(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115761778 A
(43)申请公布日 2023.03.07

(21)申请号 202211483546.8

(22)申请日 2022.11.24

(71)申请人 联仁健康医疗大数据科技股份有限公司
地址 200131 上海市浦东新区中国(上海)自由贸易试验区川和路55弄3号

(72)发明人 黎安

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 侯军洋

(51)Int.Cl.
G06V 30/414(2022.01)
G06V 30/40(2022.01)
G06F 40/109(2020.01)

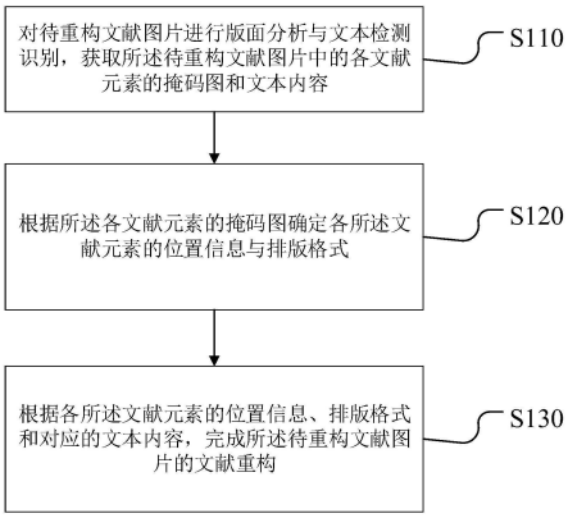
权利要求书2页 说明书9页 附图5页

(54)发明名称

一种文献重构方法、装置、设备和存储介质

(57)摘要

本发明实施例公开了一种文献重构方法、装置、设备和存储介质，其中，方法包括：对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容；根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式；根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。本发明实施例的技术方案解决了现有技术在进行文献重构时无法识别文献图片中文献元素排版格式的问题，可以确定文献图片中每一种文献元素的排版格式，提高文献重构的准确性。



CN 115761778 A

1.一种文献重构方法,其特征在于,所述方法包括:

对待重构文献图片进行版面分析与文本检测识别,获取所述待重构文献图片中的各文献元素的掩码图和文本内容;

根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式;

根据各所述文献元素的位置信息、排版格式和对应的文本内容,完成所述待重构文献图片的文献重构。

2.根据权利要求1所述的方法,其特征在于,所述根据所述各文献元素的掩码图确定各所述文献元素的排版格式,包括:

基于所述各文献元素的掩码图的中心点位置确定所述各文献元素的掩码图与所述待重构文献图片的横截距比值;

根据所述横截距比值与预设排版格式阈值标准的对比结果,确定所述各文献元素的排版格式。

3.根据权利要求1所述的方法,其特征在于,针对所述文献元素中的文本形式的文献元素,根据所述各文献元素的掩码图确定各所述文献元素的位置信息,包括:

识别所述文本形式的文献元素的文本段落,并确定各文本段落的文本框;

计算所述各文本段落的文本框与对应文本段落的掩码图之间的位置区域交并比,得到目标交并比数据;

根据所述目标交并比数据与预设交并比检测阈值的对比结果,对所述各文本段落位置进行校验,以确定对应的目标文本位置。

4.根据权利要求3所述的方法,其特征在于,根据所述目标交并比数据与预设交并比检测阈值的对比结果,对所述各文本段落位置进行校验,以确定对应的目标文本位置,包括:

当所述交并比大于预设交并比检测阈值时,将所述文本段落位置作为目标文本位置;

当所述交并比小于预设交并比检测阈值时,不将所述文本段落位置作为目标文本位置。

5.根据权利要求3所述的方法,其特征在于,所述文本形式的文献元素包括:文本、表格、表标题、表注释、图标题、图注释。

6.根据权利要求1所述的方法,其特征在于,所述方法还包括:

当文献元素为标题时,将所述文献元素与预设目录标题模板中的标题项进行匹配,确定所述文献元素的标题目录级别。

7.根据权利要求1所述的方法,其特征在于,在对所述待重构文献图片进行版面分析与文本检测识别之前,还包括:

对所述待重构文献图片进行边缘去除和内容校正处理,完成对所述待重构文献图片的预处理。

8.一种文献重构装置,其特征在于,所述装置包括:

文献元素识别模块,用于对待重构文献图片进行版面分析与文本检测识别,获取所述待重构文献图片中的各文献元素的掩码图和文本内容;

文献元素分析模块,用于根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式;

文献元素重构模块,用于根据各所述文献元素的位置信息、排版格式和对应的文本内

容,完成所述待重构文献图片的文献重构。

9.一种计算机设备,其特征在于,所述计算机设备包括:

一个或多个处理器;

存储器,用于存储一个或多个程序;

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-7中任一所述的文献重构方法。

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,该程序被处理器执行时实现如权利要求1-7中任一所述的文献重构方法。

一种文献重构方法、装置、设备和存储介质

技术领域

[0001] 本发明实施例涉及图像识别技术领域，尤其涉及一种文献重构方法、装置、设备和存储介质。

背景技术

[0002] 近些年来，图像识别技术发展迅速，图像识别技术可以通过识别和分析传统纸质文件图像中的要素，将纸质文件重构成为电子文件。现有技术通常利用OCR(Optical Character Recognition，光学字符识别)技术进行纸质文件的重构，但是对于文献这类专业性较强的文件，利用OCR技术无法很好的实现文献的重构效果，例如，OCR技术无法识别文献中的不同元素的排版格式，对文献中的元素进行重构。

