(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 3 月 7 日 (07.03.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/041526 A1

(51) 国际专利分类号:
G06K 9/00 (2006.01)

(21) 国际申请号: PCT/CN2017/108809

(22) 国际申请日: 2017 年 10 月 31 日 (31.10.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710776352.X 2017年8月31日 (31.08.2017) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。

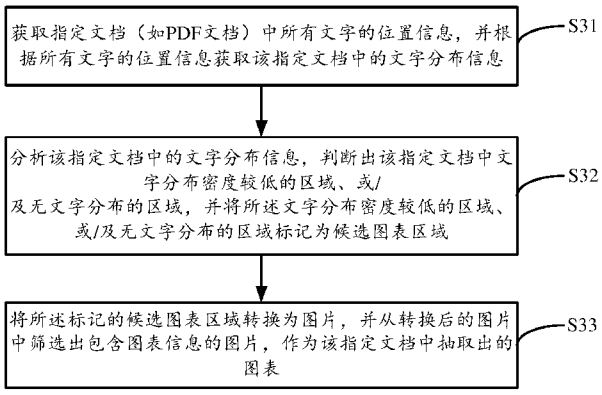
(72) 发明人: 王鸿滨 (WANG, HongBin); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 王晓伟 (WANG, XiaoWei); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 汪伟 (WANG, Wei); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 苏晓明 (SU, XiaoMing); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。 肖京 (XIAO, Jing); 中国广东省深圳福田区八卦岭八卦三路平安大厦吴东勤, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳福田区园岭街道八卦四路 10 号中浩大厦 1528-1530 室于志光, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) **Title:** METHOD OF EXTRACTING CHART IN DOCUMENT, ELECTRONIC DEVICE AND COMPUTER-READABLE STORAGE MEDIUM

(54) **发明名称:** 文档图表抽取方法、电子设备及计算机可读存储介质



S31 Acquiring position information of all texts in a specified document, and acquiring text distribution information in the specified document according to the position information of all the texts

S32 Analyzing the text distribution information in the specified document and determining an area having a lower density of text distribution in the specified document and/or an area without text distribution, and marking the area having the lower density of text distribution and/or the area without text distribution to be candidate chart areas

S33 Converting the marked candidate chart areas into images, and filtering images containing chart information from the converted images to be charts extracted from the specified document

图 3

(57) **Abstract:** A method of extracting a chart in a document. The method comprises: acquiring position information of all texts in a specified document, and acquiring text distribution information in the specified document according to the position information of all the texts (S31); analyzing the text distribution information in the specified document and determining an area having a lower density of text distribution in the specified document and/or an area without text distribution, and marking the area having the lower density of text distribution and/or the area without text distribution to be candidate chart areas (S32); and converting the marked candidate chart areas into images, and filtering images containing chart information from the converted images to be charts extracted from the specified document (S33). The above method can improve the efficiency and coverage of chart extraction.

WO 2019/041526 A1

[见续页]

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种文档图表抽取方法，该方法包括步骤：获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息（S31）；分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域（S32）；将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表（S33）。本方法可以提升图表抽取的效率和覆盖面。

文档图表抽取方法、电子设备及计算机可读存储介质

本专利申请以2017年8月31日提交的申请号为201710776352.X，名称为“文档图表抽取方法、电子设备及计算机可读存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请涉及计算机信息技术领域，尤其涉及一种文档图表抽取方法、电子设备及计算机可读存储介质。

背景技术

现有的PDF图表抽取工具及程序大多是基于PDF存储对象的，这种方式仅能抽取作为单独图片对象存储的图表，而在一个PDF文档中，含有较多的图表信息（如Office图表等），这些图表都能直观地表达出文档中的部分信息。然而，现有的PDF图表抽取工具及程序对于Office图表等由多个部分组成的图表则无法正确抽取。故，现有技术中的文档图表抽取方法设计不够合理，亟需改进。

发明内容

有鉴于此，本申请提出一种文档图表抽取方法、电子设备及计算机可读存储介质，通过文本密度分析从PDF文档中抽取图表，提升了图表抽取的效率和覆盖面。

首先，为实现上述目的，本申请提出一种电子设备，所述电子设备包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的文档图表抽取系统，所述文档图表抽取系统被所述处理器执行时实现如下步骤：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

优选地，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阈值，则判断出该行文字分布密度较低，并清洗该行文字。

