(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104346615 A
(43)申请公布日 2015.02.11

(21)申请号 201310343908.8

(22)申请日 2013.08.08

(71)申请人 北大方正集团有限公司
地址 100871 北京市海淀区成府路 298 号方正大厦 9 层
申请人 北京方正阿帕比技术有限公司
北京大学

(72)发明人 许灿辉 汤帜 陶欣 史操

(74)专利代理机构 北京友联知识产权代理事务所（普通合伙） 11343
代理人 尚志峰 汪海屏

(51)Int.Cl.
G06K 9/46(2006.01)

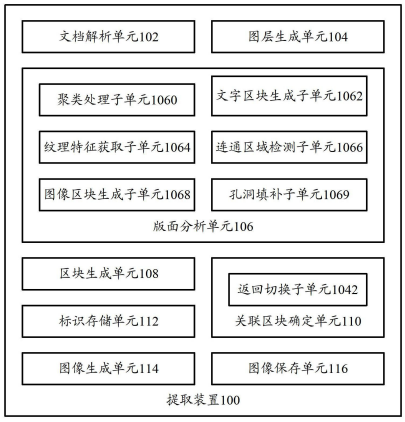
权利要求书2页 说明书10页 附图8页

(54)发明名称

版式文档中复合图的提取装置和提取方法

(57)摘要

本发明提供了一种版式文档中复合图的提取装置，包括：文档解析单元，对版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；图层生成单元，提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；版面分析单元，分别对文字图层和非文字图层进行版面分析处理；区块生成单元，生成文字图层中的文字区块和非文字图层中的图像区块；关联区块确定单元，确定与每个图像区块相关联的文字区块，以合并为复合图区块；标识存储单元，存储复合图区块包含的所有图元的标识。本发明还提出了一种版式文档中复合图的提取方法。通过本发明的技术方案，可以在复杂的文档版面布局中，尤其是图文混排版面中，实现对复合图的准确提取。



CN 104346615 A

1. 一种版式文档中复合图的提取装置，其特征在于，包括：

文档解析单元，用于对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；

图层生成单元，用于提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；

版面分析单元，用于分别对所述文字图层和所述非文字图层进行版面分析处理；

区块生成单元，用于根据所述版面分析单元的版面分析处理结果，生成所述文字图层中的文字区块和所述非文字图层中的图像区块；

关联区块确定单元，用于确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；

标识存储单元，用于存储所述复合图区块包含的所有图元的标识。

2. 根据权利要求 1 所述的版式文档中复合图的提取装置，其特征在于，所述版面分析单元包括：

聚类处理子单元，用于对所述文字图层中的文字图元进行聚类处理，以对所述文字图元进行分类；

文字区块生成子单元，对于同类别的多个文字图元，在对应的最小外接矩形相交或间隔距离小于预设距离的情况下，用于将所述多个文字图元作为文字图元集合，并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

3. 根据权利要求 1 所述的版式文档中复合图的提取装置，其特征在于，所述版面分析单元包括：

纹理特征获取子单元，用于获取所述非文字图层中的非文字图元的纹理特征；

连通区域检测子单元，用于根据所述纹理特征以及预设的特征阈值，检测出所述非文字图层中连通的非文字对象区域；

图像区块生成子单元，对于多个所述连通的非文字对象区域，在对应的最小外接矩形相交或间隔距离小于预设距离的情况下，用于将多个所述连通的非文字对象区域作为区域集合，并将所述区域集合的最小外接矩形作为所述图像区块。

4. 根据权利要求 3 所述的版式文档中复合图的提取装置，其特征在于，所述版面分析单元还包括：

孔洞填补子单元，用于对所述连通的非文字对象区域中存在的孔洞进行填补。

5. 根据权利要求 1 所述的版式文档中复合图的提取装置，其特征在于，所述关联区块确定单元包括：

位置关系检测子单元，用于检测所述图像区块与所述文字区块之间的位置关系，其中，若指定图像区块与至少一个文字区块相交，或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离，则判定所述至少一个文字区块与所述指定图像区块相关联。

