 获取待检测单据，以及人工标注位置；

[0009] 将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0010] 若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型。

[0011] 在其中一个实施例中，所述根据所述待检测单据，生成检测模型训练样本，包括：

[0012] 获取所述文本位置检测模型的第一原始训练样本；

[0013] 通过对所述待检测单据进行图像增强，得到第一增强训练样本；所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；

[0014] 根据所述第一原始训练样本和所述第一增强训练样本，得到所述检测模型训练样本。

[0015] 在其中一个实施例中，所述通过对所述待检测单据进行图像增强，得到第一增强训练样本，包括：

[0016] 通过对所述待检测单据进行图像增强，得到待检测单据增强图像；

[0017] 根据所述人工标注位置，得到所述待检测单据增强图像的文本位置；

[0018] 根据所述待检测单据增强图像和所述文本位置，得到所述第一增强训练样本。

[0019] 在其中一个实施例中，所述方法还包括：

[0020] 根据所述检测模型训练样本对所述文本位置检测模型进行训练，得到目标文本位置检测模型；

[0021] 获取待识别单据；

[0022] 将所述待识别单据输入至所述目标文本位置检测模型，得到所述目标文本位置检测模型输出的目标文本位置；

[0023] 根据所述目标文本位置，在所述待识别单据中截取目标文本图片；

[0024] 将所述目标文本图片输入至文本识别模型，得到所述文本识别模型输出的文本内容；

[0025] 根据所述文本内容，统计所述文本识别模型的识别情况；

[0026] 根据所述识别情况，生成识别模型训练样本；所述识别模型训练样本用于训练所述文本识别模型。

[0027] 在其中一个实施例中，所述根据所述识别情况，生成识别模型训练样本，包括：

[0028] 若所述识别情况符合预设的样本增强条件，则获取所述文本识别模型的第二原始训练样本；

[0029] 通过对所述目标文本图片进行图像增强，得到第二增强训练样本；所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；

[0030] 根据所述第二原始训练样本和所述第二增强训练样本，得到所述识别模型训练样本。

[0031] 在其中一个实施例中，所述识别情况包括准确率等级；所述根据所述文本内容，统计所述文本识别模型的识别情况，包括：

[0032] 将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；

[0033] 统计所述文本内容与所述交叉验证文本内容之间的匹配情况；所述匹配情况包括完全匹配、部分匹配和不匹配中的至少一种；

[0034] 根据所述匹配情况，得到所述准确率等级。

[0035] 在其中一个实施例中，所述识别情况还包括识别置信度；所述根据所述文本内容，统计所述文本识别模型的识别情况，还包括：

[0036] 将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；

[0037] 统计所述文本内容的第一置信度和所述交叉验证文本内容的第二置信度；

[0038] 根据预设的第一置信度权重和第二置信度权重，对所述第一置信度和所述第二置信度进行加权，得到所述识别置信度。

[0039] 本发明实施例提供一种训练样本生成装置，所述装置包括：

[0040] 获取模块，用于获取待检测单据，以及人工标注位置；

[0041] 文本位置检测模块，用于将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0042] 训练样本生成模块，用于若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位

置检测模型。

[0043] 本发明实施例提供一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现以下步骤：

[0044] 获取待检测单据，以及人工标注位置；

[0045] 将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0046] 若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型。

[0047] 本发明实施例提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现以下步骤：

[0048] 获取待检测单据，以及人工标注位置；

[0049] 将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0050] 若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型。

[0051] 本发明实施例中的训练样本生成方法、装置、计算机设备和存储介质，通过获取待检测单据和人工标注位置，将待检测单据输入至文本位置检测模型，得到文本位置检测模型输出的单据文本位置，可以将文本位置检测模型检测到的单据文本位置与人工标注位置相比较，若单据文本位置与人工标注位置不匹配，则可以确定模型对待检测单据检测不准确，根据待检测单据生成文本位置检测模型训练样本，可以增加检测不准确的样本在全体训练样本中的比重，提高模型对检测不准确的样本的检测准确率，由于直接根据检测不准确的待检测单据生成训练样本，可以在提高文本位置检测准确率的同时，提高训练样本的生成效率。

附图说明

[0052] 图1为一个实施例中训练样本生成方法的流程示意图；

[0053] 图2为另一个实施例中训练样本生成方法的流程示意图；

[0054] 图3为一个实施例中训练样本生成装置的结构框图；

[0055] 图4为一个实施例中计算机设备的内部结构图。

具体实施方式

[0056] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0057] 在一个实施例中，如图1所示，提供了一种训练样本生成方法，可以应用于终端或服务器中，其中，终端可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器可以用独立的服务器或者是多个服务器组成的服务器集群来实现。以该方法应用于服务器为例进行说明，包括以下步骤：