发明内容

[0003] 本发明实施例提供了一种文献重构方法、装置、设备和存储介质，可以确定文献图片中每一种文献元素的排版格式，提高文献重构的准确性。

[0004] 第一方面，本发明实施例提供了一种文献重构方法，该方法包括：

[0005] 对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容；

[0006] 根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式；

[0007] 根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。

[0008] 第二方面，本发明实施例提供了一种文献重构装置，该装置包括：

[0009] 文献元素识别模块，用于对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容；

[0010] 文献元素分析模块，用于根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式；

[0011] 文献元素重构模块，用于根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。

[0012] 第三方面，本发明实施例提供了一种计算机设备，该计算机设备包括：

[0013] 一个或多个处理器；

[0014] 存储器，用于存储一个或多个程序；

[0015] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现任一实施例所述的文献重构方法。

[0016] 第四方面，本发明实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现任一实施例所述的文献重构方法。

[0017] 本发明实施例所提供的技术方案，通过对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容；根据所述各文献

元素的掩码图确定各所述文献元素的位置信息与排版格式;根据各所述文献元素的位置信息、排版格式和对应的文本内容,完成所述待重构文献图片的文献重构。本发明实施例的技术方案解决了现有技术在进行文献重构时无法识别文献图片中文献元素排版格式的问题,可以确定文献图片中每一种文献元素的排版格式,提高文献重构的准确性。

附图说明

[0018] 图1是本发明实施例提供的一种文献重构方法流程图;

[0019] 图2是本发明实施例提供的又一种文献重构方法流程图;

[0020] 图3是本发明实施例提供的一种进行文献重构的工作流程图;

[0021] 图4是本发明实施例提供的一种文献重构装置的结构示意图;

[0022] 图5是本发明实施例提供的一种计算机设备的结构示意图。

具体实施方式

[0023] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0024] 图1是本发明实施例提供的一种文献重构方法流程图,本发明实施例可适用于对文献图片进行重构的场景中,该方法可以由文献重构装置执行,该装置可以由软件和/或硬件的方式来实现。

[0025] 如图1所示,文献重构方法包括以下步骤:

[0026] S110、对待重构文献图片进行版面分析与文本检测识别,获取所述待重构文献图片中的各文献元素的掩码图和文本内容。

[0027] 其中,待重构文献图片可以是需要进行重构处理的文献图片,通过重构处理可以将非电子版的文献内容重构为电子文件格式。文献元素可以是文献的元素组成部分,例如,文献元素可以包括页眉、标题、文本、表标题、表、表注释、图标题、图、图注释、公式、编者名、页脚等。掩码图可以是标记文献元素所在位置的图像,通过掩码图可以确定各文献元素在待重构文献图片中的位置,具体的,可以通过yolox(目标检测算法)算法对待重构文献图片进行版面分析,确定待重构文献图片中的各文献元素的掩码图。文本内容可以是文献元素中的内容,具体的,可以采用预设的文本识别算法对各文献元素进行文本检测识别,获取到各文献元素的文本内容。

[0028] S120、根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式。

[0029] 其中,位置信息可以是文献元素在待重构文献图片中位置的信息,因为掩码图可以标记文献元素在待重构文献图片的位置,所以可以根据各文献元素的掩码图确定各文献元素的位置信息。通过确定各文献元素的位置信息,可以方便后续根据文献元素的位置信息将各文献元素的文本内容依次进行排列。进一步,虽然根据各文献元素的掩码图可以确定各文献元素的位置信息,但是为了提高文献重构的精细程度,后续还需要对各文献元素的排版格式进行进一步的判断。排版格式可以是对文献元素的文本内容进行排版所采用的

格式，排版格式可以包括单栏排版格式和双栏排版格式。文献元素的排版格式也可以通过掩码图确定，例如，可以通过各文献元素的掩码图横节距与待重构文献图片横节距的比值确定文献元素的排版格式。

[0030] S130、根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。

[0031] 其中，可以根据文献元素的位置信息、排版格式，依次将各文献元素对应的文本内容进行排列，具体的，可以将不同的文献元素按照对应的形式存入JSON(Java Script Object Notation，JS对象简谱)文件中，针对图片和表格留存URL(Universal Resource Locator，统一资源定位符)地址，完成待重构文献图片的文献重构。

[0032] 本发明实施例所提供的技术方案，通过对待重构文献图片进行版面分析与文本检测识别，获取待重构文献图片中的各文献元素的掩码图和文本内容；根据各文献元素的掩码图确定各文献元素的位置信息与排版格式；根据各文献元素的位置信息、排版格式和对应的文本内容，完成待重构文献图片的文献重构。本发明实施例的技术方案解决了现有技术在进行文献重构时无法识别文献图片中文献元素排版格式的问题，可以确定文献图片中每一种文献元素的排版格式，提高文献重构的准确性。

[0033] 图2是本发明实施例提供的又一种文献重构方法流程图，本发明实施例可适用于对文献图片进行重构的场景中，本实施例在上述实施例的基础上，进一步的说明如何根据各文献元素的掩码图确定各文献元素的排版格式，以及针对文本形式的文献元素，如何根据各文献元素的掩码图确定各文献元素的位置信息，该装置可以由软件和/或硬件的方式来实现，集成于具有应用开发功能的计算机设备中。