优选地，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阈值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

优选地，所述从转换后的图片中筛选出包含图表信息的图片包括：通过

像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

优选地，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

此外，为实现上述目的，本申请还提供一种文档图表抽取方法，该方法应用于电子设备，所述方法包括：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

优选地，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阀值，则判断出该行文字分布密度较低，并清洗该行文字；及

所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阀值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

优选地，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

优选地，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

进一步地，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质存储有文档图表抽取系统，所述文档图表抽取系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述的文档图表抽取方法的步骤。

相较于现有技术，本申请所提出的电子设备、文档图表抽取方法及计算

机可读存储介质，通过文本密度分析从 PDF 文档中抽取图表，该方法除了能提取传统方法能抽取的图表外，还能提取出传统方法无法提取的 Office 图表信息等由多个部分组成的图表，提升了图表抽取的效率和覆盖面。

附图说明

图1是本申请电子设备一可选的硬件架构的示意图；

图2是本申请电子设备中文档图表抽取系统一实施例的程序模块示意图；

图3为本申请文档图表抽取方法一实施例的实施流程示意图。

附图标记：

电子设备	2
存储器	21
处理器	22
网络接口	23
文档图表抽取系统	20
获取模块	201
分析模块	202
抽取模块	203
流程步骤	S31-S33

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

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，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

需要说明的是，在本申请中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛

盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的保护范围之内。

进一步需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

首先，本申请提出一种电子设备 2。

参阅图 1 所示，是本申请电子设备 2 一可选的硬件架构的示意图。本实施例中，所述电子设备 2 可包括，但不限于，可通过系统总线相互通信连接存储器 21、处理器 22、网络接口 23。需要指出的是，图 1 仅示出了具有组件 21-23 的电子设备 2，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

其中，所述电子设备 2 可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该电子设备 2 可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

所述存储器 21 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器 21 可以是所述电子设备 2 的内部存储单元，例如该电子设备 2 的硬盘或内存。在另一些实施例中，所述存储器 21 也可以是所述电子设备 2 的外部存储设备，例如该电子设备 2 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。当然，所述存储器 21 还可以既包括所述电子设备 2 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 21 通常用于存储安装于所述电子设备 2 的操作系统和各类应用软件，例如所述文档图表抽取系统 20 的程序代码等。此外，所述存储器 21 还可以用于暂时地存储已经输出或者将要输出的各类数据。

所述处理器 22 在一些实施例中可以是中央处理器（Central Processing Unit，CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 22 通常用于控制所述电子设备 2 的总体操作，例如执行与所述电子设备 2 进行数据交互或者通信相关的控制和处理等。本实施例中，所述处理器 22 用于运行所述存储器 21 中存储的程序代码或者处理数据，例如运行所述的文档图表抽取系统 20 等。

所述网络接口 23 可包括无线网络接口或有线网络接口，该网络接口 23

通常用于在所述电子设备 2 与其他电子设备之间建立通信连接。例如，所述网络接口 23 用于通过网络将所述电子设备 2 与外部数据平台相连，在所述电子设备 2 与外部数据平台之间的建立数据传输通道和通信连接。所述网络可以是企业内部网（Intranet）、互联网（Internet）、全球移动通讯系统（Global System of Mobile communication，GSM）、宽带码分多址（Wideband Code Division Multiple Access，WCDMA）、4G 网络、5G 网络、蓝牙（Bluetooth）、Wi-Fi 等无线或有线网络。

至此，已经详细介绍了本申请各个实施例的应用环境和相关设备的硬件结构和功能。下面，将基于上述应用环境和相关设备，提出本申请的各个实施例。

参阅图 2 所示，是本申请电子设备 2 中文档图表抽取系统 20 一实施例的程序模块图。本实施例中，所述的文档图表抽取系统 20 可以被分割成一个或多个程序模块，所述一个或者多个程序模块被存储于所述存储器 21 中，并由一个或多个处理器（本实施例中为所述处理器 22）所执行，以完成本申请。例如，在图 2 中，所述的文档图表抽取系统 20 可以被分割成获取模块 201、分析模块 202、以及抽取模块 203。本申请所称的程序模块是指能够完成特定功能的一系列计算机程序指令段，比程序更适合于描述所述文档图表抽取系统 20 在所述电子设备 2 中的执行过程。以下将就各程序模块 201-203 的功能进行详细描述。