[0058] 步骤S110，获取待检测单据，以及人工标注位置。

[0059] 其中,待检测单据可以为需要对文本位置进行检测的单据。

[0060] 其中,人工标注位置可以为通过人工识别和标注,所得到的单据上的文本位置。

[0061] 具体实现中,服务器可以获取到真实单据,将其作为待检测单据。还可以对待检测单据上的文本位置进行人工识别,并将识别到的文本位置作为标签,与待检测单据对应存储在服务器上。

[0062] 例如,银行的服务器可以在日营业结束后,获取到当天交易的所有单据,作为待检测单据,模型训练人员可以识别单据上各个文本框的顶点坐标,将顶点坐标作为标签,输入服务器,服务器将标签与待检测单据对应存储。

[0063] 步骤S120,将待检测单据输入至文本位置检测模型,得到文本位置检测模型输出的单据文本位置。

[0064] 其中,文本位置检测模型可以为识别单据上文本位置的模型。单据文本位置可以为单据上各个文本区域所在的位置。

[0065] 其中,文本位置检测模型可以预先根据人工标注的单据文本位置进行训练得到,可以为DBNet(Differentiable Binarization Network,可微二值化网络)或CRAFT(Character Region Awareness for Text Detection,基于字符区域感知的文本检测)模型。

[0066] 具体实现中,服务器可以将获取到的所有真实单据输入至文本位置检测模型,文本位置检测模型可以输出各个单据上各个文本区域所在的位置。

[0067] 例如,文本位置检测模型可以识别到单据上的一个矩形文本区域,将矩形文本区域的四个顶点作为单据文本位置进行输出,还可以将矩形文本区域的左上角顶点坐标,以及矩形的长和宽作为单据文本位置进行输出。

[0068] 步骤S130,若单据文本位置与人工标注位置不匹配,则根据待检测单据,生成检测模型训练样本;检测模型训练样本用于训练文本位置检测模型。

[0069] 其中,检测模型训练样本可以为文本位置检测模型的训练样本。

[0070] 具体实现中,服务器可以将单据文本位置与人工标注位置相比较,若二者相同,则可以确定单据文本位置与人工标注位置相匹配,此时可以不对待检测单据进行处理,直接使用待检测单据作为文本位置检测模型训练样本集合中的一个元素,否则,若二者不相同,则可以确定单据文本位置与人工标注位置不匹配,此时可以对待检测单据进行图像增强处理,包括图像扭曲、图像拉伸和图像倾斜,生成待检测单据的多个副本,将待检测单据和待检测单据的多个副本作为文本位置检测模型训练样本集合中的元素。

[0071] 需要说明的是,当单据文本位置与人工标注位置不匹配时,由于多个副本是通过对待检测单据进行图像增强得到的,且人工标注位置为对待检测单据中文本位置的准确标注,可以用人工标注位置对待检测单据及其多个副本的文本位置进行标注,得到样本标签。

[0072] 实际应用中,可以通过人工方式对单据文本位置进行纠正,若单据文本位置无人工纠正,则可以直接使用待标注单据作为训练样本,否则,若单据文本位置有人工纠正,则可以对待检测单据进行图像增强,生成待检测单据的多个副本,将待标注单据和待标注单据的多个副本作为训练样本,并可以使用人工纠正的文本位置作为训练样本标签。通过对所有的待检测单据进行上述处理,得到训练样本集合,可以对文本位置检测模型进行训练。

[0073] 例如,对于一张当天交易的原始单据,若通过文本位置检测模型检测到原始单据

上一个文本框的左上角顶点坐标为(100,200)，而人工识别的坐标为(80,190)，二者不匹配，则可以通过对原始单据进行扭曲、拉伸和倾斜等形变处理，制造出多张单据副本，原始单据和多张单据副本可以作为文本位置检测模型的训练样本，其中，可以使用人工识别的坐标(80,190)对原始单据和多张单据副本进行标注。

[0074] 上述训练样本生成方法，通过获取待检测单据和人工标注位置，将待检测单据输入至文本位置检测模型，得到文本位置检测模型输出的单据文本位置，可以将文本位置检测模型检测到的单据文本位置与人工标注位置相比较，若单据文本位置与人工标注位置不匹配，则可以确定模型对待检测单据检测不准确，根据待检测单据生成文本位置检测模型训练样本，可以增加检测不准确的样本在全体训练样本中的比重，提高模型对检测不准确的样本的检测准确率，由于直接根据检测不准确的待检测单据生成训练样本，可以在提高文本位置检测准确率的同时，提高训练样本的生成效率。

[0075] 在一个实施例中，上述步骤S130，可以具体包括：

[0076] 步骤S132，获取文本位置检测模型的第一原始训练样本；

[0077] 步骤S134，通过对待检测单据进行图像增强，得到第一增强训练样本；图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；

