(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111695518 B
(45)授权公告日 2023.09.29

(21)申请号 202010538181.9

(22)申请日 2020.06.12

(65)同一申请的已公布的文献号
申请公布号 CN 111695518 A

(43)申请公布日 2020.09.22

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 李乔伊 黄相凯 李煜林 黄聚 钦夏孟 秦铎浩 刘明浩 韩钧宇

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 朱颖 刘芳

(51)Int.Cl.
G06V 30/40(2022.01)
G06V 30/19(2022.01)
G06F 40/186(2020.01)
G06F 40/30(2020.01)

(56)对比文件
CN 103186633 A,2013.07.03
CN 108984683 A,2018.12.11
CN 110689447 A,2020.01.14
CN 109086756 A,2018.12.25
US 2011307435 A1,2011.12.15
US 2012254730 A1,2012.10.04
CN 110674627 A,2020.01.10
US 2020065857 A1,2020.02.27
US 2016239473 A1,2016.08.18
US 2019243841 A1,2019.08.08
DE 10162155 A1,2002.07.25
US 2002169803 A1,2002.11.14
Capobianco, Samuele等.DocEmul a Toolkit to Generate Structured Historical Documents.《14th IAPR International Conference on Document Analysis and Recognition (ICDAR)》.2017,1186-1191.
Xavier Holt等.Extracting structured data from invoices.《Proceedings of the Australasian Language Technology Association Workshop 2018》.2018,53-59.
(续)

审查员 母润发

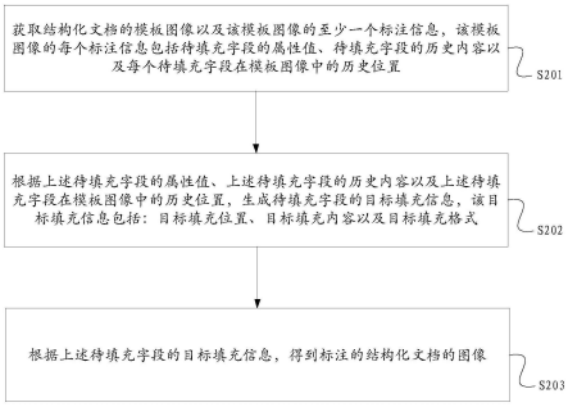
权利要求书3页 说明书12页 附图4页

(54)发明名称

结构化文档信息标注的方法、装置及电子设备

(57)摘要

本申请公开了结构化文档信息标注的方法、装置及电子设备，涉及人工智能领域、深度学习领域以及大数据领域。具体实现方案为：获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息，所述标注信息包括所述待填充字段的属性值、历史内容以及所述待填充字段在所述模板图像中的历史位置。根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置，生成所述待填充字段的目标填充信息。根据所述待填充字段的目标填充信息，得到标注的结构化文档的图像。该方法能够实现结构化文档的快速准确的标注。



CN 111695518 B

[转续页]

[接上页]

(56)对比文件

姜鹏;许峰;戚荣志.一种基于云平台的防汛文档智能生成模型构建.水利信息化.2013,(03),25-32.

高宁;刘洋.数字研发系统中结构化与非结构化数据的融合及实现.计算机应用.2017,(S2),241-243.

1.一种结构化文档信息标注的方法,包括:

获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息,所述标注信息包括所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在所述模板图像中的历史位置;

根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置,生成所述待填充字段的目标填充信息,所述目标填充信息包括:目标填充位置、目标填充内容以及目标填充格式;

根据所述待填充字段的目标填充信息,得到标注的结构化文档的图像;

所述根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置,生成所述待填充字段的目标填充信息,包括:

根据所述待填充字段的属性值和历史内容,生成所述待填充字段的目标填充内容;

根据所述待填充字段在模板图像中的历史位置进行位置调整,得到所述待填充字段的目标填充位置。

2.根据权利要求1所述的方法,其中,所述根据所述待填充字段的属性值和历史内容,生成所述待填充字段的目标填充内容,包括:

若预设语料库中存在所述待填充字段的属性值,则从所述预设语料库中所述待填充字段的属性值对应的内容中随机选择一参考内容,将所述参考内容作为所述待填充字段的所述目标填充内容;

若预设语料库中不存在所述待填充字段的属性值,同时,存在与所述待填充字段的属性值的差异小于预设值的参考属性值,则从所述预设语料库中所述参考属性值对应的内容中随机选择一参考内容,将所述参考内容作为所述待填充字段的所述目标填充内容。

3.根据权利要求1或2所述的方法,其中,所述根据所述待填充字段的属性值和历史内容,生成所述待填充字段的目标填充内容,包括:

若预设语料库中既不存在所述待填充字段的属性值也不存在与所述待填充字段的属性值的差异小于预设值的参考属性值,则确定所述待填充字段的历史内容的词性,并根据所述预设语料库中是否存在所述待填充字段的历史内容的词性对应的内容,生成所述待填充字段的目标填充内容。

4.根据权利要求3所述的方法,其中,所述根据所述预设语料库中是否存在所述待填充字段的历史内容的词性对应的语料,生成所述待填充字段的目标填充内容,包括:

若所述预设语料库中存在所述待填充字段的历史内容的词性对应的内容,则从所述待填充字段的历史内容的词性对应的内容中随机选择一参考内容,将所述参考内容作为所述待填充字段的所述目标填充内容。

5.根据权利要求3所述的方法,其中,所述根据所述预设语料库中是否存在所述待填充字段的历史内容的词性对应的语料,生成所述待填充字段的目标填充内容,包括:

若所述预设语料库中不存在所述待填充字段的历史内容的词性对应的内容,则将所述待填充字段的历史内容作为所述待填充字段的所述目标填充内容。

6.根据权利要求2、4-5任一项所述的方法,其中,所述根据所述待填充字段在模板图像中的历史位置进行位置调整,得到所述待填充字段的目标填充位置,包括:

在所述待填充字段在模板图像中的历史位置所在的区域内,对所述待填充字段的位置

进行调整；

对所述模板图像中至少一个待填充字段按照相同方向和距离进行位置调整；

若所述待填充字段的位置调整后的位置与所述待填充字段之外的字段存在重叠，则对存在重叠的待填充字段的位置进行调整，得到所述待填充字段的目标填充位置。

7.根据权利要求6所述的方法，其中，所述对所述模板图像中至少一个待填充字段按照相同方向和距离进行位置调整，包括：

若所述待填充字段为表格中的字段，则执行如下至少一项：

对所述表格的所有行按照相同方向和距离进行位置调整、对所述表格的每一列分别按照相同方向和距离进行位置调整。

8.根据权利要求6所述的方法，其中，所述对所述模板图像中至少一个待填充字段按照相同方向和距离进行位置调整，包括：

对所述模板图像中的所有待填充字段按照相同方向和距离进行位置调整。

9.根据权利要求1-2、4-5、7-8任一项所述的方法，其中，所述根据所述待填充字段的目标填充信息，得到标注的结构化文档的图像，包括：

在所述模板图像的所述目标填充位置上，按照所述目标填充格式填充所述目标填充内容，得到标注的结构化文档的图像。

10.根据权利要求9所述的方法，其中，所述标注信息还包括：待填充字段的颜色；

所述在所述模板图像的所述目标填充位置上，按照所述目标填充格式填充所述目标填充内容，得到标注的结构化文档的图像，包括：

在所述模板图像的所述目标填充位置上，按照所述目标填充格式以及所述待填充字段的颜色，填充所述目标填充内容，得到标注的结构化文档的图像。

11.根据权利要求1-2、4-5、7-8、10任一项所述的方法，所述获取结构化文档的模板图像以及所述模板图像的至少一个标注信息之前，还包括：

根据用户指示的待填充字段的属性值、待填充字段的历史内容以及每个待填充字段在原始图像中的历史位置，对所述原始图像进行涂抹处理，得到所述模板图像以及所述模板图像的至少一个标注信息。

12.一种结构化文档信息标注的装置，包括：

获取模块，用于获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息，所述标注信息包括所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在所述模板图像中的历史位置；

处理模块，用于根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置，生成所述待填充字段的目标填充信息，所述目标填充信息包括：目标填充位置、目标填充内容以及目标填充格式；以及，

根据所述待填充字段的目标填充信息，得到标注的结构化文档的图像；

所述处理模块，具体用于根据所述待填充字段的属性值和历史内容，生成所述待填充字段的目标填充内容；

根据所述待填充字段在模板图像中的历史位置进行位置调整，得到所述待填充字段的目标填充位置。

13.一种电子设备，包括：

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-11中任一项所述的方法。

14.一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行权利要求1-11中任一项所述的方法。

结构化文档信息标注的方法、装置及电子设备

技术领域

[0001] 本申请实施例涉及计算机技术领域中的人工智能技术、深度学习技术以及大数据技术，尤其涉及一种结构化文档信息标注的方法、装置及电子设备。

背景技术

[0002] 发票、收据、明细单、卡证等是日常工作生活中常见的文档形式，其特点为包括大量键值(key-value)对应关系的文字结构以及表格形式的文字结构，即包含结构化文字。通常，用户只能得到这些文档的纸质版或者这些文档的照片。而在一些场景下，需要这些文档的纸质文档或文档照片中提取出能够结构化存储的关键信息，从而将这些文档电子化。其中，提取能够结构化存储的关键信息时涉及到图像文本识别的相关技术，例如文字检测、结构化解析、端到端文字检测识别、表格提取等等。而图像文本识别的相关技术的实现通常需要使用大量的标注数据进行算法训练。而由于发票、收据、明细单、卡证等文档形式版式繁杂且同版式间变化大，因此，如果使用人工标注，则需要投入较大的人工成本以及时间进行标注。为解决该问题，可以使用自动生成标注数据的自动标注方法。

[0003] 现有技术中，提出了一种自动生成标注数据的方法，该方法中首先选择一帧背景图像，再通过在背景图像中的任意位置随机写入文字，经过多次随机写入，可以得到多个标注数据。

[0004] 但是，现有技术的方法无法适用于发票、收据、明细单、卡证等包含结构化数据的文档自动标注。

发明内容

[0005] 本申请提供了一种一种结构化文档信息标注的方法、装置及电子设备。

[0006] 根据本申请的一方面，提供了一种结构化文档信息标注的方法，包括：

[0007] 获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息，所述标注信息包括所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在所述模板图像中的历史位置。

[0008] 根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置，生成所述待填充字段的目标填充信息，所述目标填充信息包括：目标填充位置、目标填充内容以及目标填充格式。

[0009] 根据所述待填充字段的目标填充信息，得到标注的结构化文档的图像。

[0010] 根据本申请的另一方面，提供了一种结构化文档信息标注的装置，包括：

[0011] 获取模块，用于获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息，所述标注信息包括所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在所述模板图像中的历史位置。

[0012] 处理模块，用于根据所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在模板图像中的历史位置，生成所述待填充字段的目标填充信息，所述目

标填充信息包括:目标填充位置、目标填充内容以及目标填充格式;以及,根据所述待填充字段的目标填充信息,得到标注的结构化文档的图像。

[0013] 根据本申请的又一方面,提供了一种电子设备,包括:

[0014] 至少一个处理器;以及,与所述至少一个处理器通信连接的存储器;其中,所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行上述第一方面所述的方法。

[0015] 根据本申请的又一方面,提供了一种存储有计算机指令的非瞬时计算机可读存储介质,所述计算机指令用于使所述计算机执行上述第一方面所述的方法。

[0016] 根据本申请的又一方面,提供了一种计算机程序产品,所述程序产品包括:计算机程序,所述计算机程序存储在可读存储介质中,电子设备的至少一个处理器可以从所述可读存储介质读取所述计算机程序,所述至少一个处理器执行所述计算机程序使得电子设备执行第一方面所述的方法。该电子设备例如可以是服务器。

[0017] 根据本申请的技术,基于预先生成的模板图像以及待填充字段的属性值、历史内容以及历史位置,可以生成与历史内容和历史位置不同同时又语义一致或相近的目标填充信息,基于该目标填充信息,可以得到标注的结构化文档的图像,从而实现结构化文档的快速准确的标注。当本实施例执行多次后,可以实现仅提供一个原始文档的图像即可生成大量语义一致或相近的虚拟结构化文档,以用于后续的算法训练等过程中。

[0018] 应当理解,本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征,也不用于限制本申请的范围。本申请的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0019] 附图用于更好地理解本方案,不构成对本申请的限定。其中:

[0020] 图1为本申请实施例提供的结构化文档信息标注的方法的示例性场景示意图;

[0021] 图2为本申请实施例提供的机构化文档信息标注的方法的流程示意图;

[0022] 图3为本申请实施例提供的结构化文档信息标注的方法的流程示意图;

[0023] 图4为基于预设语料库生成待填充字段的目标填充内容的流程示意图;

[0024] 图5为本申请实施例提供的结构化文档信息标注的方法的流程示意图;

[0025] 图6为本申请实施例提供的一种结构化文档信息标注的装置的模块结构图;

[0026] 图7是根据本申请实施例的结构化文档信息标注的方法的电子设备的框图。

具体实施方式

[0027] 以下结合附图对本申请的示范性实施例做出说明,其中包括本申请实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本申请的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0028] 现有的自动生成标注数据的方法,首先选择一帧背景图像,再通过在背景图像中的任意位置随机写入文字,经过多次随机写入,可以得到多个标注数据。

[0029] 而对于发票、收据、明细单、卡证等文档来说,其中均包含结构化数据,这类结构化数据在文档中具有特定的位置以及特定的格式。例如,因此,如果使用现有技术的方法,无