所述获取模块 201，用于获取指定文档（如 PDF 文档）中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息。

优选地，在本实施例中，所述文字的位置信息包括，但不限于，文字的横向坐标、纵向坐标、与上一行文字的纵向距离、及与下一行文字的纵向距离等。所述文字分布信息包括，但不限于，每一行文字的左上角坐标，该行文字的长度和宽度等。

所述分析模块 202，用于分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域。

优选地，在本实施例中，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阈值（如 5 个字符单位长度），则判断出该行文字分布密度较低，并清洗（删除）该行文字。清洗后的该行文字变成了一个无文字分布的区域。

优选地，在本实施例中，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阈值宽度（如 2 个字符单位宽度）的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

所述抽取模块203，用于将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。在本实施例中，可以使用特定的图片处理工具（如 ImageMagick 工具）将所述标记的候选图表区域转换为图片。

优选地，在本实施例中，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析（或内容丰富程度分析），对转换后的图片进行筛选，选择出包含图表信息（如 PDF 图表信息）的图片。由于无文字区域有两种情况：一种是图表，一种是页面的空白区域，通过对图片的像素分布分析，可以判断出是这两种情况中的哪一种。

具体而言，通过像素分布分析从转换后的图片中筛选出包含图表信息的图片包括：

（1）对该转换后的图片进行灰度处理（如通过应用程序 Python 中的 Opencv 模块进行灰度处理），将该转换后的图片转换为灰度图。在该灰度图中，图片的每个像素点都被表示为 0 或 255。其中，0 代表黑色，为图片中有信息内容的像素点，255 代表白色，为图片中空白的像素点。

（2）按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值（如数量超过 5，比例超过 50%），则判定该行包含有具体内容。

（3）统计包含有具体内容的行的数量，以此来判定图片中内容的丰富程度，包含有具体内容的行越多，则代表该图片的内容越丰富。若包含有具体内容的行数大于或等于设定阈值（如 2 行），则判定该转换后的图片内容丰富，是一张包含图表信息的图片。反之，若包含有具体内容的行数小于该设定阈值（如 2 行），则判定该转换后的图片内容不够丰富，是一张没有包含图表信息的空白图片。

通过上述程序模块 201-203，本申请所提出的文档图表抽取系统 20，通过文本密度分析从 PDF 文档中抽取图表，该方法除了能提取传统方法能抽取的图表外，还能提取出传统方法无法提取的 Office 图表信息等由多个部分组成的图表，提升了图表抽取的效率和覆盖面。

此外，本申请还提出一种文档图表抽取方法。

参阅图 3 所示，是本申请文档图表抽取方法一实施例的实施流程示意图。在本实施例中，根据不同的需求，图 3 所示的流程图中的步骤的执行顺序可以改变，某些步骤可以省略。

步骤S31，获取指定文档（如PDF文档）中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息。

优选地，在本实施例中，所述文字的位置信息包括，但不限于，文字的

横向坐标、纵向坐标、与上一行文字的纵向距离、及与下一行文字的纵向距离等。所述文字分布信息包括，但不限于，每一行文字的左上角坐标，该行文字的长度和宽度等。

步骤S32，分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域。

优选地，在本实施例中，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阀值（如5个字符单位长度），则判断出该行文字分布密度较低，并清洗（删除）该行文字。清洗后的该行文字变成了一个无文字分布的区域。

优选地，在本实施例中，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阀值宽度（如2个字符单位宽度）的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

步骤S33，将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。在本实施例中，可以使用特定的图片处理工具（如ImageMagick工具）将所述标记的候选图表区域转换为图片。

优选地，在本实施例中，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析（或内容丰富程度分析），对转换后的图片进行筛选，选择出包含图表信息（如PDF图表信息）的图片。由于无文字区域有两种情况：一种是图表，一种是页面的空白区域，通过对图片的像素分布分析，可以判断出是这两种情况中的哪一种。

具体而言，通过像素分布分析从转换后的图片中筛选出包含图表信息的图片包括：

（1）对该转换后的图片进行灰度处理（如通过应用程序Python中的Opencv模块进行灰度处理），将该转换后的图片转换为灰度图。在该灰度图中，图片的每个像素点都被表示为0或255。其中，0代表黑色，为图片中有信息内容的像素点，255代表白色，为图片中空白的像素点。