[0078] 步骤S136，根据第一原始训练样本和第一增强训练样本，得到检测模型训练样本。

[0079] 其中，第一原始训练样本可以为输入文本位置检测模型的各个待检测单据。

[0080] 其中，第一增强训练样本可以为当文本位置检测模型输出的单据文本位置与人工标注位置不匹配时，对输入文本位置检测模型的待检测单据进行图像增强处理所得到的训练样本。

[0081] 具体实现中，可以将输入文本位置检测模型的所有待检测单据均作为第一原始训练样本，在将待检测单据输入文本位置检测模型后，若检测到的单据文本位置与人工标注位置不相同，则可以对待检测单据进行图像增强处理，包括图像扭曲、图像拉伸和图像倾斜，生成待检测单据的多个图像增强副本，作为第一增强训练样本，检测模型训练样本的集合可以由多个第一原始训练样本和多个第一增强训练样本组成。

[0082] 实际应用中，若无需对文本位置检测模型检测到的单据文本位置进行纠正，则可以将相应的待检测单据作为第一原始训练样本，若需要对文本位置检测模型检测到的单据文本位置进行纠正，则可以将相应的待检测单据作为第一原始训练样本，待检测单据的图像增强副本作为第一增强训练样本，通过对当天交易的所有单据进行上述处理，可以得到包括多个第一原始训练样本和多个第一增强训练样本的检测模型训练样本集合。

[0083] 本实施例中，通过获取文本位置检测模型的第一原始训练样本，可以直接将原始的待检测单据作为训练样本，提高训练样本生成效率，通过对待检测单据进行图像增强，得到第一增强训练样本，可以在文本位置检测模型检测错误时，使用待检测单据的多个副本作为训练样本，根据第一原始训练样本和第一增强训练样本得到检测模型训练样本，可以在提高训练样本生成效率的同时，增加检测错误样本在全体样本中的比重，提高文本位置检测模型的检测准确率。

[0084] 在一个实施例中，上述步骤S134，可以具体包括：通过对待检测单据进行图像增强，得到待检测单据增强图像；根据人工标注位置，得到待检测单据增强图像的文本位置；根据待检测单据增强图像和文本位置，得到第一增强训练样本。

[0085] 其中，待检测单据增强图像可以为对待检测单据进行图像增强后的图像。

[0086] 具体实现中，若检测到的单据文本位置与人工标注位置不相同，则可以对输入文本位置检测模型的待检测单据进行图像增强处理，包括图像扭曲、图像拉伸和图像倾斜，得到待检测单据增强图像，待检测单据增强图像中的文本位置可以为人工标注位置，可以用人工标注位置对待检测单据增强图像进行标注，得到待检测单据增强图像的标签，并根据待检测单据增强图像及其标签，得到第一增强训练样本。

[0087] 本实施例中，通过对待检测单据进行图像增强得到待检测单据增强图像，根据人工标注位置得到待检测单据增强图像的文本位置，可以快速生成待检测单据的增强图像和获取增强图像的文本位置，根据待检测单据增强图像和文本位置得到第一增强训练样本，可以提高训练样本生成效率。

[0088] 在一个实施例中，上述训练样本生成方法，具体还可以包括：

[0089] 步骤S140，根据检测模型训练样本对文本位置检测模型进行训练，得到目标文本位置检测模型；

[0090] 步骤S141，获取待识别单据；

[0091] 步骤S142，将待识别单据输入至目标文本位置检测模型，得到目标文本位置检测模型输出的目标文本位置；

[0092] 步骤S143，根据目标文本位置，在待识别单据中截取目标文本图片；

[0093] 步骤S144，将目标文本图片输入至文本识别模型，得到文本识别模型输出的文本内容；

[0094] 步骤S145，根据文本内容，统计文本识别模型的识别情况；

[0095] 步骤S146，根据识别情况，生成识别模型训练样本；识别模型训练样本用于训练文本识别模型。

[0096] 其中，目标文本位置检测模型可以为使用检测模型训练样本对文本位置检测模型进行训练所得到的文本位置检测模型。待识别单据可以为需要对文本内容进行识别的单据。目标文本位置可以为目标文本位置检测模型检测到的文本位置。目标文本图片可以为在待识别单据上目标文本位置处截取的图片，图片中可以包含一个或多个文本。文本识别模型可以为识别文本图片中文本内容的模型，可以预先根据人工标注的文本图片进行训练得到，可以为EfficientNet模型。文本识别模型的识别情况可以为文本识别模型的识别准确等级和识别置信度。识别模型训练样本可以为文本识别模型的训练样本。