法保证结构化数据被在正确的位置填入属性值对应的正确的内容，因此，现有技术的方法无法适用于发票、收据、明细单、卡证等包含结构化数据的文档自动标注。

[0030] 考虑到现有技术中在整帧图像中随机写入文字进行标注的方式无法适用于包含结构化数据的文档的问题，本申请实施例基于结构化数据的文档的实际结构，基于字段在文档的图像中的实际位置以及实际内容，对图像进行标注，从而实现结构化文档的快速准确的标注。

[0031] 为便于描述，本申请实施例以下将包含结构化数据的文档统一简称为结构化文档。

[0032] 图1为本申请实施例提供的结构化文档信息标注的方法的示例性场景示意图，如图1所示，该方法可以应用在提取结构化文档的信息的场景中。在该场景中，可以首先使用本申请的结构化文档信息标注方法得到某一种结构化文档的大量标注文档，这些文档可以均为图像的形式。进而，使用本申请所得到的大量标注文档对用于提取结构化文档的图像文本识别算法进行训练。进而，基于训练好的图像文本识别算法对实际的结构化文档的图像进行信息提取。

[0033] 该场景中所涉及的对结构化文档信息标注、训练图像文本识别算法以及基于训练好的图像文本识别算法对实际的结构化文档的图像进行信息提取，可以在同一电子设备中完成，例如均在同一服务器中完成，或者，也可以在不同的电子设备中完成，本申请实施例对此不作具体限定。

[0034] 图2为本申请实施例提供的机构化文档信息标注的方法的流程示意图，该方法的执行主体可以为前述的电子设备，例如前述的服务器。如图2所示，该方法包括：

[0035] S201、获取结构化文档的模板图像以及该模板图像的至少一个标注信息，该模板图像的每个标注信息包括待填充字段的属性值、待填充字段的历史内容以及每个待填充字段在模板图像中的历史位置。

[0036] 如前文所述，结构化文档是指包含结构化数据的文档，结构化数据可以是指具有特定位置以及特定的格式的数据。示例性，一张发票作为一个结构化文档，其中每项数据均有固定的位置以及固定的格式。

[0037] 上述结构化文档的模板图像以及该模板图像中的标注信息可以预先生成，具体过程将在下述实施例中详细说明。

[0038] 针对每一个模板图像，均可以具有至少一个标注信息。可选的，一个待填充字段可以对应一个标注信息。在生成上述模板图像时，会将待填充字段涂抹掉，在下述处理过程中，再向待填充字段的位置上写入新的内容，从而完成数据的自动标注。

[0039] 其中，一个标注信息包括待填充字段的属性值、历史内容以及待填充字段在模板图像中的历史位置。

[0040] 待填充字段的属性值还可以称为待填充字段的类别。待填充字段的内容可以指待填充字段的实际的值。示例性的，某个发票文档中的某个待填充字段为“张三”，即人的姓名，则该待填充字段的属性值为“姓名”，内容为“张三”。

[0041] 待填充字段的历史内容可以在生成模板图像时直接从原始图像中读取到，待填充字段的属性值可以在生成模板图像时根据字段间的关系得到。

[0042] 在结构化文档中，可能涉及键(key)、值(value)、表头(table_key)、表头内容

(table_value)这几种属性。其中,key和value可以成对出现,table_key和table_value可以成对出现。示例性的,如果结构化文档中的某个待填充字段的属性为value,则该字段的属性值可以是指与该字段关联的属性为key的字段的内容。例如,待填充字段为“张三”,该字段与“姓名”字段关联,则待填充字段的属性值为“姓名”。

[0043] 待填充字段的历史位置可以指示待填充字段在模板图像中所占区域的位置,例如,该历史位置可以包括待填充字段在模板图像中的所占区域的左上角像素位置、右下角像素位置或者中心位置。

[0044] S202、根据上述待填充字段的属性值、上述待填充字段的历史内容以及上述待填充字段在模板图像中的历史位置,生成待填充字段的目标填充信息,该目标填充信息包括:目标填充位置、目标填充内容以及目标填充格式。

[0045] 可选的,目标填充位置可以指待填充字段在需要生成的标注的结构化文档中所占的区域的位置,例如,区域的左上角像素位置和右下角像素位置。目标填充内容是指待填充字段在需要生成的标注的结构化文档中需要填充文字。目标填充格式可以包括:待填充字段所填充文字的字体、字号等。

[0046] 可选的,基于待填充字段的属性值、历史内容以及历史位置,即待填充字段的实际内容和实际位置,可以对历史内容和历史位置进行相应的调整,从而得到与原始文档中字段信息不同,同时又语义一致或相近的目标填充信息。

[0047] S203、根据上述待填充字段的目标填充信息,得到标注的结构化文档的图像。

[0048] 可选的,基于上述目标填充信息,可以在模板图像中写入与原有的待填充字段不同,同时又语义一致或相近的字段信息,从而得到一个标注的结构化文档的图像。

[0049] 上述过程执行多次之后,即可得到大量的互相不同,同时又与原始机构化文档又语义一致或相近的标注文档。

[0050] 本实施例中,基于预先生成的模板图像以及待填充字段的属性值、历史内容以及历史位置,可以生成与历史内容和历史位置不同同时又语义一致或相近的目标填充信息,基于该目标填充信息,可以得到标注的结构化文档的图像,从而实现结构化文档的快速准确的标注。当本实施例执行多次后,可以实现仅提供一个原始文档的图像即可生成大量语义一致或相近的虚拟结构化文档,以用于后续的算法训练等过程中。

[0051] 以下说明上述步骤S202中基于待填充字段的属性值、历史内容以及历史位置,生成目标填充信息的过程。

[0052] 图3为本申请实施例提供的结构化文档信息标注的方法的流程示意图,如图3所示,上述步骤S202的一种可选方式包括:

[0053] S301、根据上述待填充字段的属性值和历史内容,生成上述待填充字段的目标填充内容。

[0054] 待填充字段的属性值可以代表字段的类别,例如年龄、姓名等。待填充字段的历史内容可以代表待填充字段在原始文档中的实际内容,例如“张三”等。基于这两个信息,生成待填充字段的目标填充内容,可以使得目标填充内容与待填充字段的属性值一致同时在内容上可以存在区别。

[0055] S302、根据上述待填充字段在模板图像中的历史位置进行位置调整,得到上述待填充字段的目标填充位置。

[0056] 值的说明的是，上述步骤S301和S302的执行顺序可以不分先后。

[0057] 示例性的，可以基于待填充字段在模板图像中的历史位置，将该历史位置向上或向下调整一定的值，从而使得所得到的目标填充位置以历史位置为基准同时又存在一定差异。

[0058] 本实施例中，基于历史内容和属性值生成目标填充内容，基于历史位置生成目标填充位置，从而可以使得基于该目标填充内容和目标填充位置所生成的标注的结构化文档与原始文档内容语义一致或相近同时又存在一定差异，进而使得后续所得到的多个标注的结构化文档的多样性。