[0034] 如图2所示，文献重构方法包括以下步骤：

[0035] S210、对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容。

[0036] 其中，待重构文献图片可以是需要进行重构处理的文献图片，通过重构处理可以将待非电子版的文献内容重构为电子文件格式。文献元素可以是文献的元素组成部分，例如，文献元素可以包括页眉、标题、文本、表标题、表、表注释、图标题、图、图注释、公式、编者名、页脚等。掩码图可以是标记文献元素所在位置的图像，通过掩码图可以确定各文献元素在待重构文献图片中的位置，具体的，可以通过yolox(目标检测算法)算法对待重构文献图片进行版面分析，确定待重构文献图片中的各文献元素的掩码图。文本内容可以是文献元素中的内容，具体的，可以采用预设的文本识别算法对各文献元素进行文本检测识别，获取到各文献元素的文本内容。

[0037] 在一种可选的实施方式中，在对待重构文献图片进行版面分析与文本检测识别之前，还可以对待重构文献图片进行边缘去除和内容校正处理，完成对待重构文献图片的预处理。具体的，可以利用Marior算法中的MRM(Margin Remove Module，边界去除模块)对待重构文献图片进行边缘去除处理，再利用ICRM(Iterative Content Remove Module，迭代式内容矫正模块)对待重构文献图片进行内容校正处理。通过对待重构文献图片的预处理，可以解决待重构文献图片中存在的环境边界过大、环境边界缺失和文献元素形变等问题，提高针对待重构文献图片进行文献重构的准确性。

[0038] S220、根据所述各文献元素的掩码图确定各所述文献元素的位置信息。

[0039] 其中，位置信息可以是文献元素在待重构文献图片中位置的信息，因为掩码图可以标记文献元素在待重构文献图片的位置，所以可以根据各文献元素的掩码图确定各文献元素的位置信息。通过确定各文献元素的位置信息，可以方便后续根据文献元素的位置信息将各文献元素的文本内容依次进行排列。进一步，虽然根据各文献元素的掩码图可以确定各文献元素的位置信息，但是为了提高文献重构的精细程度，后续还需要对各文献元素的排版格式进行进一步的判断。

[0040] 在一种可选的实施方式中，针对文献元素中的文本形式的文献元素，可以识别文本形式的文献元素的文本段落，并确定各文本段落的文本框；计算各文本段落的文本框与对应文本段落的掩码图之间的位置区域交并比，得到目标交并比数据；根据目标交并比数据与预设交并比检测阈值的对比结果，对各文本段落位置进行校验，以确定对应的目标文本位置。

[0041] 其中，文本形式的文献元素包括文本、表格、表标题、表注释、图标题、图注释，相应的，文本段落可以是文本形式的文献元素的文本内容。文本段落的文本框可以是预设的用于对各文本段落位置进行校验的文本框，文本段落的文本框可以由预设的文本检测算法生成。

[0042] 目标交并比数据可以需要进行校验的交并比数据，通过计算各文本段落的文本框与对应文本段落的掩码图之间的交集位置区域，再将交集位置区域除以文本段落的文本框区域或者除以对应文本段落的掩码图区域的比值作为目标交并比数据。预设交并比检测阈值可以是预设的用于对目标交并比数据进行校验的阈值，根据目标交并比数据与预设交并比检测阈值的对比结果，可以确定各文本段落位置是否正确，即实现对各文本段落位置的校验，示例性的，可以将0.8作为预设交并比检测阈值。

[0043] 目标文本位置可以是最终确定的文本形式的文献元素的位置，示例性的，当目标交并比数据大于预设交并比检测阈值时，可以将文本段落位置作为目标文本位置；当目标交并比数据小于预设交并比检测阈值时，不可以将文本段落位置作为目标文本位置。其中，当目标交并比数据小于预设交并比检测阈值时，表示文本段落可能存在其他文献元素的文本内容，因此不可以将文本段落位置作为目标文本位置。通过计算各文本段落的文本框与对应文本段落的掩码图之间的位置区域交并比，并将位置区域交并比与预设交并比检测阈值进行对比，可以对文本形式的文献元素位置进行校验，提高文献重构的准确性。

[0044] S230、基于所述各文献元素的掩码图的中心点位置确定所述各文献元素的掩码图与所述待重构文献图片的横截距比值。

[0045] 其中，中心点位置可以是文献元素的掩码图的中心点在待重构文献图片中的位置，通过各文献元素的掩码图的中心点位置可以确定各文献元素的掩码图与待重构文献图片的横截距比值，进而确定各文献元素的排版格式。横截距比值可以是各文献元素的掩码图横截距与待重构文献图片横截距的比值，横截距比值可以根据文献元素的掩码图的中心点位置确定。例如，当文献元素的掩码图的中心点位置在待重构文献图片横截距的二分之一处时，可以确定文献元素的掩码图与待重构文献图片的横截距比值为1；当文献元素的掩码图的中心点位置在待重构文献图片横截距的四分之一或者四分之三处时，可以确定文献元素的掩码图与待重构文献图片的横截距比值为0.5。