6. 根据权利要求 1 至 5 中任一项所述的版式文档中复合图的提取装置，其特征在于，还包括：

图像生成单元，用于将所述复合图区块生成为图像文件；

图像保存单元，用于保存所述图像文件。

7. 一种版式文档中复合图的提取方法，其特征在于，包括：

对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；

提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；

分别对所述文字图层和所述非文字图层进行版面分析处理，以生成所述文字图层中的文字区块和所述非文字图层中的图像区块；

确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；

存储所述复合图区块包含的所有图元的标识。

8. 根据权利要求 7 所述的版式文档中复合图的提取方法，其特征在于，对所述文字图层进行版面分析处理的步骤包括：

对所述文字图层中的文字图元进行聚类处理，以对所述文字图元进行分类，其中，

对于同类别的多个文字图元，若对应的最小外接矩形相交或间隔距离小于预设距离，则将所述多个文字图元作为文字图元集合，并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

9. 根据权利要求 7 所述的版式文档中复合图的提取方法，其特征在于，对所述非文字图层进行版面分析处理的步骤包括：

获取所述非文字图层中的非文字图元的纹理特征，并根据预设的特征阈值，检测出所述非文字图层中连通的非文字对象区域，其中，

对于多个所述连通的非文字对象区域，若对应的最小外接矩形相交或间隔距离小于预设距离，则将多个所述连通的非文字对象区域作为区域集合，并将所述区域集合的最小外接矩形作为所述图像区块。

10. 根据权利要求 9 所述的版式文档中复合图的提取方法，其特征在于，还包括：

对所述连通的非文字对象区域中存在的孔洞进行填补。

11. 根据权利要求 7 所述的版式文档中复合图的提取方法，其特征在于，所述确定与每个所述图像区块相关联的文字区块的步骤包括：

检测所述图像区块与所述文字区块之间的位置关系，若指定图像区块与至少一个文字区块相交，或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离，则判定所述至少一个文字区块与所述指定图像区块相关联。

12. 根据权利要求 7 至 11 中任一项所述的版式文档中复合图的提取方法，其特征在于，还包括：

将所述复合图区块保存为图像文件。

版式文档中复合图的提取装置和提取方法

技术领域

[0001] 本发明涉及电子文档格式转换技术领域，具体而言，涉及一种版式文档中复合图的提取装置和一种版式文档中复合图的提取方法。

背景技术

[0002] 将纸张文档转换为电子文档，大多采用扫描仪扫描或者相机拍摄的方式，获取文档的数字图像，对其进行一系列图像处理后，将字符切分出来，输入 OCR（Optical Character Recognition，光学字符识别）系统。而由文档处理软件，如排版软件，直接生成的版式文档，正在取代从纸质文档转化而来的图像文档成为数字出版物的主要文档来源。

[0003] 结构信息的自动提取，主要包括版面分析和版面理解，其研究皆停留在图像文档版面的物理结构的提取，而针对通过 OCR 转化或者直接生成的版式文档的研究才刚刚起步。文档版面布局的复杂性和多样性导致插图的准确分割成为公开性难题，尤其是文字环绕型的插图。另外，版式文档中，复合图都由多个子图像、大量路径操作、文字图元等子对象构成，不能在逆向工程的版面结构分析中作为复合图的完整体被正确的提取出来。因而版式文档不仅在描述上要大量路径来描述，造成很大程度的冗余，更不利于版式文档流式重排时复合图的正常显示，难以满足日益增长数字化阅读的现实需求。

[0004] 因此，需要一种新的版式文档中复合图的提取技术，可以在复杂的文档版面布局中，尤其是图文混排版面中，实现对复合图的准确提取。

发明内容

[0005] 本发明正是基于上述问题，提出了一种新的版式文档中复合图的提取技术，可以在复杂的文档版面布局中，尤其是图文混排版面中，实现对复合图的准确提取。

[0006] 有鉴于此，本发明提出了一种版式文档中复合图的提取装置，包括：文档解析单元，用于对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；图层生成单元，用于提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；版面分析单元，用于分别对所述文字图层和所述非文字图层进行版面分析处理；区块生成单元，用于根据所述版面分析单元的版面分析处理结果，生成所述文字图层中的文字区块和所述非文字图层中的图像区块；关联区块确定单元，用于确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；标识存储单元，用于存储所述复合图区块包含的所有图元的标识。