（2）按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值（如数量超过5，比例超过50%），则判定该行包含有具体内容。

（3）统计包含有具体内容的行的数量，以此来判定图片中内容的丰富程度，包含有具体内容的行越多，则代表该图片的内容越丰富。若包含有具体内容的行数大于或等于设定阈值（如2行），则判定该转换后的图片内容丰富，

是一张包含图表信息的图片。反之，若包含有具体内容的行数小于该设定阈值（如2行），则判定该转换后的图片内容不够丰富，是一张没有包含图表信息的空白图片。

通过上述步骤S31-S33，本申请所提出的文档图表抽取方法，通过文本密度分析从PDF文档中抽取图表，该方法除了能提取传统方法能抽取的图表外，还能提取出传统方法无法提取的Office图表信息等由多个部分组成的图表，提升了图表抽取的效率和覆盖面。

进一步地，为实现上述目的，本申请还提供一种计算机可读存储介质（如ROM/RAM、磁碟、光盘），所述计算机可读存储介质存储有文档图表抽取系统20，所述文档图表抽取系统20可被至少一个处理器22执行，以使所述至少一个处理器22执行如上所述的文档图表抽取方法的步骤。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件来实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

以上参照附图说明了本申请的优选实施例，并非因此局限本申请的权利范围。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。另外，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

本领域技术人员不脱离本申请的范围和实质，可以有多种变型方案实现本申请，比如作为一个实施例的特征可用于另一实施例而得到又一实施例。凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1. 一种电子设备，其特征在于，所述电子设备包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的文档图表抽取系统，所述文档图表抽取系统被所述处理器执行时实现如下步骤：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

2. 如权利要求 1 所述的电子设备，其特征在于，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阈值，则判断出该行文字分布密度较低，并清洗该行文字。

3. 如权利要求 2 所述的电子设备，其特征在于，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阈值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

4. 如权利要求 3 所述的电子设备，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

5. 如权利要求 4 所述的电子设备，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

6. 一种文档图表抽取方法，应用于电子设备，其特征在于，所述方法包括：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度

较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

7. 如权利要求 6 所述的文档图表抽取方法，其特征在于，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阀值，则判断出该行文字分布密度较低，并清洗该行文字。

8. 如权利要求 7 所述的文档图表抽取方法，其特征在于，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阀值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

9. 如权利要求 8 所述的文档图表抽取方法，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

10. 如权利要求 9 所述的文档图表抽取方法，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

11. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有文档图表抽取系统，所述文档图表抽取系统可被至少一个处理器执行，以使所述至少一个处理器执行如下步骤：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

12. 如权利要求 11 所述的计算机可读存储介质，其特征在于，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一

阀值，则判断出该行文字分布密度较低，并清洗该行文字。

13. 如权利要求 12 所述的计算机可读存储介质，其特征在于，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阀值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

14. 如权利要求 13 所述的计算机可读存储介质，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：通过像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

15. 如权利要求 14 所述的计算机可读存储介质，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

16. 一种文档图表抽取系统，其特征在于，所述文档图表抽取系统可被至少一个处理器执行，以使所述至少一个处理器执行如下步骤：

获取指定文档中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息；

分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域；及

将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表。

17. 如权利要求 16 所述的文档图表抽取系统，其特征在于，所述判断出该指定文档中文字分布密度较低的区域包括：若一行文字的长度小于第一阀值，则判断出该行文字分布密度较低，并清洗该行文字。

18. 如权利要求 17 所述的文档图表抽取系统，其特征在于，所述判断出该指定文档中无文字分布的区域包括：对该指定文档中每一页从上到下进行扫描，若超过第二阀值宽度的区域没有扫描到文字，则判断出该区域为无文字分布的区域。