[0046] S240、根据所述横截距比值与预设排版格式阈值标准的对比结果，确定所述各文

献元素的排版格式。

[0047] 其中，预设排版格式阈值标准可以是预设的用于确定文献元素的排版格式的阈值标准，根据横截距比值与预设排版格式阈值标准的对比结果，可以确定各文献元素的排版格式。例如，可以将横截距比值是否大于0.5作为预设排版格式阈值标准，当横截距比值大于0.5时，可以确定文献元素的排版格式为单栏排版；当横截距比值小于0.5时，可以确定文献元素的排版格式为双栏排版。进一步的，当文献元素的排版格式为双栏排版时，可以根据文献元素的掩码图的中心点位置确定文献元素在双栏排版的左侧或者右侧，例如，当文献元素的掩码图的中心点位置在待重构文献图片横截距的四分之一处时，可以确定文献元素在双栏排版的左侧；当文献元素的掩码图的中心点位置在待重构文献图片横截距的四分之三处时，可以确定文献元素在双栏排版的右侧。

[0048] S250、根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。

[0049] 其中，可以根据文献元素的位置信息、排版格式，依次将各文献元素对应的文本内容进行排列，具体的，可以将不同的文献元素按照对应的形式存入JSON文件中，针对图片和表格留存URL地址，完成待重构文献图片的文献重构。

[0050] 在一种可选的实施方式中，当文献元素为标题时，还可以将文献元素与预设目录标题模板中的标题项进行匹配，确定文献元素的标题目录级别。

[0051] 其中，可以是预设目录标题模板预设的用于识别标题的目录级别的模板，当文献元素为标题时，通过将文献元素与预设目录标题模板中的标题项进行匹配，再通过确定与文献元素匹配成功的标题项在预设目录标题模板中的目录级别，可以确定文献元素的标题目录级别，进而得到一个完整的标题目录。

[0052] 示例性的，图3是本发明实施例提供的一种进行文献重构的工作流程图，如图3所示，文献重构的工作流程为：首先输入待重构文献图片，随后对待重构文献图片进行预处理，接着对待重构文献图片进行版面分析；经过版面分析后再识别待重构文献图片的文献元素，其中，针对公式和表格，可以分别选用预设的公式识别算法和公式识别算法进行识别，针对文本可以通过OCR技术进行识别；随后对识别后的文献元素的内容进行段落重构、图表重构、目录重构、排版重构。其中，段落重构可以是对文献元素的位置进行校验；图表重构可以将待重构文献图片中的图片与对应的图注释和图标题进行合并，将待重构文献图片中的表格与对应的表注释和表标题进行合并；目录重构可以是识别待重构文献图片中标题的目录级别，进而得到标题目录；排版重构可以确定待重构文献图片中各文献元素的排版格式。

[0053] 本发明实施例所提供的技术方案，通过对待重构文献图片进行版面分析与文本检测识别，获取待重构文献图片中的各文献元素的掩码图和文本内容；根据各文献元素的掩码图确定各文献元素的位置信息；基于各文献元素的掩码图的中心点位置确定各文献元素的掩码图与待重构文献图片的横截距比值；根据横截距比值与预设排版格式阈值标准的对比结果，确定各文献元素的排版格式；根据各文献元素的位置信息、排版格式和对应的文本内容，完成待重构文献图片的文献重构。本发明实施例的技术方案解决了现有技术在进行文献重构时无法识别文献图片中文献元素排版格式的问题，可以确定文献图片中每一种文献元素的排版格式，提高文献重构的准确性。

[0054] 图5是本发明实施例提供的一种文献重构装置的结构示意图，本发明实施例可适用于对文献图片进行重构的场景中，该装置可以由软件和/或硬件的方式来实现，集成于具有应用开发功能的计算机设备中。

[0055] 如图5所示，文献重构装置包括：文献元素识别模块310、文献元素分析模块320和文献元素重构模块330。

[0056] 其中，文献元素识别模块310，用于对待重构文献图片进行版面分析与文本检测识别，获取待重构文献图片中的各文献元素的掩码图和文本内容；文献元素分析模块320，用于根据各文献元素的掩码图确定各文献元素的位置信息与排版格式；文献元素重构模块330，用于根据各文献元素的位置信息、排版格式和对应的文本内容，完成待重构文献图片的文献重构。

[0057] 本发明实施例所提供的技术方案，通过对待重构文献图片进行版面分析与文本检测识别，获取待重构文献图片中的各文献元素的掩码图和文本内容；根据各文献元素的掩码图确定各文献元素的位置信息与排版格式；根据各文献元素的位置信息、排版格式和对应的文本内容，完成待重构文献图片的文献重构。本发明实施例的技术方案解决了现有技术在进行文献重构时无法识别文献图片中文献元素排版格式的问题，可以确定文献图片中每一种文献元素的排版格式，提高文献重构的准确性。

[0058] 在一种可选的实施方式中，文献元素分析模块320具体用于：基于各文献元素的掩码图的中心点位置确定各文献元素的掩码图与待重构文献图片的横截距比值；根据横截距比值与预设排版格式阈值标准的对比结果，确定各文献元素的排版格式。