[0007] 在该技术方案中，通过对版式文档进行解析后，将得到的图元分别构成文字图层(包含文字图元)和非文字图层(包含图像图元等)，然后分别对每个图层进行区块分类，最终利用区块之间的关系判定复合图区块，以实现对复合图区块的分割，并确保对文字图元和非文字图元的妥善处理。在生成多个图层时，具体地，可以先提取所有的文字图元以形成文字图层，然后将文字图元过滤以利用剩余的元素构成非文字图元。本方案可以对图文混排、包含图像和图注信息等复杂情况进行有效地分析，从而准确地分割出其中的复合图区块。

复合图区块中可以包含一个或多个复合图，还可以包含复合图中或周围的图注等文字。通过记录所有构成该复合图区块的图元的标识，如图元 ID，从而能够利用这些图元 ID 来对应出该复合图区块，实现了将该区块与整个版面的分离，方便进行流式重排等处理。

[0008] 在上述技术方案中，优选地，所述版面分析单元包括：聚类处理子单元，用于对所述文字图层中的文字图元进行聚类处理，以对所述文字图元进行分类；文字区块生成子单元，对于同类别的多个文字图元，在对应的最小外接矩形相交或间隔距离小于预设距离的情况下，用于将所述多个文字图元作为文字图元集合，并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

[0009] 在该技术方案中，通过基于页面内文字图元邻域特征相似性的聚类算法处理，可以有效地对文字图元进行分类，从而确定每个文字图元应该属于正文部分还是复合图部分。通过对距离的判断及相应的处理，从而确定多个文字图元的构成关系，比如用于构成一个文字区块，该文字区块对应于一个完整的字符。

[0010] 在上述技术方案中，优选地，所述版面分析单元包括：纹理特征获取子单元，用于获取所述非文字图层中的非文字图元的纹理特征；连通区域检测子单元，用于根据所述纹理特征以及预设的特征阈值，检测出所述非文字图层中连通的非文字对象区域；图像区块生成子单元，对于多个所述连通的非文字对象区域，在对应的最小外接矩形相交或间隔距离小于预设距离的情况下，用于将多个所述连通的非文字对象区域作为区域集合，并将所述区域集合的最小外接矩形作为所述图像区块。

[0011] 在该技术方案中，利用基于纹理分析和形态学处理的页面非文字对象的连通域检测，从而识别出版面中的连通的非文字对象区域，该区域实际上对应于版面中的一幅图像或该图像中的一部分；再通过对距离的判断及相应的处理，即可将构成同一幅图像的多个连通区域进行合并，从而实现对某一幅图像的完整的识别。

[0012] 在上述技术方案中，优选地，所述版面分析单元还包括：孔洞填补子单元，用于对所述连通的非文字对象区域中存在的孔洞进行填补。

[0013] 在该技术方案中，通过对连通的非文字对象区域中存在的孔洞进行填补，从而能够以整体为对象来处理对应的区域，避免了孔洞为处理过程带来的难度和可能造成的意外。

[0014] 在上述技术方案中，优选地，所述关联区块确定单元包括：位置关系检测子单元，用于检测所述图像区块与所述文字区块之间的位置关系，其中，若指定图像区块与至少一个文字区块相交，或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离，则判定所述至少一个文字区块与所述指定图像区块相关联。

[0015] 在该技术方案中，由于图像往往存在一些文字描述，比如图标题、图中的标注文字等等，这些文字与图像之间是相关联的，应该划分至相同的区块。通过上述处理，使得分割出来的复合图区块更加准确。

[0016] 在上述技术方案中，优选地，还包括：图像生成单元，用于将所述复合图区块生成为图像文件；图像保存单元，用于保存所述图像文件。

[0017] 在该技术方案中，直接将分割出来的复合图区块以图像文件的形式进行保存，从而不必对图元 ID 进行管理，尤其是当这些复合图区块中包含有数量很多的图元时，以图像文件进行处理的方式，显然有利于提升处理效率。

[0018] 根据本发明的又一方面，还提出了一种版式文档中复合图的提取方法，包括：步骤 202，对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；步骤 204，提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；步骤 206，分别对所述文字图层和所述非文字图层进行版面分析处理，以生成所述文字图层中的文字区块和所述非文字图层中的图像区块；步骤 208，确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；步骤 210，存储所述复合图区块包含的所有图元的标识。