[0097] 具体实现中，在通过步骤S130生成检测模型训练样本后，可以使用检测模型训练样本对文本位置检测模型进行训练，得到目标文本位置检测模型，将待识别单据输入目标文本位置检测模型，检测得到待识别单据中的文本位置，服务器可以根据文本位置在待识别单据中截取一个或多个文本图片，作为目标文本图片，将目标文本图片输入至文本识别模型，通过文本识别模型可以识别出目标文本图片中的文本内容。还可以通过交叉验证模型对目标文本图片进行识别，得到交叉验证模型输出的交叉验证文本内容，或者，采用人工方式对文本图片进行识别，得到人工标注文本内容，通过将文本识别模型识别到的文本内容与交叉验证文本内容或人工标注文本内容相比较，可以得到文本识别模型的识别情况，并根据识别情况生成识别模型训练样本，使用识别模型训练样本对文本识别模型进行训练。

[0098] 本实施例中，通过根据检测模型训练样本对文本位置检测模型进行训练，得到目标文本位置检测模型，可以得到位置检测准确率较高的文本位置检测模型，获取待识别单据，将待识别单据输入至目标文本位置检测模型，得到目标文本位置检测模型输出的目标文本位置，并根据目标文本位置在待识别单据中截取目标文本图片，可以在准确确定文本位置的基础上，准确截取文本图片，进而，将目标文本图片输入至文本识别模型，得到文本识别模型输出的文本内容，可以避免文本位置确定不准确对文本识别带来的干扰，根据文本内容统计文本识别模型的识别情况，根据识别情况生成识别模型训练样本，可以根据识别情况调整训练样本，提高文本识别模型的识别准确率。

[0099] 在一个实施例中，上述步骤S146，可以具体包括：若识别情况符合预设的样本增强条件，则获取文本识别模型的第二原始训练样本；通过对目标文本图片进行图像增强，得到第二增强训练样本；图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；根据第二原始训练样本和第二增强训练样本，得到识别模型训练样本。

[0100] 其中，第二原始训练样本可以为输入文本识别模型的各个目标文本图片。

[0101] 其中，第二增强训练样本可以为当文本识别模型输出的文本内容不准确或准确率较低时，对输入文本识别模型的目标文本图片进行图像增强处理所得到的训练样本。

[0102] 具体实现中，可以将输入文本识别模型的所有目标文本图片均作为第二原始训练样本，在将目标文本图片输入文本识别模型后，可以根据文本识别模型输出的文本内容，统计文本识别模型的识别情况，可以统计识别准确率，若识别准确率低于预设的准确率阈值，则可以对目标文本图片进行图像增强，包括图像扭曲、图像拉伸和图像倾斜，生成目标文本图片的多个图像增强副本，作为第二增强训练样本，识别模型训练样本的集合可以由多个第二原始训练样本和多个第二增强训练样本组成。

[0103] 本实施例中，通过若识别情况符合预设的样本增强条件，则获取文本识别模型的第二原始训练样本，通过对目标文本图片进行图像增强，得到第二增强训练样本，可以在识别情况符合预设条件时，直接将目标文本图片及其图像增强副本作为训练样本，提高训练样本生成效率，根据第二原始训练样本和第二增强训练样本，得到识别模型训练样本，可以在提高训练样本生成效率的同时，提高文本识别模型的识别准确率。

[0104] 在一个实施例中，识别情况包括准确率等级，上述步骤S145，可以具体包括：将目标文本图片输入至交叉验证模型，得到交叉验证模型输出的交叉验证文本内容；统计文本内容与交叉验证文本内容之间的匹配情况；匹配情况包括完全匹配、部分匹配和不匹配中的至少一种；根据匹配情况，得到准确率等级。

[0105] 其中，交叉验证内容可以为通过交叉验证模型识别到的文本内容。

[0106] 具体实现中，可以通过第三方付费接口对文本图片进行识别，作为交叉验证文本内容，还可以通过开源的识别引擎对文本图片进行识别，作为交叉验证文本内容。还可以将文本内容与交叉验证文本内容相比较，得到二者的匹配情况，匹配情况可以为完全匹配、部分匹配或不匹配，可以预先设置匹配情况与准确率等级之间的对应关系，基于对应关系，可以根据匹配情况确定文本内容的准确率等级。若准确率等级低于预设的等级阈值，则需要对目标文本图片进行图像增强。

[0107] 例如，可以以EfficientNet模型为文本识别模型，以第三方付费接口和开源识别引擎为交叉验证模型，对目标文本图片分别进行识别，得到三个识别结果，当三个识别结果

完全一致时，可以用绿色进行标记，当识别结果部分相一致时，可以用黄色进行标记，当识别结果全都不同时，可以用红色进行标记。还可以设置颜色标记与准确率等级之间的对应关系，例如，可以设置绿、黄、红分别对应准确率等级的高、中、低。通过颜色标注，可以使用户对准确率等级较低的识别结果进行重点关注。