[0059] 在一种可选的实施方式中，文献元素分析模块320还用于：针对文献元素中的文本形式的文献元素，识别文本形式的文献元素的文本段落，并确定各文本段落的文本框；计算各文本段落的文本框与对应文本段落的掩码图之间的位置区域交并比，得到目标交并比数据；根据目标交并比数据与预设交并比检测阈值的对比结果，对各文本段落位置进行校验，以确定对应的目标文本位置。

[0060] 在一种可选的实施方式中，文献元素分析模块320还用于：当目标交并比数据大于预设交并比检测阈值时，将文本段落位置作为目标文本位置；当目标交并比数据小于预设交并比检测阈值时，不将文本段落位置作为目标文本位置。

[0061] 在一种可选的实施方式中，文本形式的文献元素包括：文本、表格、表标题、表注释、图标题、图注释。

[0062] 在一种可选的实施方式中，文献元素重构模块330还用于当文献元素为标题时，将文献元素与预设目录标题模板中的标题项进行匹配，确定文献元素的标题目录级别。

[0063] 在一种可选的实施方式中，文献重构装置还包括预处理模块，用于在对待重构文献图片进行版面分析与文本检测识别之前，对待重构文献图片进行边缘去除和内容校正处理，完成对待重构文献图片的预处理。

[0064] 本发明实施例所提供的文献重构装置可执行本发明任意实施例所提供的文献重构方法，具备执行方法相应的功能模块和有益效果。

[0065] 图5为本发明实施例提供的一种计算机设备的结构示意图。图5示出了适于用来实现本发明实施方式的示例性计算机设备12的框图。图5显示的计算机设备12仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。计算机设备12可以任意具有计

算能力的终端设备,可以与配置于文献重构设备中。

[0066] 如图5所示,计算机设备12以通用计算设备的形式表现。计算机设备12的组件可以包括但不限于:一个或者多个处理器或者处理单元16,系统存储器28,连接不同系统组件(包括系统存储器28和处理单元16)的总线18。

[0067] 总线18可以是几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(ISA)总线,微通道体系结构(MAC)总线,增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0068] 计算机设备12典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机设备12访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0069] 系统存储器28可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(RAM)30和/或高速缓存32。计算机设备12可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存储系统34可以用于读写不可移动的、非易失性磁介质(图5未显示,通常称为“硬盘驱动器”)。尽管图5中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线18相连。系统存储器28可以包括至少一个程序产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0070] 具有一组(至少一个)程序模块42的程序/实用工具40,可以存储在例如系统存储器28中,这样的程序模块42包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42通常执行本发明所描述的实施例中的功能和/或方法。

[0071] 计算机设备12也可以与一个或多个外部设备14(例如键盘、指向设备、显示器24等)通信,还可与一个或者多个使得用户能与该计算机设备12交互的设备通信,和/或与使得该计算机设备12能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22进行。并且,计算机设备12还可以通过网络适配器20与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器20通过总线18与计算机设备12的其它模块通信。应当明白,尽管图5中未示出,可以结合计算机设备12使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0072] 处理单元16通过运行存储在系统存储器28中的程序,从而执行各种功能应用以及数据处理,例如实现本发实施例所提供的文献重构方法,该方法包括:

[0073] 对待重构文献图片进行版面分析与文本检测识别,获取所述待重构文献图片中的各文献元素的掩码图和文本内容;

[0074] 根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式;

[0075] 根据各所述文献元素的位置信息、排版格式和对应的文本内容,完成所述待重构

文献图片的文献重构。

[0076] 本实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现如本发明任意实施例所提供的文献重构方法，包括：

[0077] 对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容；

[0078] 根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式；

[0079] 根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构。

[0080] 本发明实施例的计算机存储介质，可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是但不限于：电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0081] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0082] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0083] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，程序设计语言包括面向对象的程序设计语言，诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言，诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络，包括局域网(LAN)或广域网(WAN)，连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0084] 本领域普通技术人员应该明白，上述的本发明的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个计算装置上，或者分布在多个计算装置所组成的网络上，可选地，他们可以用计算机装置可执行的程序代码来实现，从而可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件的结合。