[0019] 在该技术方案中，通过对版式文档进行解析后，将得到的图元分别构成文字图层(包含文字图元)和非文字图层(包含图像图元等)，然后分别对每个图层进行区块分类，最终利用区块之间的关系判定复合图区块，以实现对复合图区块的分割，并确保对文字图元和非文字图元的妥善处理。在生成多个图层时，具体地，可以先提取所有的文字图元以形成文字图层，然后将文字图元过滤以利用剩余的元素构成非文字图元。本方案可以对图文混排、包含图像和图注信息等复杂情况进行有效地分析，从而准确地分割出其中的复合图区块。复合图区块中可以包含一个或多个复合图，还可以包含复合图中或周围的图注等文字。通过记录所有构成该复合图区块的图元的标识，如图元 ID，从而能够利用这些图元 ID 来对应出该复合图区块，实现了将该区块与整个版面的分离，方便进行流式重排等处理。

[0020] 在上述技术方案中，优选地，对所述文字图层进行版面分析处理的步骤包括：对所述文字图层中的文字图元进行聚类处理，以对所述文字图元进行分类，其中，对于同类别的多个文字图元，若对应的最小外接矩形相交或间隔距离小于预设距离，则将所述多个文字图元作为文字图元集合，并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

[0021] 在该技术方案中，通过基于页面内文字图元邻域特征相似性的聚类算法处理，可以有效地对文字图元进行分类，从而确定每个文字图元应该属于正文部分还是复合图部分。通过对距离的判断及相应的处理，从而确定多个文字图元的构成关系，比如用于构成一个文字区块，该文字区块对应于一个完整的字符。

[0022] 在上述技术方案中，优选地，对所述非文字图层进行版面分析处理的步骤包括：获取所述非文字图层中的非文字图元的纹理特征，并根据预设的特征阈值，检测出所述非文字图层中连通的非文字对象区域，其中，对于多个所述连通的非文字对象区域，若对应的最小外接矩形相交或间隔距离小于预设距离，则将多个所述连通的非文字对象区域作为区域集合，并将所述区域集合的最小外接矩形作为所述图像区块。

[0023] 在该技术方案中，利用基于纹理分析和形态学处理的页面非文字对象的连通域检测，从而识别出版面中的连通的非文字对象区域，该区域实际上对应于版面中的一幅图像或该图像中的一部分；再通过对距离的判断及相应的处理，即可将构成同一幅图像的多个连通区域进行合并，从而实现对某一幅图像的完整的识别。

[0024] 在上述技术方案中，优选地，还包括：对所述连通的非文字对象区域中存在的孔洞进行填补。

[0025] 在该技术方案中，通过对连通的非文字对象区域中存在的孔洞进行填补，从而能够以整体为对象来处理对应的区域，避免了孔洞为处理过程带来的难度和可能造成的意外。

[0026] 在上述技术方案中，优选地，所述确定与每个所述图像区块相关联的文字区块的步骤包括：检测所述图像区块与所述文字区块之间的位置关系，若指定图像区块与至少一

个文字区块相交，或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离，则判定所述至少一个文字区块与所述指定图像区块相关联。

[0027] 在该技术方案中，由于图像往往存在一些文字描述，比如图标题、图中的标注文字等等，这些文字与图像之间是相关联的，应该划分至相同的区块。通过上述处理，使得分割出来的复合图区块更加准确。

[0028] 在上述技术方案中，优选地，还包括：将所述复合图区块保存为图像文件。

[0029] 在该技术方案中，直接将分割出来的复合图区块以图像文件的形式进行保存，从而不必对图元 ID 进行管理，尤其是当这些复合图区块中包含有数量很多的图元时，以图像文件进行处理的方式，显然有利于提升处理效率。

[0030] 通过以上技术方案，可以在复杂的文档版面布局中，尤其是图文混排版面中，实现对复合图的准确提取。

附图说明

[0031] 图 1 示出了根据本发明的实施例的版式文档中复合图的提取装置的框图；

[0032] 图 2 示出了根据本发明的实施例的版式文档中复合图的提取方法的流程图；

[0033] 图 3 示出了根据本发明的实施例的对版式文档中的复合图进行提取的具体流程图；

[0034] 图 4A 至图 4D 示出了根据本发明的一个实施例的对版式文档中的复合图进行提取的示意图；

[0035] 图 5A 至图 5D 示出了根据本发明的另一个实施例的对版式文档中的复合图进行提取的示意图。

具体实施方式

[0036] 为了能够更清楚地理解本发明的上述目的、特征和优点，下面结合附图和具体实施方式对本发明进行进一步的详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