19. 如权利要求 18 所述的文档图表抽取系统，其特征在于，所述从转换

后的图片中筛选出包含图表信息的图片包括：通过像素分布分析，对转换后的图片进行筛选，选择出包含图表信息的图片。

20. 如权利要求 19 所述的文档图表抽取系统，其特征在于，所述从转换后的图片中筛选出包含图表信息的图片包括：

对该转换后的图片进行灰度处理，将该转换后的图片转换为灰度图；

按行统计该灰度图中黑色像素点的数量和比例，若一行中黑色像素点的数量和比例超过指定阈值，则判定该行包含有具体内容；及

统计包含有具体内容的行的数量，若包含有具体内容的行数大于或等于设定阈值，则判定该转换后的图片为一张包含图表信息的图片。

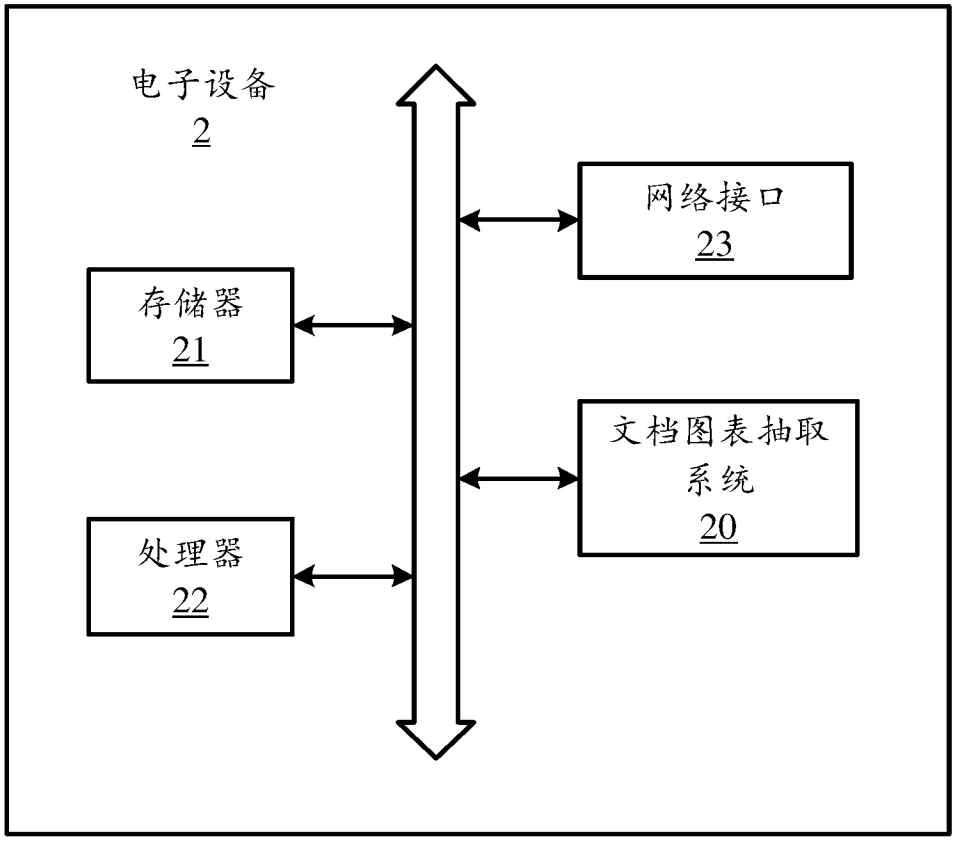


图 1

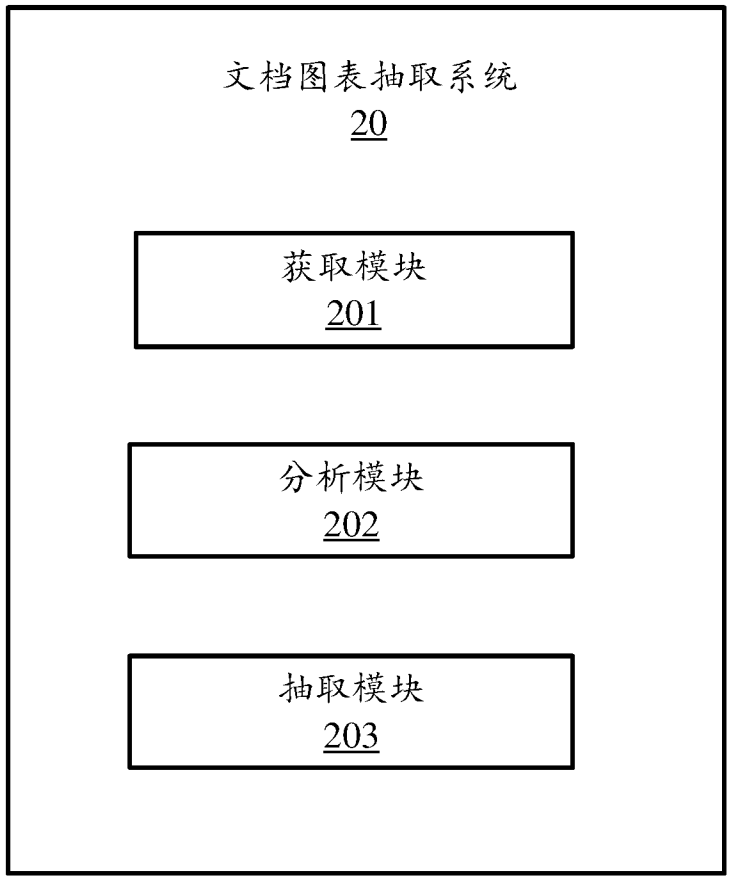


图 2

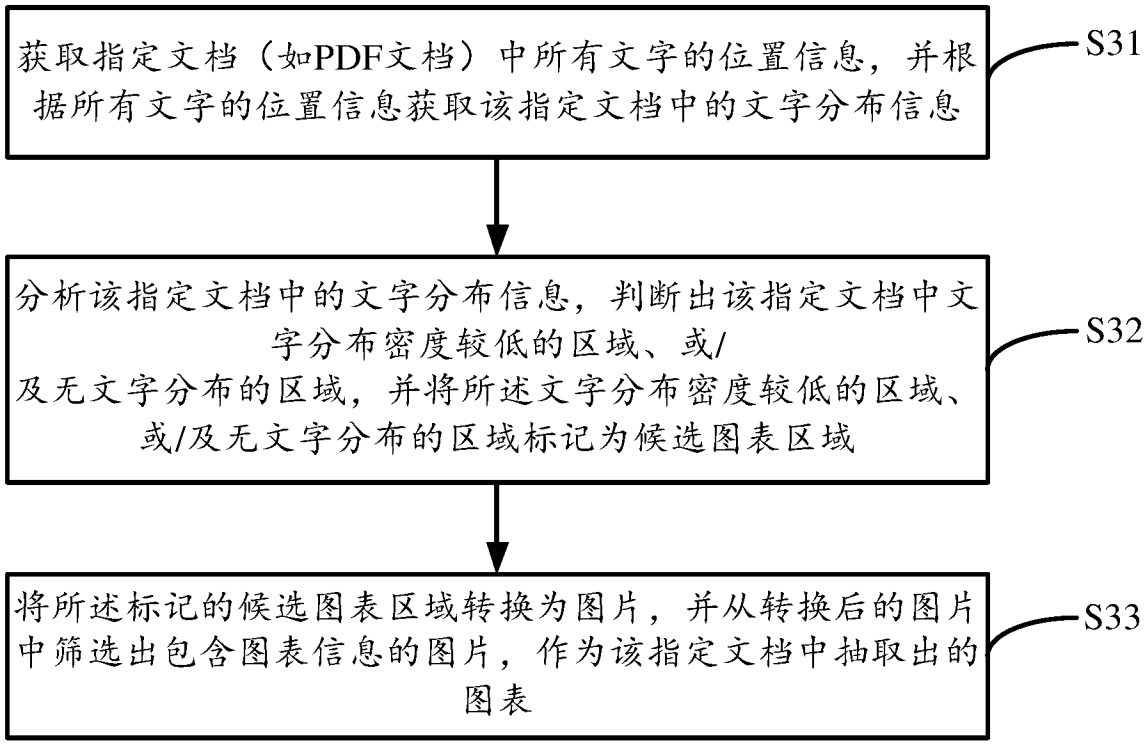

获取指定文档（如PDF文档）中所有文字的位置信息，并根据所有文字的位置信息获取该指定文档中的文字分布信息
S31
分析该指定文档中的文字分布信息，判断出该指定文档中文字分布密度较低的区域、或/及无文字分布的区域，并将所述文字分布密度较低的区域、或/及无文字分布的区域标记为候选图表区域
S32
将所述标记的候选图表区域转换为图片，并从转换后的图片中筛选出包含图表信息的图片，作为该指定文档中抽取出的图表
S33

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/108809

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, WPI, EPODOC: 文档, 图像, 图片, 图表, 文字, 文本, 字符, 分布, 位置, 密度, 无文字, 非文字, 非文本, 少, 区域, 密度, block?, area?, document?, file?, image?, character?, word?, text, chart?, draw+, line?, density