[0085] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，

本发明不限于这里的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

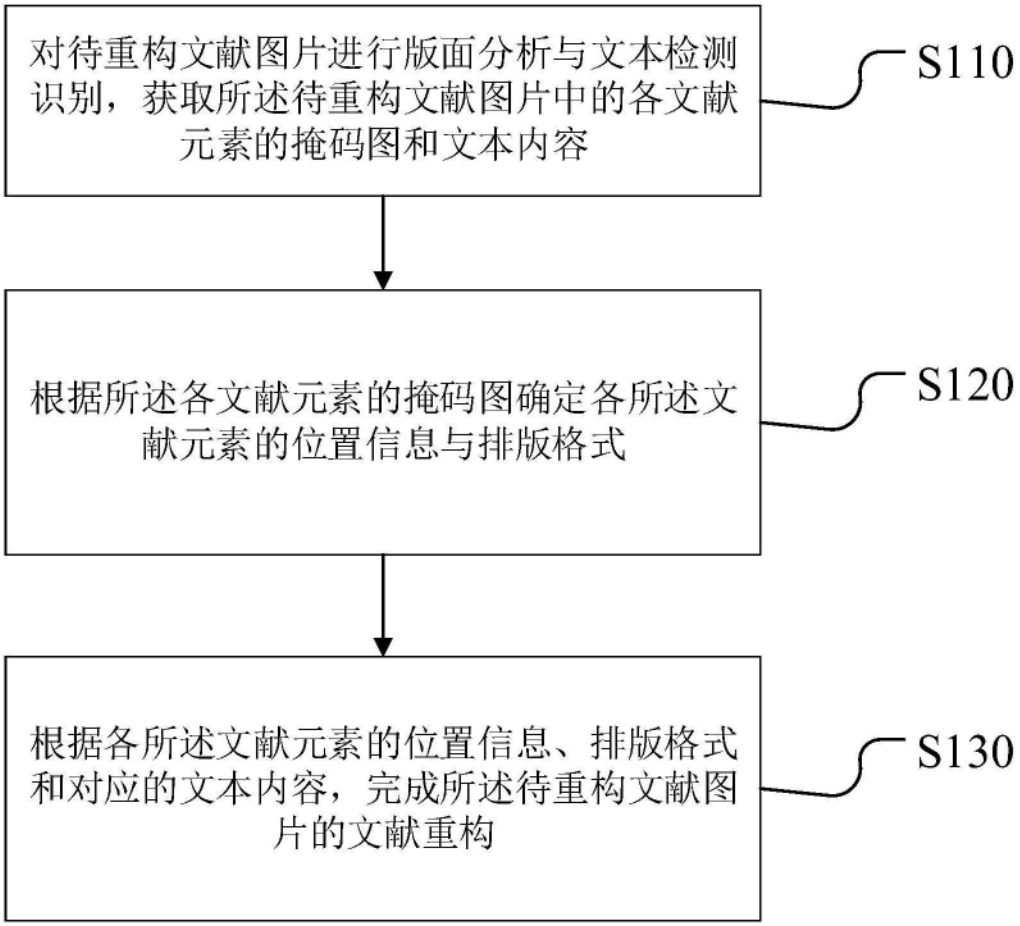

对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容
S110
根据所述各文献元素的掩码图确定各所述文献元素的位置信息与排版格式
S120
根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构
S130