[0037] 在下面的描述中阐述了很多具体细节以便于充分理解本发明，但是，本发明还可以采用其他不同于在此描述的其他方式来实施，因此，本发明并不限于下面公开的具体实施例的限制。

[0038] 图 1 示出了根据本发明的实施例的版式文档中复合图的提取装置的框图。

[0039] 如图 1 所示，根据本发明的实施例的版式文档中复合图的提取装置 100，包括：文档解析单元 102，用于对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；图层生成单元 104，用于提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；版面分析单元 106，用于分别对所述文字图层和所述非文字图层进行版面分析处理；区块生成单元 108，用于根据所述版面分析单元 106 的版面分析处理结果，生成所述文字图层中的文字区块和所述非文字图层中的图像区块；关联区块确定单元 110，用于确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；标识存储单元 112，用于存储所述复合图区块包含的所有图元的标识。

[0040] 在该技术方案中，通过对版式文档进行解析后，将得到的图元分别构成文字图层

(包含文字图元)和非文字图层(包含图像图元等),然后分别对每个图层进行区块分类,最终利用区块之间的关系判定复合图区块,以实现对复合图区块的分割,并确保对文字图元和非文字图元的妥善处理。在生成多个图层时,具体地,可以先提取所有的文字图元以形成文字图层,然后将文字图元过滤以利用剩余的元素构成非文字图元。本方案可以对图文混排、包含图像和图注信息等复杂情况进行有效地分析,从而准确地分割出其中的复合图区块。复合图区块中可以包含一个或多个复合图,还可以包含复合图中或周围的图注等文字。通过记录所有构成该复合图区块的图元的标识,如图元 ID,从而能够利用这些图元 ID 来对应出该复合图区块,实现了将该区块与整个版面的分离,方便进行流式重排等处理。

[0041] 在上述技术方案中,优选地,所述版面分析单元 106 包括 :聚类处理子单元 1060,用于对所述文字图层中的文字图元进行聚类处理,以对所述文字图元进行分类 ;文字区块生成子单元 1062,对于同类别的多个文字图元,在对应的最小外接矩形相交或间隔距离小于预设距离的情况下,用于将所述多个文字图元作为文字图元集合,并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

[0042] 在该技术方案中,通过基于页面内文字图元邻域特征相似性的聚类算法处理,可以有效地对文字图元进行分类,从而确定每个文字图元应该属于正文部分还是复合图部分。通过对距离的判断及相应的处理,从而确定多个文字图元的构成关系,比如用于构成一个文字区块,该文字区块对应于一个完整的字符。

[0043] 在上述技术方案中,优选地,所述版面分析单元 106 包括 :纹理特征获取子单元 1064,用于获取所述非文字图层中的非文字图元的纹理特征 ;连通区域检测子单元 1066,用于根据所述纹理特征以及预设的特征阈值,检测出所述非文字图层中连通的非文字对象区域 ;图像区块生成子单元 1068,对于多个所述连通的非文字对象区域,在对应的最小外接矩形相交或间隔距离小于预设距离的情况下,用于将多个所述连通的非文字对象区域作为区域集合,并将所述区域集合的最小外接矩形作为所述图像区块。

[0044] 在该技术方案中,利用基于纹理分析和形态学处理的页面非文字对象的连通域检测,从而识别出版面中的连通的非文字对象区域,该区域实际上对应于版面中的一幅图像或该图像中的一部分 ;再通过对距离的判断及相应的处理,即可将构成同一幅图像的多个连通区域进行合并,从而实现对某一幅图像的完整的识别。

[0045] 在上述技术方案中,优选地,所述版面分析单元 106 还包括 :孔洞填补子单元 1069,用于对所述连通的非文字对象区域中存在的孔洞进行填补。

[0046] 在该技术方案中,通过对连通的非文字对象区域中存在的孔洞进行填补,从而能够以整体为对象来处理对应的区域,避免了孔洞为处理过程带来的难度和可能造成的意外。

[0047] 在上述技术方案中,优选地,所述关联区块确定单元 110 包括 :位置关系检测子单元 1100,用于检测所述图像区块与所述文字区块之间的位置关系,其中,若指定图像区块与至少一个文字区块相交,或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离,则判定所述至少一个文字区块与所述指定图像区块相关联。