[0059] 以下对上述步骤S301中生成待填充字段的目标填充内容的过程进行说明。

[0060] 作为一种可选的实施方式，可以首先判断预设语料库中是否存在上述待填充字段的属性值或与上述待填充字段的属性值的差异小于预设值的参考属性值，如果存在，则可以从预设语料库中上述待填充字段的属性值对应的内容中或上述参考属性值对应的内容中随机选择一参考内容，将该参考内容作为上述待填充字段的目标填充内容。

[0061] 可选的，上述预设语料库可以是预先采集大量的实际语料所生成的，该预设语料库中可以包括属于各种属性值的内容。示例性的，针对“姓名”这一属性值，预设语料库中可以记录大量的实际姓名，例如“张三”、“李四”、“王五”等。这些实际姓名均与“姓名”这一属性值关联，标识这些实际姓名属于“姓名”。

[0062] 在获取到待填充字段的属性值后，判断预设语料库中是否存在该待填充字段的属性值，如果存在，可以从该属性值对应的内容中随机选择一个参考内容作为待填充内容。如果不存在，则继续判断预设语料库中是否存在与待填充字段的属性值的差异小于预设值的参考属性值，如果存在，可以从该参考属性值对应的内容中随机选择一个参考内容作为待填充内容。

[0063] 其中，与待填充字段的属性值的差异小于预设值的参考属性值可以是待填充字段的属性值的近义词。示例性的，待填充字段的属性值为“姓名”，则参考属性值可以为“名称”等。

[0064] 本实施例中，当预设语料库中存在待填充字段的属性值或者与该属性值相近的参考属性值时，可以随机选择该属性值或该参考属性值对应的内容作为目标填充内容，既能使得目标填充内容符合待填充字段的属性，即与待填充字段语义一致或相近，同时，又可以保证目标填充内容的多样性。

[0065] 作为另一种可选的实施方式，如果经过前述的判断过程确认预设语料库中既不存在待填充字段的属性值也不存在与待填充字段的属性值的差异小于预设值的参考属性值，则确定待填充字段的历史内容的词性，并根据预设语料库中是否存在待填充字段的历史内容的词性对应的内容，生成待填充字段的目标填充内容。

[0066] 可选的，可以对待填充字段的历史内容进行词性分析，得到待填充字段的词性。

[0067] 进而，可以根据预设语料库中是否存在待填充字段的历史内容的词性对应的内容，生成待填充字段的目标填充内容。

[0068] 本实施例中，通过确定历史内容的词性，以及根据预设语料库中是否存在待填充字段的历史内容的词性对应的内容生成目标填充内容，可以使得在预设语料库中不存在待填充字段的属性值的内容时尽可能得到与待填充字段的实际内容较为接近的填充内容。

[0069] 在根据预设语料库中是否存在待填充字段的历史内容的词性对应的内容，生成待填充字段的目标填充内容时，可以通过如下过程实现。

[0070] 一种可选方式中，如果预设语料库中存在待填充字段的历史内容的词性对应的内容，则从待填充字段的历史内容的词性对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的所述目标填充内容。

[0071] 可选的，预设语料库中可以记录所有语料的词性。当确定待填充字段的历史内容的词性后，从历史内容的词性对应的内容中随机选择出参考内容作为目标填充内容。

[0072] 示例性的，待填充字段的历史内容的词性为名词，则通过该方式所选择出的目标填充内容为一名词，而不会是其他词性，例如动词或形容词。

[0073] 通过上述方式，使得在预设语料库中不存在待填充字段的属性值的内容时尽可能得到与待填充字段的实际内容较为接近的填充内容，而避免生成的目标填充内容与待填充字段的实际内容偏差较大。

[0074] 另一种可选方式中，如果预设语料库中不存在待填充字段的历史内容的词性对应的内容，则将待填充字段的历史内容作为待填充字段的目标填充内容。

[0075] 如果前述的条件均不满足，则说明无法从预设语料库中找到可以适用于待填充字段的内容，在这种情况下，可以直接将历史内容作为待填充字段的目标填充内容。

[0076] 通过这种方式，可以避免生成的目标填充内容与待填充字段的实际内容偏差较大。

[0077] 图4为基于预设语料库生成待填充字段的目标填充内容的流程示意图，如图4所示，生成流程可以包括：

[0078] S401、判断预设语料库中是否存在待填充字段的属性值，若是，则执行步骤S402，否则，执行步骤S403。

[0079] S402、从预设语料库中待填充字段的属性值对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容。

[0080] S403、判断预设语料库中是否存在与待填充字段的属性值的差异小于预设值的参考属性值，若是，则执行步骤S404，否则，执行步骤S405。

[0081] S404、从预设语料库中参考属性值对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容。

[0082] S405、确定待填充字段的历史内容的词性。

[0083] S406、判断预设语料库中是否存在待填充字段的历史内容的词性对应的内容，若是，则执行步骤S407，否则，执行步骤S408。

[0084] S407、从待填充字段的历史内容的词性对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容。

[0085] S408、将待填充字段的历史内容作为待填充字段的目标填充内容。

[0086] 上述各步骤的具体执行过程可以参见前述实施例，此处不再赘述。

[0087] 以下对上述步骤S302中根据上述待填充字段在模板图像中的历史位置进行位置调整，得到上述待填充字段的目标填充位置的过程进行说明。

[0088] 图5为本申请实施例提供的结构化文档信息标注的方法的流程示意图，如图5所示，上述步骤S302的一种可选方式包括：

[0089] S501、在待填充字段在模板图像中的历史位置所在的区域内，对待填充字段的位置进行调整。

[0090] 示例性的，待填充字段的历史位置在所在区域的中心位置，则可以以将该中心位置向左上方移动一定数量的像素，并记录移动后的位置。

[0091] 本步骤为对每一个待填充字段单独进行位置调整。具体实施过程中，可以选择部分或全部的待填充字段，并针对每个字段分别进行上述的调整。

[0092] S502、对模板图像中至少一个待填充字段按照相同方向和距离进行位置调整。

[0093] 本步骤为对模板图像中部分或全部的待填充字段统一进行位置调整。

[0094] 在统一进行位置调整时，将这些待填充字段的位置按照相同方向和距离进行移动，即每个待填充字段的位置移动值相同。

[0095] S503、判断待填充字段的位置调整后的位置与待填充字段之外的字段是否存在重叠，若是，则执行步骤S504，否则，将上述步骤S502所得到的调整后的位置作为待填充字段的目标填充位置。

[0096] S504、对存在重叠的待填充字段的位置进行调整，得到待填充字段的目标填充位置。

[0097] 在经过上述步骤S501和S501的位置调整后，可能存在某个待填充字段的位置与其他待填充字段的位置重叠的现象，因此，可以对存在重叠的待填充字段的位置进行调整，得到待填充字段的目标填充位置。