[0108] 本实施例中，通过将目标文本图片输入至交叉验证模型得到交叉验证模型输出的交叉验证文本内容，统计文本内容与交叉验证文本内容之间的匹配情况，并根据匹配情况得到准确率等级，可以通过对文本识别模型识别到的文本内容进行交叉验证，准确确定文本识别模型的识别准确率等级，进而根据准确率等级确定是否需要通过图像增强生成文本识别模型的训练样本，提高训练样本生成效率。

[0109] 在一个实施例中，识别情况还包括识别置信度，上述步骤S145，具体还可以包括：将目标文本图片输入至交叉验证模型，得到交叉验证模型输出的交叉验证文本内容；统计文本内容的第一置信度和交叉验证文本内容的第二置信度；根据预设的第一置信度权重和第二置信度权重，对第一置信度和第二置信度进行加权，得到识别置信度。

[0110] 具体实现中，可以通过文本识别模型和交叉验证模型对目标文本图片进行识别，分别输出文本内容和交叉验证文本内容，并统计文本内容的第一置信度，和交叉验证文本内容的第二置信度。还可以预先设置文本内容和交叉验证文本内容的置信度权重，例如，可以将文本内容的权重设置为第一置信度权重，将交叉验证文本内容的权重设置为第二置信度权重，根据第一置信度权重和第二置信度权重对第一置信度和第二置信度进行加权，可以得到识别置信度。若识别置信度低于预设的置信度阈值，则需要对目标文本图片进行图像增强。

[0111] 例如，以EfficientNet模型为文本识别模型，以第三方付费接口和开源识别引擎为交叉验证模型，对目标文本图片进行识别，分别得到识别结果“你好”，“你好”，“伲好”，相应地，可以确定置信度分别为1,1,0。预先设置置信度权重系数分别为0.5,0.3,0.2，通过进行加权，可以得到识别置信度为0.5×1+0.3×1+0.2×0＝0.8，即识别结果的置信度为0.8×100％＝80％。通过统计置信度，可以使用户对置信度较低的识别结果进行重点关注。

[0112] 本实施例中，通过将目标文本图片输入至交叉验证模型，得到交叉验证模型输出的交叉验证文本内容，统计文本内容的第一置信度和交叉验证文本内容的第二置信度，根据预设的第一置信度权重和第二置信度权重，对第一置信度和第二置信度进行加权，得到识别置信度，可以通过对文本识别模型识别到的文本内容进行交叉验证，准确确定文本识别模型的识别置信度，进而根据识别置信度确定是否需要通过图像增强生成文本识别模型的训练样本，提高训练样本生成效率。

[0113] 在一个实施例中，如图2所示，提供了一种训练样本生成方法，以该方法应用于服务器为例进行说明，包括以下步骤：

[0114] 步骤S201，获取待检测单据，以及人工标注位置；

[0115] 步骤S202，将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0116] 步骤S203，若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；

[0117] 步骤S204，根据所述检测模型训练样本对所述文本位置检测模型进行训练，得到

目标文本位置检测模型；

[0118] 步骤S205，获取待识别单据；

[0119] 步骤S206，将所述待识别单据输入至所述目标文本位置检测模型，得到所述目标文本位置检测模型输出的目标文本位置；

[0120] 步骤S207，根据所述目标文本位置，在所述待识别单据中截取目标文本图片；

[0121] 步骤S208，将所述目标文本图片输入至文本识别模型，得到所述文本识别模型输出的文本内容；

[0122] 步骤S209，根据所述文本内容，统计所述文本识别模型的识别情况；

[0123] 步骤S210，根据所述识别情况，生成识别模型训练样本；所述识别模型训练样本用于训练所述文本识别模型。

[0124] 为了便于本领域技术人员深入理解本申请实施例，以下将结合一个具体示例进行说明。

[0125] 在单据识别过程中，需要通过检测模型对单据中的文字位置进行识别和标记，以使文字识别模型在识别单据的文字内容时，可以将单据按照文字位置的标记拆分为多个包含文字内容的小图片，进而降低识别难度和提高识别效率。然而，在训练检测模型时，每一次的训练数据都需要通过人工进行标记，及人工标识出待训练单据图片中的文字位置，效率较低。

[0126] 通过采用晚上批量自动进行文本检测，批量识别，将人工标注过程，转变为人工纠正过程，可以极大地提高标注效率。具体地，可以分别对文本位置和字符识别进行标注：

[0127] (a)文本位置标注：使用晚上批量进行文本位置检测和人工纠正，在人工纠正过程中，对于无人工纠正的标注和有人工纠正的标注，分别进行动态统计和分析。其中，对于文字位置检测错误的数据进行数据增强扩展和训练，即对于文字检测错误率比较高的单据，进行数据增强，具体地，可以通过各种变换，把一张图片制造出多张，用于增加训练数据个数。使得训练过程和标注过程不断结合，极大地减少了训练数据量，同时又保障了训练效果。