[0048] 在该技术方案中,由于图像往往存在一些文字描述,比如图标题、图中的标注文字等等,这些文字与图像之间是相关联的,应该划分至相同的区块。通过上述处理,使得分割出来的复合图区块更加准确。

[0049] 在上述技术方案中，优选地，还包括：图像生成单元 114，用于将所述复合图区块生成为图像文件；图像保存单元 116，用于保存所述图像文件。

[0050] 在该技术方案中，直接将分割出来的复合图区块以图像文件的形式进行保存，从而不必对图元 ID 进行管理，尤其是当这些复合图区块中包含有数量很多的图元时，以图像文件进行处理的方式，显然有利于提升处理效率。

[0051] 图 2 示出了根据本发明的实施例的版式文档中复合图的提取方法的流程图。

[0052] 如图 2 所示，根据本发明的实施例的版式文档中复合图的提取方法，包括：步骤 202，对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型；步骤 204，提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层；步骤 206，分别对所述文字图层和所述非文字图层进行版面分析处理，以生成所述文字图层中的文字区块和所述非文字图层中的图像区块；步骤 208，确定与每个所述图像区块相关联的文字区块，以合并为复合图区块；步骤 210，存储所述复合图区块包含的所有图元的标识。

[0053] 在该技术方案中，通过对版式文档进行解析后，将得到的图元分别构成文字图层(包含文字图元)和非文字图层(包含图像图元等)，然后分别对每个图层进行区块分类，最终利用区块之间的关系判定复合图区块，以实现对复合图区块的分割，并确保对文字图元和非文字图元的妥善处理。在生成多个图层时，具体地，可以先提取所有的文字图元以形成文字图层，然后将文字图元过滤以利用剩余的元素构成非文字图元。本方案可以对图文混排、包含图像和图注信息等复杂情况进行有效地分析，从而准确地分割出其中的复合图区块。复合图区块中可以包含一个或多个复合图，还可以包含复合图中或周围的图注等文字。通过记录所有构成该复合图区块的图元的标识，如图元 ID，从而能够利用这些图元 ID 来对应出该复合图区块，实现了将该区块与整个版面的分离，方便进行流式重排等处理。

[0054] 在上述技术方案中，优选地，对所述文字图层进行版面分析处理的步骤包括：对所述文字图层中的文字图元进行聚类处理，以对所述文字图元进行分类，其中，对于同类别的多个文字图元，若对应的最小外接矩形相交或间隔距离小于预设距离，则将所述多个文字图元作为文字图元集合，并将所述文字图元集合的最小外接矩形作为一个所述文字区块。

[0055] 在该技术方案中，通过基于页面内文字图元邻域特征相似性的聚类算法处理，可以有效地对文字图元进行分类，从而确定每个文字图元应该属于正文部分还是复合图部分。通过对距离的判断及相应的处理，从而确定多个文字图元的构成关系，比如用于构成一个文字区块，该文字区块对应于一个完整的字符。

[0056] 在上述技术方案中，优选地，对所述非文字图层进行版面分析处理的步骤包括：获取所述非文字图层中的非文字图元的纹理特征，并根据预设的特征阈值，检测出所述非文字图层中连通的非文字对象区域，其中，对于多个所述连通的非文字对象区域，若对应的最小外接矩形相交或间隔距离小于预设距离，则将多个所述连通的非文字对象区域作为区域集合，并将所述区域集合的最小外接矩形作为所述图像区块。

[0057] 在该技术方案中，利用基于纹理分析和形态学处理的页面非文字对象的连通域检测，从而识别出版面中的连通的非文字对象区域，该区域实际上对应于版面中的一幅图像或该图像中的一部分；再通过对距离的判断及相应的处理，即可将构成同一幅图像的多个连通区域进行合并，从而实现对某一幅图像的完整的识别。

[0058] 在上述技术方案中，优选地，还包括：对所述连通的非文字对象区域中存在的孔洞

进行填补。

[0059] 在该技术方案中，通过对连通的非文字对象区域中存在的孔洞进行填补，从而能够以整体为对象来处理对应的区域，避免了孔洞为处理过程带来的难度和可能造成的意外。

[0060] 在上述技术方案中，优选地，所述确定与每个所述图像区块相关联的文字区块的步骤包括：检测所述图像区块与所述文字区块之间的位置关系，若指定图像区块与至少一个文字区块相交，或所述指定图像区块与所述至少一个文字区块的间隔距离小于预设距离，则判定所述至少一个文字区块与所述指定图像区块相关联。