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 107133566 A (CHANG, Cheng) 05 September 2017 (05.09.2017), description, paragraphs [0016]-[0032], and figure 2	1-20
X	CN 101008960 A (RICOH CO., LTD.) 01 August 2007 (01.08.2007), description, page 6, paragraph 5 to page 8, the last paragraph	1, 6, 11, 16
A	CN 1967567 A (SAMSUNG ELECTRONICS CO., LTD.) 23 May 2007 (23.05.2007), entire document	1-20
A	CN 101004792 A (RICOH CO., LTD.) 25 July 2007 (25.07.2007), entire document	1-20
A	US 5680479 A (CANON KABUSHIKI KAISHA) 21 October 1997 (21.10.1997), entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 May 2018	30 May 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WANG, Wei Telephone No. (86-10) 53961525

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/108809

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 107133566 A	05 September 2017	None	
CN 101008960 A	01 August 2007	JP 2007200014 A	09 August 2007
		US 2007171473 A1	26 July 2007
CN 1967567 A	23 May 2007	KR 100664311 B1	04 January 2007
		US 8369623 B2	05 February 2013
		US 2007116359 A1	24 May 2007
		US 7860316 B2	28 December 2010
		US 2011064310 A1	17 March 2011
CN 101004792 A	25 July 2007	US 2007165950 A1	19 July 2007
		JP 2007193528 A	02 August 2007
		JP 4768451 B2	07 September 2011
US 5680479 A	21 October 1997	US 6115497 A	05 September 2000
		JP H0668301 A	11 March 1994
		DE 69332459 D1	12 December 2002
		JP 3359095 B2	24 December 2002
		EP 0567344 A2	27 October 1993
		US 5680478 A	21 October 1997
		US 6081616 A	27 June 2000
		US 2001001435 A1	24 May 2001
		US 6475350 B2	05 November 2002

国际检索报告

国际申请号
PCT/CN2017/108809

A. 主题的分类

G06K 9/00(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNKI, CNPAT, WPI, EPODOC: 文档，图像，图片，图表，文字，文本，字符，分布，位置，密度，无文字，非文字，非文本，少，区域，密度，block?, area?, document?, file?, image?, character?, word?, text, chart?, draw+, line?, density

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 107133566 A (常诚) 2017年 9月 5日 (2017-09-05) 说明书第[0016]-[0032]段，附图2	1-20
X	CN 101008960 A (株式会社理光) 2007年 8月 1日 (2007-08-01) 说明书第6页第5段-第8页最后一段	1，6，11，16
A	CN 1967567 A (三星电子株式会社) 2007年 5月 23日 (2007-05-23) 全文	1-20
A	CN 101004792 A (株式会社理光) 2007年 7月 25日 (2007-07-25) 全文	1-20
A	US 5680479 A (CANON KABUSHIKI KAISHA) 1997年 10月 21日 (1997-10-21) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2018年 5月 18日	2018年 5月 30日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	王伟
传真号 (86-10)62019451	电话号码 86-(10)-53961525

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/108809

检索报告引用的专利文件			公布日（年/月/日）	同族专利			公布日（年/月/日）
CN	107133566	A	2017年 9月 5日	无			
CN	101008960	A	2007年 8月 1日	JP	2007200014	A	2007年 8月 9日
				US	2007171473	A1	2007年 7月 26日
CN	1967567	A	2007年 5月 23日	KR	100664311	B1	2007年 1月 4日
				US	8369623	B2	2013年 2月 5日
				US	2007116359	A1	2007年 5月 24日
				US	7860316	B2	2010年 12月 28日
				US	2011064310	A1	2011年 3月 17日
CN	101004792	A	2007年 7月 25日	US	2007165950	A1	2007年 7月 19日
				JP	2007193528	A	2007年 8月 2日
				JP	4768451	B2	2011年 9月 7日
US	5680479	A	1997年 10月 21日	US	6115497	A	2000年 9月 5日
				JP	H0668301	A	1994年 3月 11日
				DE	69332459	D1	2002年 12月 12日
				JP	3359095	B2	2002年 12月 24日
				EP	0567344	A2	1993年 10月 27日
				US	5680478	A	1997年 10月 21日
				US	6081616	A	2000年 6月 27日
				US	2001001435	A1	2001年 5月 24日
				US	6475350	B2	2002年 11月 5日