图1

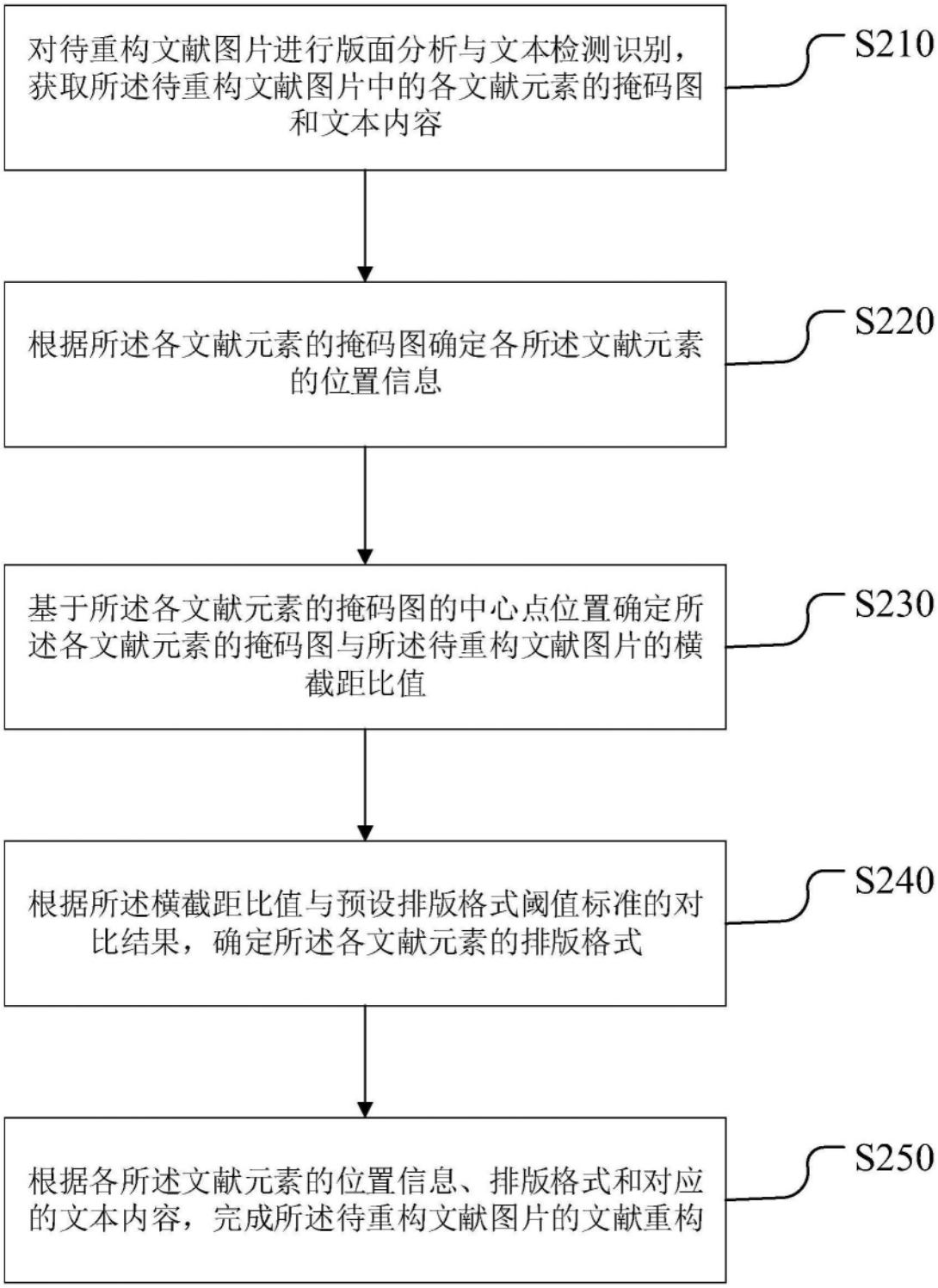

对待重构文献图片进行版面分析与文本检测识别，获取所述待重构文献图片中的各文献元素的掩码图和文本内容
S210
根据所述各文献元素的掩码图确定各所述文献元素的位置信息
S220
基于所述各文献元素的掩码图的中心点位置确定所述各文献元素的掩码图与所述待重构文献图片的横截距比值
S230
根据所述横截距比值与预设排版格式阈值标准的对比结果，确定所述各文献元素的排版格式
S240
根据各所述文献元素的位置信息、排版格式和对应的文本内容，完成所述待重构文献图片的文献重构
S250