[0128] (b)文字识别训练数据标注：针对人工纠正的文字区域，动态统计和累积，错误率高的进行数据增强和强化训练。

[0129] 其中，上述标注均基于真实单据进行标注。

[0130] 上述训练样本生成方法，通过检测模型直接检测并标记未标注的单据图片，在人工对检测结果进行纠错，将纠错后的识别结果再作为训练数据对检测模型进行训练，以将原来对单据图片的人工标注过程转换为人工纠错过程，使得训练过程和标注过程不断结合，极大地减少了训练数据量，同时又保障了训练效果。

[0131] 应该理解的是，虽然图1-2的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图1-2中的至少一部分步骤可以包括多个步骤或者多个阶段，这些步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

[0132] 在一个实施例中，如图3所示，提供了一种训练样本生成装置，包括：获取模块310、

文本位置检测模块320和训练样本生成模块330，其中：

[0133] 获取模块310，用于获取待检测单据，以及人工标注位置；

[0134] 文本位置检测模块320，用于将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置；

[0135] 训练样本生成模块330，用于若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型。

[0136] 在一个实施例中，上述训练样本生成模块330，还用于获取所述文本位置检测模型的第一原始训练样本；通过对所述待检测单据进行图像增强，得到第一增强训练样本；所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；根据所述第一原始训练样本和所述第一增强训练样本，得到所述检测模型训练样本。

[0137] 在一个实施例中，上述训练样本生成模块330，还用于通过对所述待检测单据进行图像增强，得到待检测单据增强图像；根据所述人工标注位置，得到所述待检测单据增强图像的文本位置；根据所述待检测单据增强图像和所述文本位置，得到所述第一增强训练样本。

[0138] 在一个实施例中，上述训练样本生成装置，还包括：

[0139] 文本位置检测模型训练模块，用于根据所述检测模型训练样本对所述文本位置检测模型进行训练，得到目标文本位置检测模型；

[0140] 待识别单据获取模块，用于获取待识别单据；

[0141] 文本位置检测模块，用于将所述待识别单据输入至所述目标文本位置检测模型，得到所述目标文本位置检测模型输出的目标文本位置；

[0142] 截取模块，用于根据所述目标文本位置，在所述待识别单据中截取目标文本图片；

[0143] 文本识别模块，用于将所述目标文本图片输入至文本识别模型，得到所述文本识别模型输出的文本内容；

[0144] 识别情况统计模块，用于根据所述文本内容，统计所述文本识别模型的识别情况；

[0145] 识别模型训练样本生成模块，用于根据所述识别情况，生成识别模型训练样本；所述识别模型训练样本用于训练所述文本识别模型。

[0146] 在一个实施例中，上述识别模型训练样本生成模块，还用于若所述识别情况符合预设的样本增强条件，则获取所述文本识别模型的第二原始训练样本；通过对所述目标文本图片进行图像增强，得到第二增强训练样本；所述图像增强包括图像扭曲、图像拉伸和图像倾斜中的至少一种；根据所述第二原始训练样本和所述第二增强训练样本，得到所述识别模型训练样本。

[0147] 在一个实施例中，上述识别情况统计模块，还用于将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；统计所述文本内容与所述交叉验证文本内容之间的匹配情况；所述匹配情况包括完全匹配、部分匹配和不匹配中的至少一种；根据所述匹配情况，得到所述准确率等级。

[0148] 在一个实施例中，上述识别情况统计模块，还用于将所述目标文本图片输入至交叉验证模型，得到所述交叉验证模型输出的交叉验证文本内容；统计所述文本内容的第一置信度和所述交叉验证文本内容的第二置信度；根据预设的第一置信度权重和第二置信度

权重，对所述第一置信度和所述第二置信度进行加权，得到所述识别置信度。

[0149] 关于训练样本生成装置的具体限定可以参见上文中对于训练样本生成方法的限定，在此不再赘述。上述训练样本生成装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0150] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图4所示。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储训练样本生成数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种训练样本生成方法。

[0151] 本领域技术人员可以理解，图4中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0152] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，存储器存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述一种训练样本生成方法的步骤。此处一种训练样本生成方法的步骤可以是上述各个实施例的一种训练样本生成方法中的步骤。

[0153] 在一个实施例中，提供了一种计算机可读存储介质，存储有计算机程序，计算机程序被处理器执行时，使得处理器执行上述一种训练样本生成方法的步骤。此处一种训练样本生成方法的步骤可以是上述各个实施例的一种训练样本生成方法中的步骤。

[0154] 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，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器(Read-Only Memory，ROM)、磁带、软盘、闪存或光存储器等。易失性存储器可包括随机存取存储器(Random Access Memory，RAM)或外部高速缓冲存储器。作为说明而非局限，RAM可以是多种形式，比如静态随机存取存储器(Static Random Access Memory，SRAM)或动态随机存取存储器(Dynamic Random Access Memory，DRAM)等。