[0098] 可选的，当待填充字段某其他某个字段存在位置重叠时，可以调整待填充字段的位置，或者，也可以调整与其重叠的其他某个字段的位置。

[0099] 值得说明的是，上述步骤S501-S503中的位置调整均可以是随机调整，即进行随机的位置调整。

[0100] 本实施例中，通过对待填充字段进行单独位置调整、统一位置调整以及避免重叠的位置调整，可以实现标注的结构化文档中各字段的位置的多样性。

[0101] 在上述步骤S502中对各待填充字段进行统一位置调整时，如果待填充字段为表格中的字段，则执行如下至少一项：

[0102] 对该表格的所有行按照相同方向和距离进行位置调整、对该表格的每一列分别按照相同方向和距离进行位置调整。

[0103] 通过这种方式，可以使得标注的结构化文档中的表格信息呈现多样化，有助于后续的算法训练。

[0104] 以下说明上述步骤S203中基于目标填充信息得到标注的结构化文档的图像的过程。

[0105] 一种可选方式中，可以在模板图像的目标填充位置上，填充目标填充内容，得到标注的结构化文档的图像。

[0106] 这种方式中无需考虑待填充字段的格式。

[0107] 另一种可选方式中，可以在模板图像的目标填充位置上，按照目标填充格式填充目标填充内容，得到标注的结构化文档的图像。

[0108] 在这种方式中，目标填充内容以特定的目标填充格式填充至目标填充位置上，即考虑了填充内容的格式。

[0109] 使用这种方式，可以使得标注的结构化文档的图像中字段的格式多样化，有助于后续的算法训练。

[0110] 其中，目标填充格式可以在生成目标填充内容和目标填充位置的同时通过随机的方式生成。

[0111] 作为一种可选的实施方式，前述的标注信息中还可以包括待填充字段的颜色。相应的，在基于目标填充信息得到标注的结构化文档的图像时，可以在模板图像的目标填充位置上，按照目标填充格式以及待填充字段的颜色，填充目标填充内容，得到标注的结构化文档的图像。

[0112] 示例性的，标注的待填充字段的颜色为红色，则在标注的结构化文档的图像中所填充的目标填充内容的颜色为红色。

[0113] 通过这种方式，使得标注的结构化文档的图像中字段的信息更加丰富，有助于后续的算法训练。

[0114] 上述在得到标注的结构化文档的图像时，可以对每个待填充字段按照字体、内容等生成文字切片灰度图。再将二值文字切片灰度图随机选取模糊、直方图均衡、灰度变换等图像处理操作进行图像变换，并通过柏松融合或者alpha通道融合将二值文字切片灰度图按照字体颜色与模板图像的背景相融合，从而达到“写”文字的效果。

[0115] 进一步的，还可以对上述得到的标注的结构化文档的图像进行数据增强。示例性的，进行添加噪声、缩放、弹性形变、颜色变换等数据增强操作。

[0116] 以下说明前述实施例中涉及的模板图像以及标注信息的生成过程。

[0117] 可选的，可以由用户预先选择一个结构化文档并拍摄该文档的图像，得到原始图像。进而，基于人工或者工具标注，标注出在该原始文档中需要进行填充的待填充字段的属性值、历史内容以及历史位置。

[0118] 进而，用户可以将这些信息输入电子设备。

[0119] 电子设备接收到这些信息后，一方面，可以根据待填充字段的历史位置，对该历史位置所在区域进行涂抹处理，得到模板图像。另一方面，将这些输入的信息作为模板图像的标注信息。

[0120] 其中，待填充字段在原始图像中的历史位置即为待填充字段在模板图像中的历史位置。

[0121] 示例性的，可以通过自适应阈值对原始图像灰度图中历史位置所在待涂抹区域附近进行二值化。在灰度级小值区域且不在待涂抹区域内的像素记为参考区域。进而，获取待涂抹区域轮廓周围属于参考区域的像素作为参考像素，根据参考像素按照计算权重对待涂抹区域轮廓上像素点进行替换，达到擦除的效果。替换后的像素合并入参考区域。不断重复，直至待涂抹区域全部替换完毕。

[0122] 图6为本申请实施例提供的一种结构化文档信息标注的装置的模块结构图，如图6所示，该装置包括：

[0123] 获取模块601，用于获取结构化文档的模板图像以及所述模板图像的至少一个待填充字段的标注信息，所述标注信息包括所述待填充字段的属性值、所述待填充字段的历史内容以及所述待填充字段在所述模板图像中的历史位置。

[0124] 处理模块602，用于根据所述待填充字段的属性值、所述待填充字段的历史内容以

及所述待填充字段在模板图像中的历史位置，生成所述待填充字段的目标填充信息，所述目标填充信息包括：目标填充位置、目标填充内容以及目标填充格式；以及，根据所述待填充字段的目标填充信息，得到标注的结构化文档的图像。

[0125] 在一种可选的实施方式中，处理模块602具体用于：

[0126] 根据所述待填充字段的属性值和历史内容，生成所述待填充字段的目标填充内容；以及，根据所述待填充字段在模板图像中的历史位置进行位置调整，得到所述待填充字段的目标填充位置。

[0127] 在一种可选的实施方式中，处理模块602具体用于：

[0128] 若预设语料库中存在所述待填充字段的属性值，则从所述预设语料库中所述待填充字段的属性值对应的内容中随机选择一参考内容，将所述参考内容作为所述待填充字段的所述目标填充内容；若预设语料库中不存在所述待填充字段的属性值，同时，存在与所述待填充字段的属性值的差异小于预设值的参考属性值，则从所述预设语料库中所述参考属性值对应的内容中随机选择一参考内容，将所述参考内容作为所述待填充字段的所述目标填充内容。

[0129] 在一种可选的实施方式中，处理模块602具体用于：

[0130] 若预设语料库中既不存在所述待填充字段的属性值也不存在与所述待填充字段的属性值的差异小于预设值的参考属性值，则确定所述待填充字段的历史内容的词性，并根据所述预设语料库中是否存在所述待填充字段的历史内容的词性对应的内容，生成所述待填充字段的目标填充内容。

[0131] 在一种可选的实施方式中，处理模块602具体用于：

[0132] 若所述预设语料库中存在所述待填充字段的历史内容的词性对应的内容，则从所述待填充字段的历史内容的词性对应的内容中随机选择一参考内容，将所述参考内容作为所述待填充字段的所述目标填充内容。

[0133] 在一种可选的实施方式中，处理模块602具体用于：

[0134] 若所述预设语料库中不存在所述待填充字段的历史内容的词性对应的内容，则将所述待填充字段的历史内容作为所述待填充字段的所述目标填充内容。

[0135] 在一种可选的实施方式中，处理模块602具体用于：

[0136] 在所述待填充字段在模板图像中的历史位置所在的区域内，对所述待填充字段的位置进行调整；以及，对所述模板图像中至少一个待填充字段按照相同方向和距离进行位置调整；以及，若所述待填充字段的位置调整后的位置与所述待填充字段之外的字段存在重叠，则对存在重叠的待填充字段的位置进行调整，得到所述待填充字段的目标填充位置。