图2

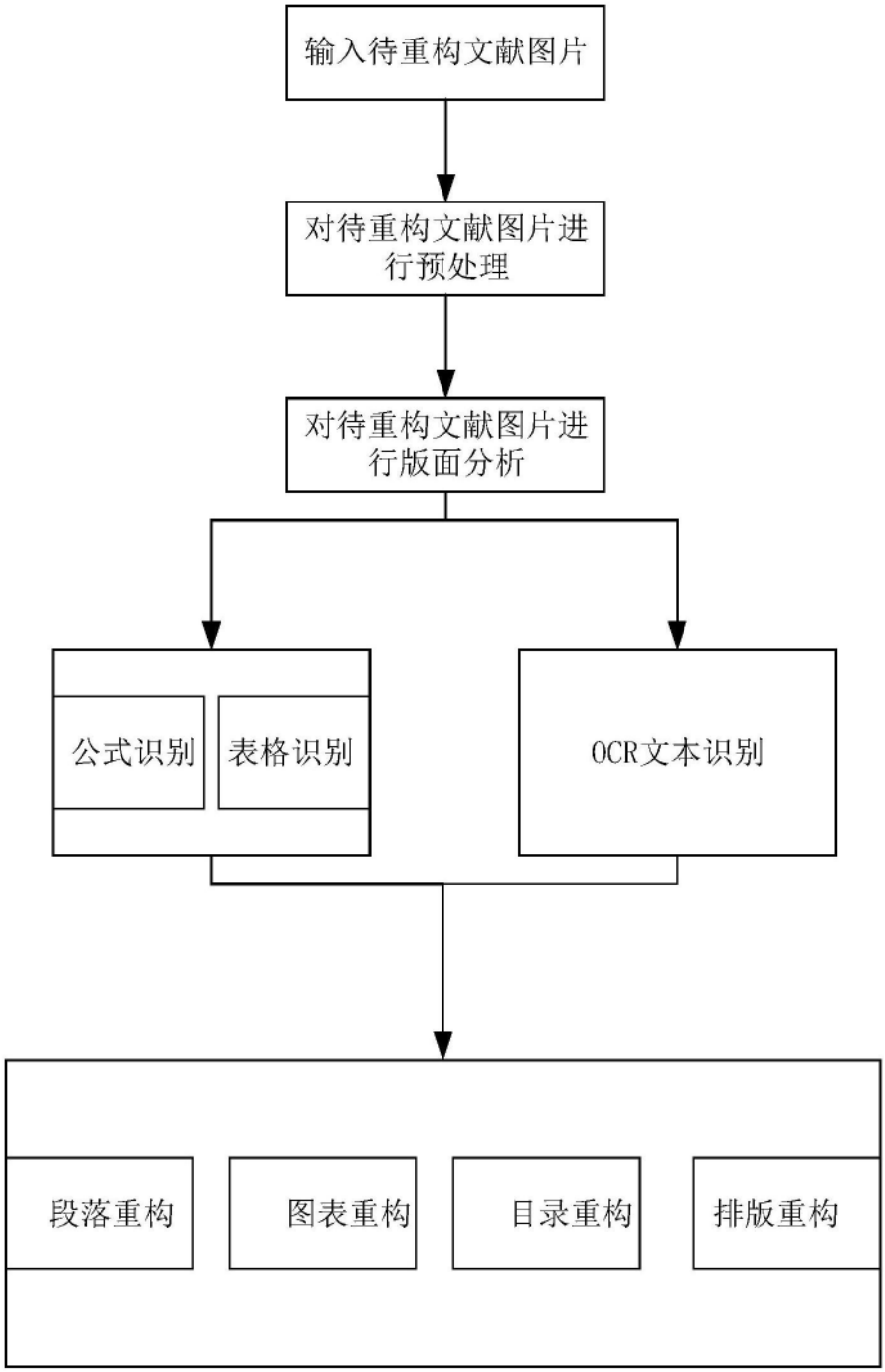

输入待重构文献图片
对待重构文献图片进行预处理
对待重构文献图片进行版面分析
公式识别
表格识别
OCR文本识别
段落重构
图表重构
目录重构
排版重构

图3

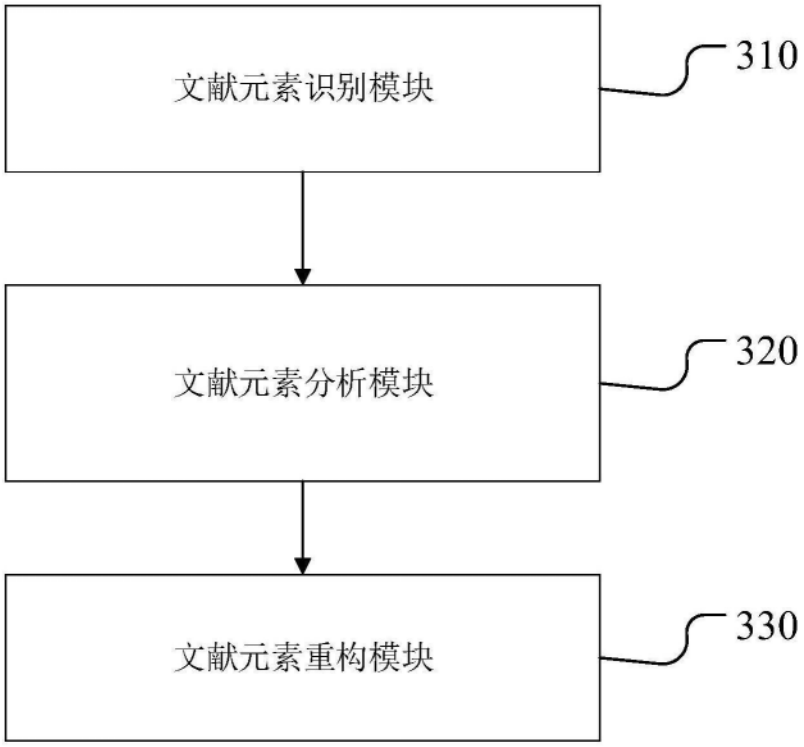


图4

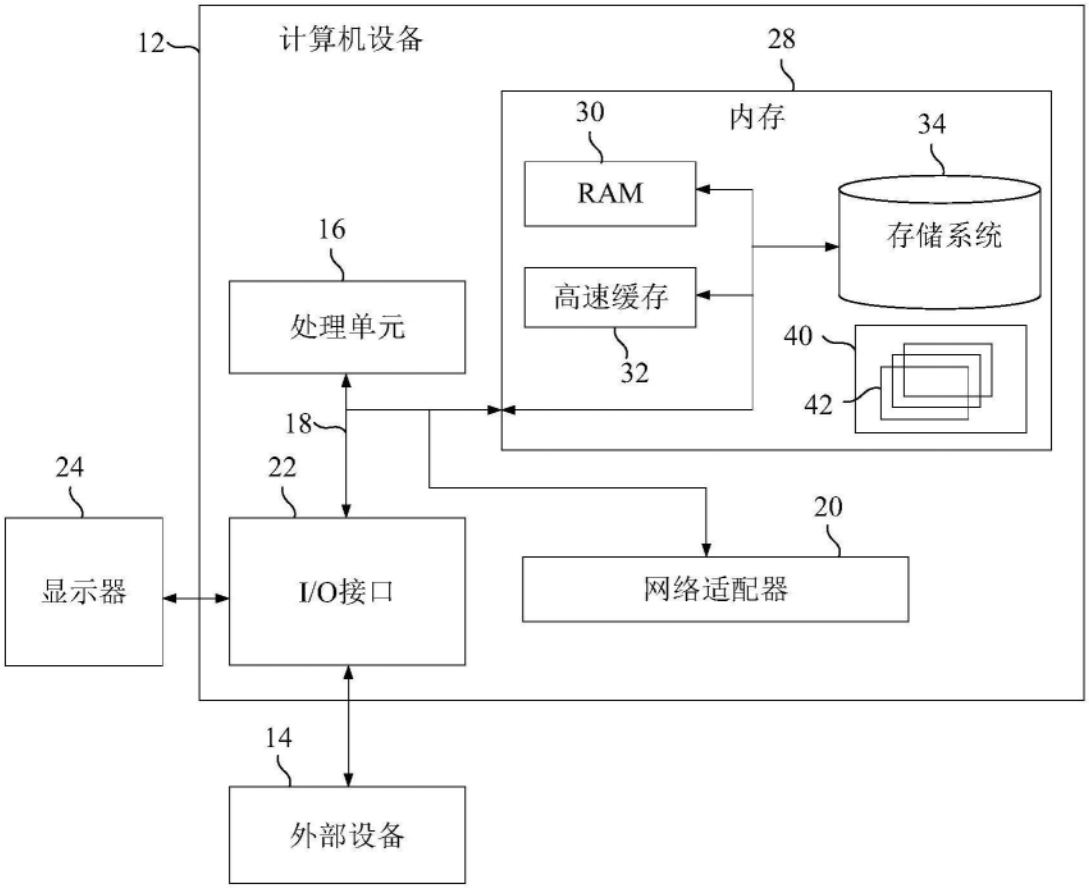

12
计算机设备
28
内存
30
RAM
34
存储系统
16
处理单元
高速缓存
32
40
42
18
24
22
20
显示器
I/O接口
网络适配器
14
外部设备

图5