[0155] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0156] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对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，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

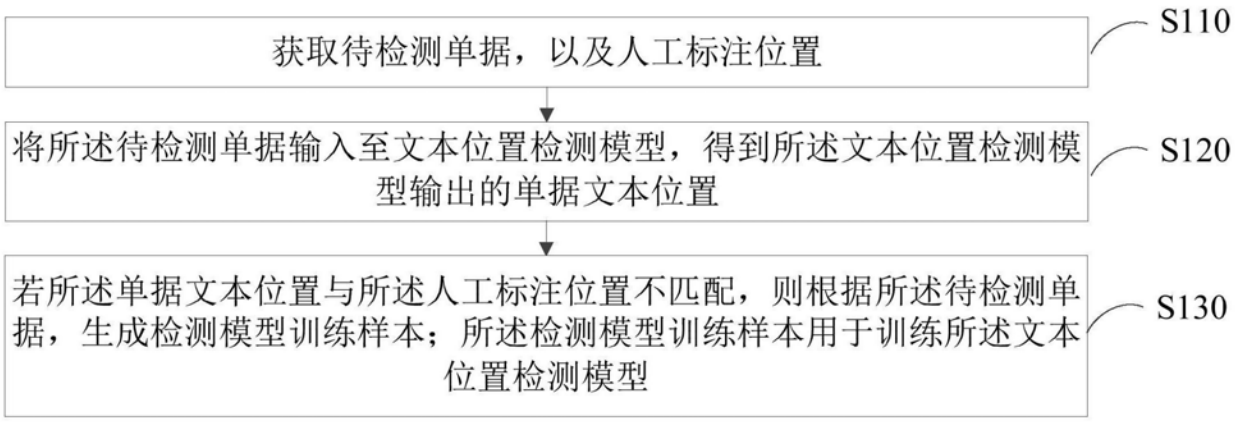

获取待检测单据，以及人工标注位置
S110
将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置
S120
若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；所述检测模型训练样本用于训练所述文本位置检测模型
S130