[0137] 在一种可选的实施方式中，处理模块602具体用于：

[0138] 若所述待填充字段为表格中的字段，则执行如下至少一项：

[0139] 对所述表格的所有行按照相同方向和距离进行位置调整、对所述表格的每一列分别按照相同方向和距离进行位置调整。

[0140] 在一种可选的实施方式中，处理模块602具体用于：

[0141] 对所述模板图像中的所有待填充字段按照相同方向和距离进行位置调整。

[0142] 在一种可选的实施方式中，处理模块602具体用于：

[0143] 在所述模板图像的所述目标填充位置上，按照所述目标填充格式填充所述目标填

充内容，得到标注的结构化文档的图像。

[0144] 在一种可选的实施方式中，所述标注信息还包括：待填充字段的颜色。

[0145] 处理模块602具体用于：

[0146] 在所述模板图像的所述目标填充位置上，按照所述目标填充格式以及所述待填充字段的颜色，填充所述目标填充内容，得到标注的结构化文档的图像。

[0147] 在一种可选的实施方式中，处理模块602还用于：

[0148] 根据用户指示的待填充字段的属性值、待填充字段的历史内容以及每个待填充字段在原始图像中的历史位置，对所述原始图像进行涂抹处理，得到所述模板图像以及所述模板图像的至少一个标注信息。

[0149] 根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

[0150] 根据本申请的实施例，本申请还提供了一种计算机程序产品，程序产品包括：计算机程序，计算机程序存储在可读存储介质中，电子设备的至少一个处理器可以从可读存储介质读取计算机程序，至少一个处理器执行计算机程序使得电子设备执行上述任一实施例提供的方案。该电子设备例如可以是服务器。

[0151] 如图7所示，是根据本申请实施例的结构化文档信息标注的方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0152] 如图7所示，该电子设备包括：一个或多个处理器701、存储器702，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如，耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作(例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图7中以一个处理器701为例。

[0153] 存储器702即为本申请所提供的非瞬时计算机可读存储介质。其中，所述存储器存储有可由至少一个处理器执行的指令，以使所述至少一个处理器执行本申请所提供的结构化文档信息标注的方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的结构化文档信息标注的方法。

[0154] 存储器702作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的结构化文档信息标注的方法对应的程序指令/模块(例如，附图6所示的获取模块601和处理模块602)。处理器701通过运行存储在存储器702中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的结构化文档信息标注的方法。

[0155] 存储器702可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据结构化文档信息标注的电子

设备的使用所创建的数据等。此外，存储器702可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器702可选包括相对于处理器701远程设置的存储器，这些远程存储器可以通过网络连接至结构化文档信息标注的电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0156] 结构化文档信息标注的方法的电子设备还可以包括：输入装置703和输出装置704。处理器701、存储器702、输入装置703和输出装置704可以通过总线或者其他方式连接，图7中以通过总线连接为例。

[0157] 输入装置703可接收输入的数字或字符信息，以及产生与结构化文档信息标注的电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置704可以包括显示设备、辅助照明装置(例如，LED)和触觉反馈装置(例如，振动电机)等。该显示设备可以包括但不限于，液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

[0158] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0159] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令，并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的，术语“机器可读介质”和“计算机可读介质”指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如，磁盘、光盘、存储器、可编程逻辑装置(PLD))，包括，接收作为机器可读信号的机器指令的机器可读介质。术语“机器可读信号”指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0160] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0161] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数

字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0162] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0163] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0164] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和其他因素,可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

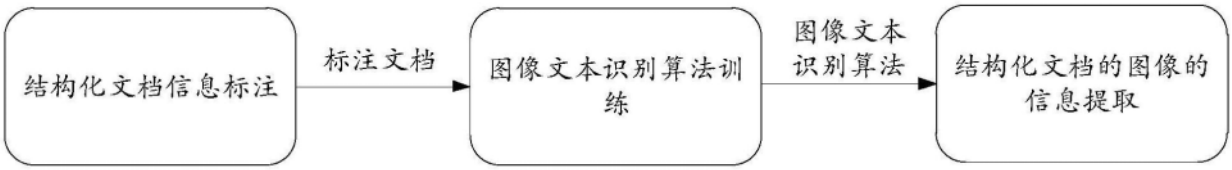

结构化文档信息标注
标注文档
图像文本识别算法训练
图像文本识别算法
结构化文档的图像的信息提取

图1

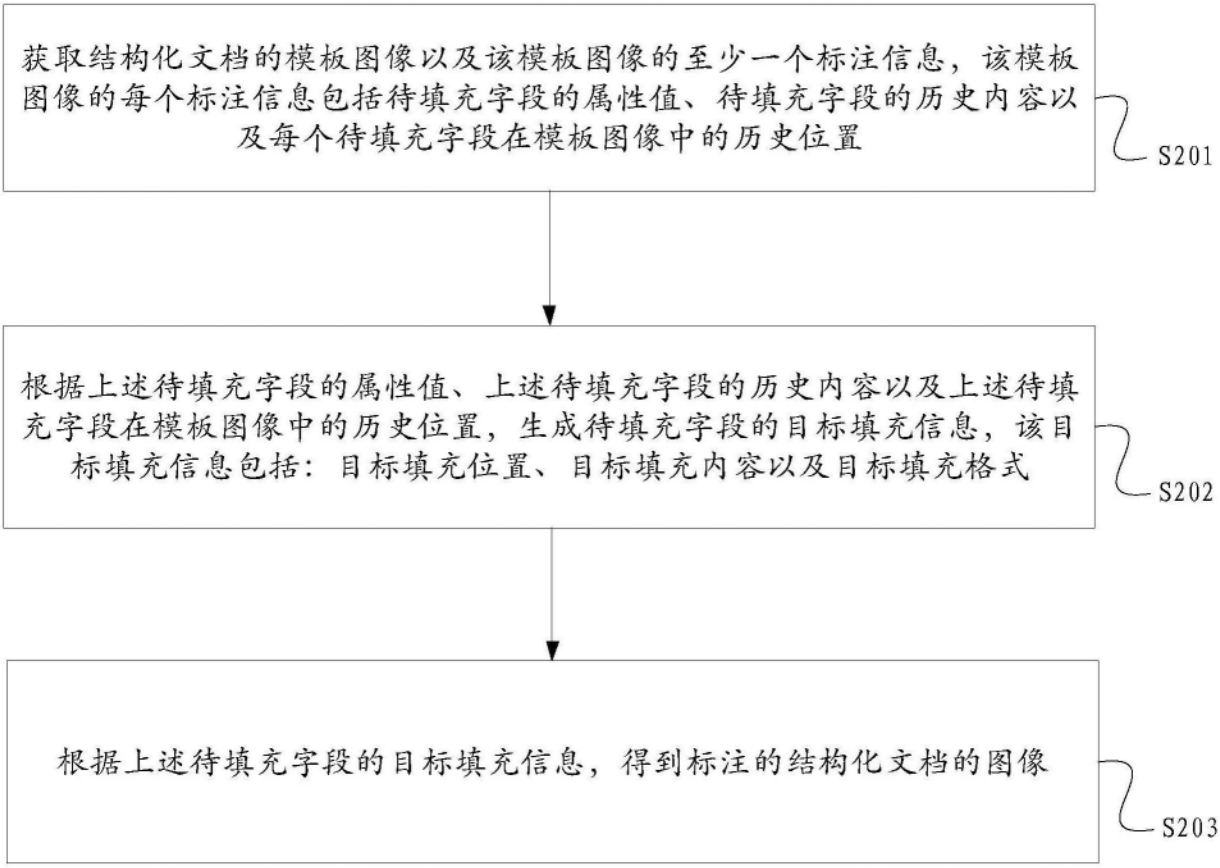

获取结构化文档的模板图像以及该模板图像的至少一个标注信息，该模板图像的每个标注信息包括待填充字段的属性值、待填充字段的历史内容以及每个待填充字段在模板图像中的历史位置
S201
根据上述待填充字段的属性值、上述待填充字段的历史内容以及上述待填充字段在模板图像中的历史位置，生成待填充字段的目标填充信息，该目标填充信息包括：目标填充位置、目标填充内容以及目标填充格式
S202
根据上述待填充字段的目标填充信息，得到标注的结构化文档的图像
S203

图2

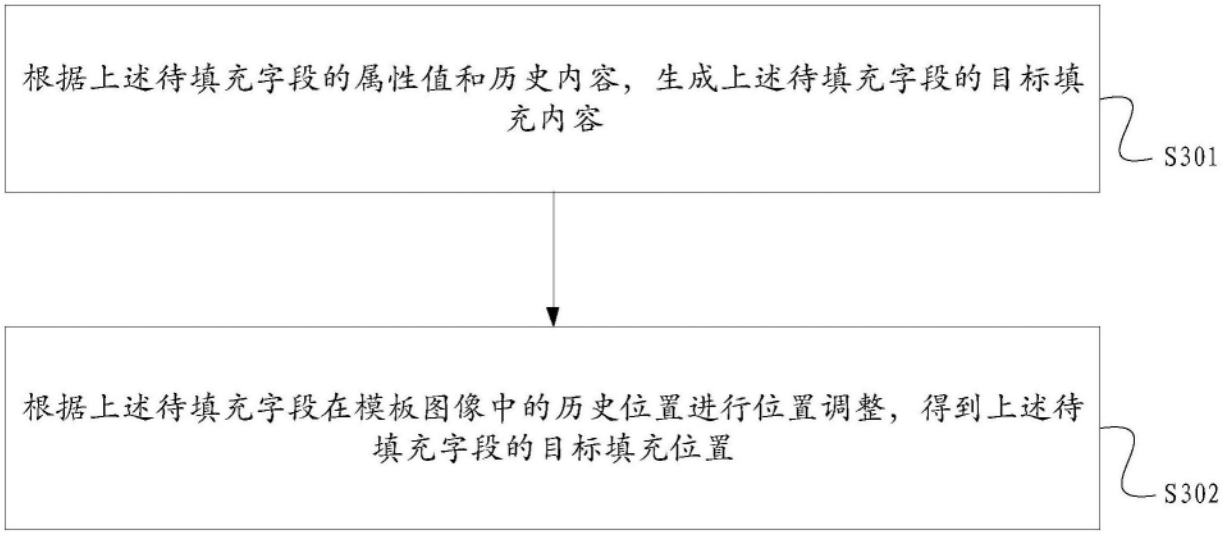

根据上述待填充字段的属性值和历史内容，生成上述待填充字段的目标填充内容
S301
根据上述待填充字段在模板图像中的历史位置进行位置调整，得到上述待填充字段的目标填充位置
S302

图3

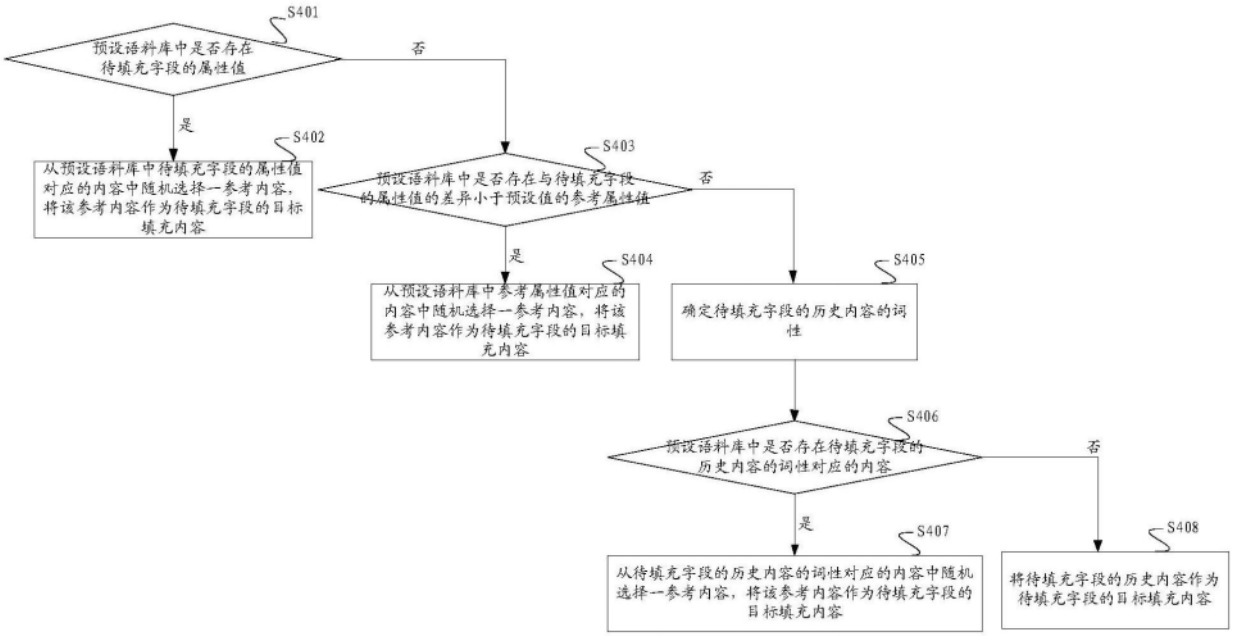

S401
预设语料库中是否存在待填充字段的属性值
否
是
S402
从预设语料库中待填充字段的属性值对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容
S403
预设语料库中是否存在与待填充字段的属性值的差异小于预设值的参考属性值
否
是
S404
从预设语料库中参考属性值对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容
S405
确定待填充字段的历史内容的词性
S406
预设语料库中是否存在待填充字段的历史内容的词性对应的内容
否
是
S407
从待填充字段的历史内容的词性对应的内容中随机选择一参考内容，将该参考内容作为待填充字段的目标填充内容
S408
将待填充字段的历史内容作为待填充字段的目标填充内容