图1

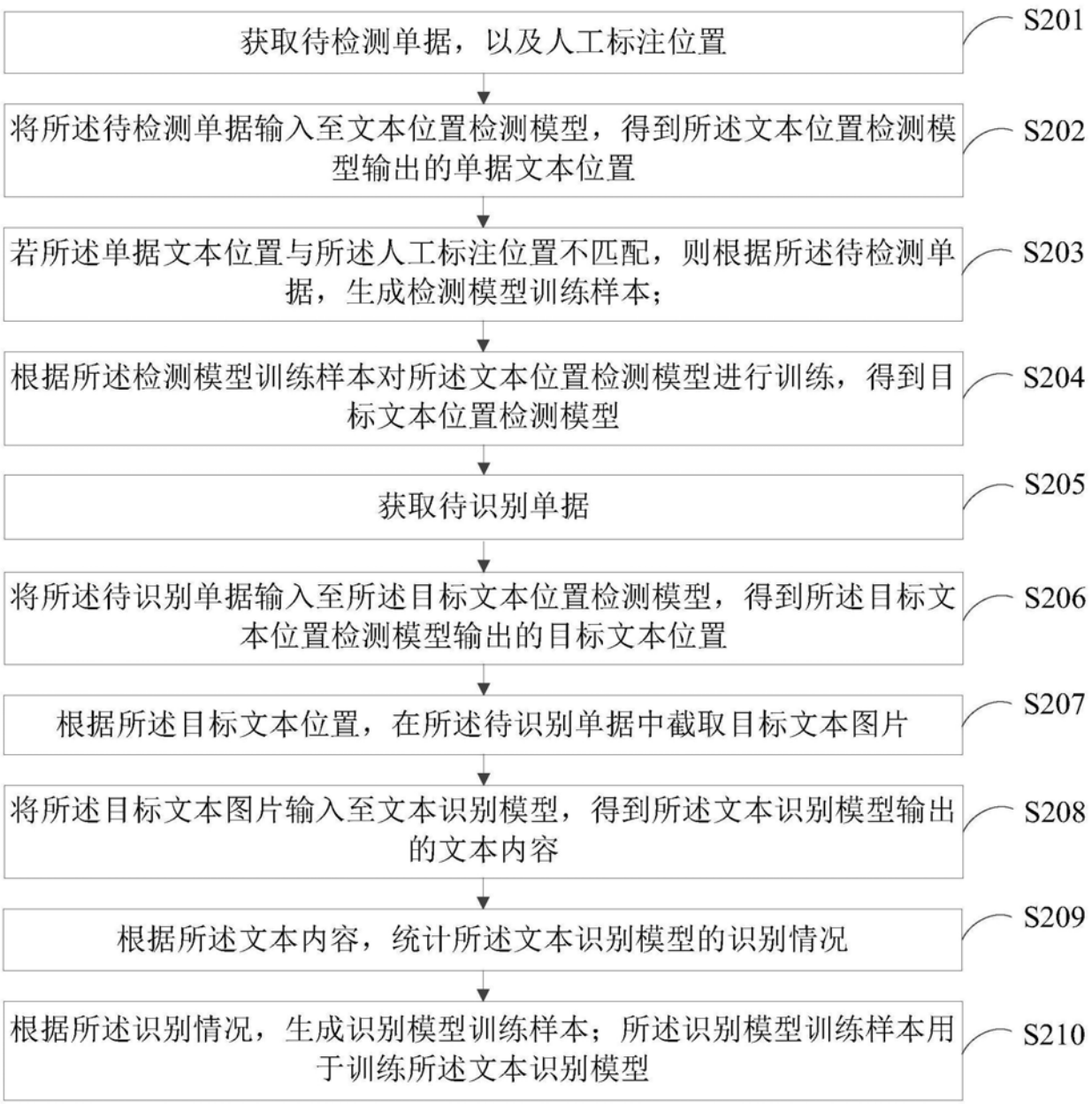

获取待检测单据，以及人工标注位置
S201
将所述待检测单据输入至文本位置检测模型，得到所述文本位置检测模型输出的单据文本位置
S202
若所述单据文本位置与所述人工标注位置不匹配，则根据所述待检测单据，生成检测模型训练样本；
S203
根据所述检测模型训练样本对所述文本位置检测模型进行训练，得到目标文本位置检测模型
S204
获取待识别单据
S205
将所述待识别单据输入至所述目标文本位置检测模型，得到所述目标文本位置检测模型输出的目标文本位置
S206
根据所述目标文本位置，在所述待识别单据中截取目标文本图片
S207
将所述目标文本图片输入至文本识别模型，得到所述文本识别模型输出的文本内容
S208
根据所述文本内容，统计所述文本识别模型的识别情况
S209
根据所述识别情况，生成识别模型训练样本；所述识别模型训练样本用于训练所述文本识别模型
S210

图2

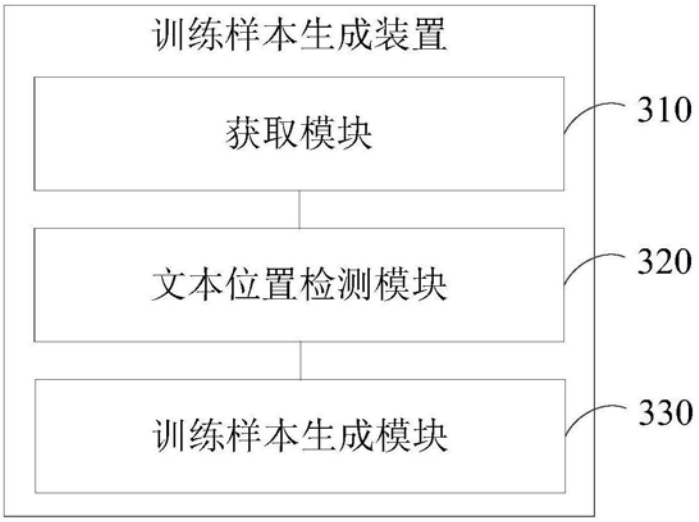

训练样本生成装置
获取模块
310
文本位置检测模块
320
训练样本生成模块
330

图3

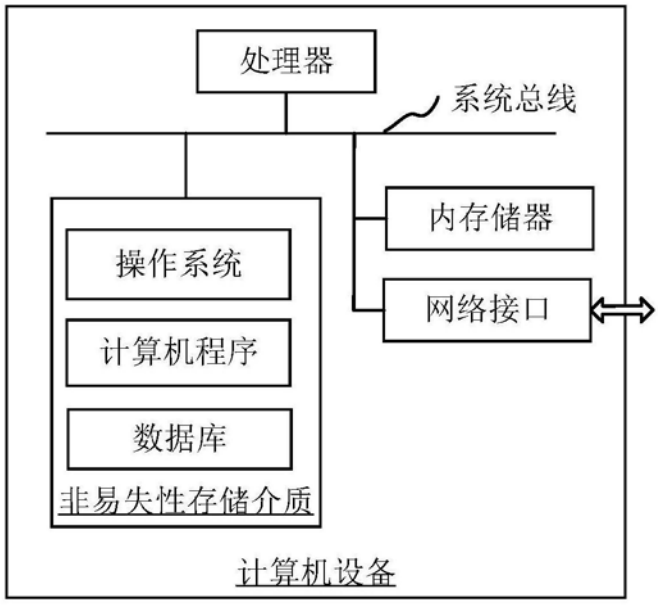

处理器
系统总线
内存储器
网络接口
操作系统
计算机程序
数据库
非易失性存储介质
计算机设备

图4