图4

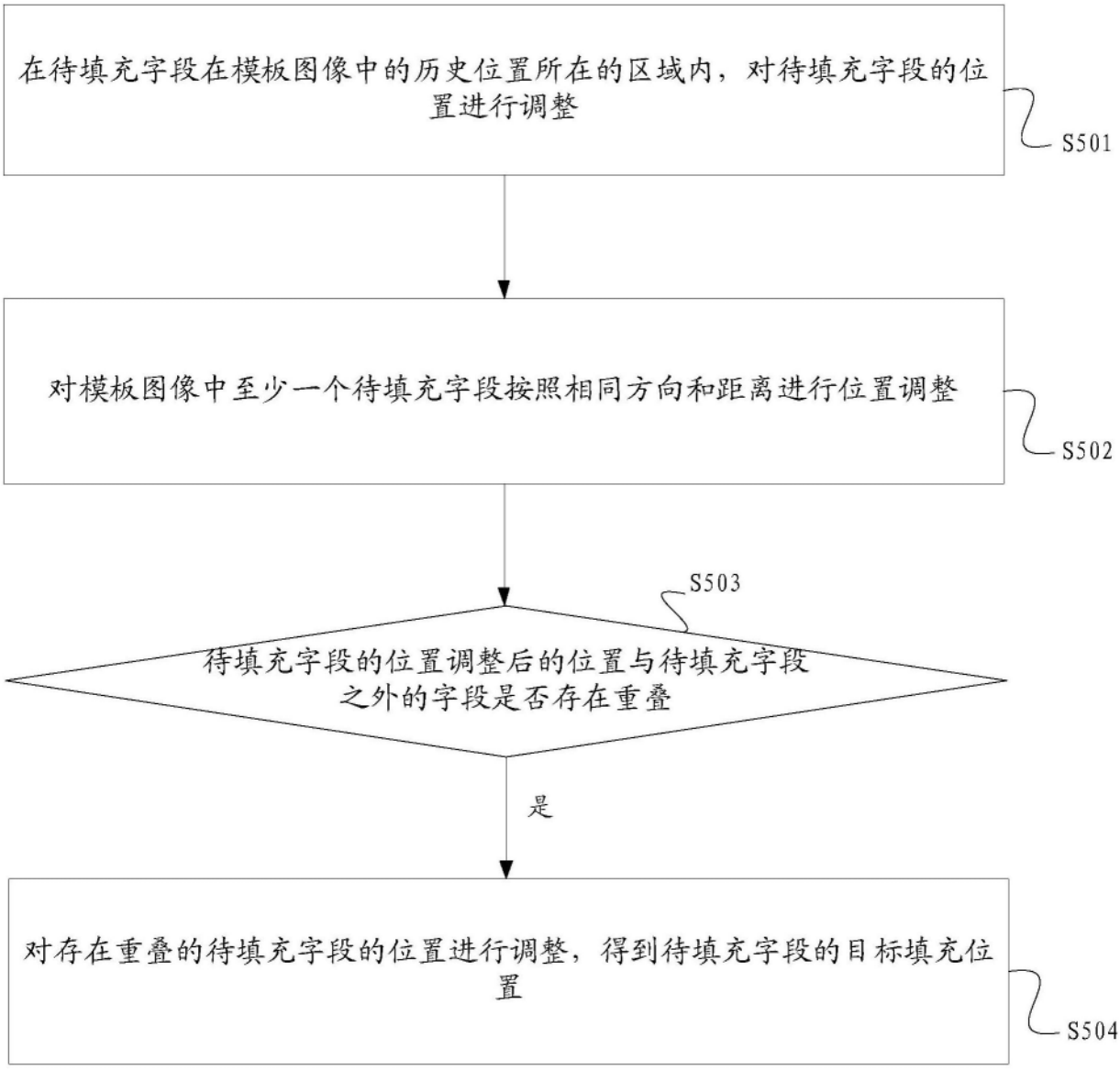

在待填充字段在模板图像中的历史位置所在的区域内，对待填充字段的位置进行调整
S501
对模板图像中至少一个待填充字段按照相同方向和距离进行位置调整
S502
S503
待填充字段的位置调整后的位置与待填充字段之外的字段是否存在重叠
是
对存在重叠的待填充字段的位置进行调整，得到待填充字段的目标填充位置
S504

图5

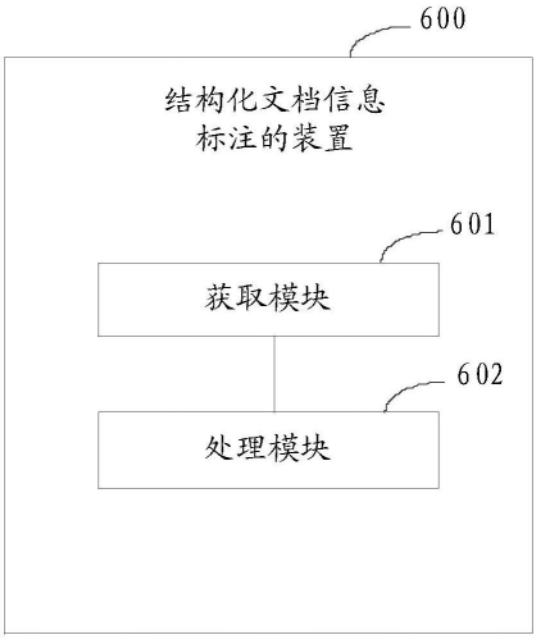

600
结构化文档信息
标注的装置
601
获取模块
602
处理模块

图6

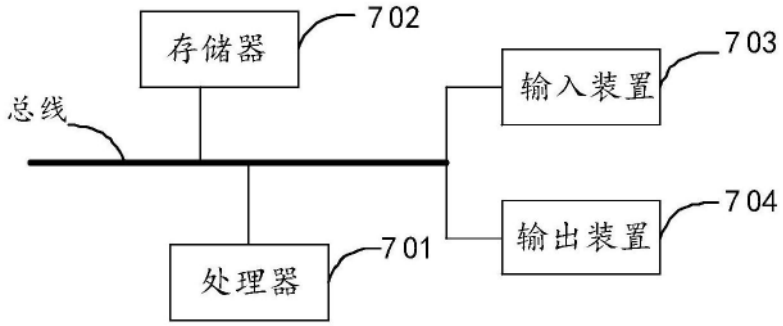

存储器
702
总线
输入装置
703
输出装置
704
处理器
701

图7