[0061] 在该技术方案中，由于图像往往存在一些文字描述，比如图标题、图中的标注文字等等，这些文字与图像之间是相关联的，应该划分至相同的区块。通过上述处理，使得分割出来的复合图区块更加准确。

[0062] 在上述技术方案中，优选地，还包括：将所述复合图区块保存为图像文件。

[0063] 在该技术方案中，直接将分割出来的复合图区块以图像文件的形式进行保存，从而不必对图元 ID 进行管理，尤其是当这些复合图区块中包含有数量很多的图元时，以图像文件进行处理的方式，显然有利于提升处理效率。

[0064] 图 3 示出了根据本发明的实施例的对版式文档中的复合图进行提取的具体流程图。

[0065] 如图 3 所示，根据本发明的实施例的对版式文档中的复合图进行提取的具体流程包括：

[0066] 步骤 302，利用解析引擎对原始的版式文档进行解析。

[0067] 步骤 304，根据解析结果，获取该版式文档中包含的图元。

[0068] 步骤 306，判断图元的类型，比如可以根据解析出来的图元类型进行分辨，其中，若为文字类型，则获取该文字图元并进入步骤 310，否则进入步骤 308。

[0069] 步骤 308，依据该图元的类型进行相应的处理。

[0070] 步骤 310，对页面进行分层处理，具体地，根据步骤 306 获取的文字图元，将所有的文字图元构成文字图层，然后将所有的文字图元过滤后，剩余的图元构成非文字图层。

[0071] 当然，这种通过对文字图元进行获取、分层、过滤、再分层的方式仅为图层构建的一种方式，实际上，也可以通过对非文字图元进行获取来实现，或是分别获取文字图元和非文字图元以同时分别构成图层等。

[0072] 下面将分别对文字图层和非文字图层进行处理，其中，步骤 312 至步骤 316 对文字图层进行了处理，而步骤 318 至步骤 322 对非文字图层进行处理，以下分别进行详细说明。

[0073] 步骤 312，构建 Delaunay 三角剖分的邻域关系。具体地，以页面内文字图元的外接矩形的质心为顶点 V，通过采用 Delaunay 三角剖分，构建页面内文字图元的邻域关系 G=(V,E)。

[0074] 步骤 314，采用基于图的并查集算法对文字图元聚类。具体地，包括：

[0075] 1、对构建的无向图中连接相邻节点 v_i 和 v_j 的边 E，计算其权重 $w(v_i, v_j)$：

[0076] $$w(v_i, v_j) = \sum_k \lambda_k f_k(v_i, v_j)$$

[0077] 其中，k 为相邻节点 v_i 和 v_j 的特征相似度函数 $f_k(v_i, v_j)$ 的维数，可以视不同的应

用场景选择特征函数的维数，λ_k 为选择的特征函数的权系数。

[0078] 2、为将所有的文字图元进行聚类，根据页面内节点间的统计分布，定义节点集合间的类内距离 Int(C) 和类间距离 $Dif(C_1,C_2)$。具体的聚类过程采用基于图的并查集算法：

[0079] 1）将页面内每个节点，即每个文字图元，当成一个集合，遍历无向图的边；

[0080] 2）查询连接边的两个节点分别属于哪个集合；

[0081] 3)如果节点集合 C_1 和 C_2 的类间距离满足条件 $Dif(C_1,C_2) \leqslant \min(Int(C_1),Int(C_2))$，则合并这两个集合，形成新的集合 C'_1，并删去集合 C_1 和 C_2；而当 $Dif(C_1,C_2)>\min(Int(C_1),Int(C_2))$，则不进行合并操作；

[0082] 4）遍历完所有的边，完成对文字图元的聚类，计算相近且同类文字图元集合的外接矩形框。

[0083] 步骤 318，计算纹理特征，检测连通区域。具体地，包括：计算该图层的图像纹理特征，采用灰度共生矩阵捕捉非文字对象的纹理特征，主要包括图像局部熵和局部标准差，设定与页面大小相关的阈值，检测出页面图像中连通的非文字对象区域。

[0084] 步骤 320，利用形态学处理填充连通区域内的孔洞。具体地，可以采用基于形态学腐蚀算子的孔洞填充算法，将连通区域中的孔洞进行填补。

[0085] 步骤 322，检测连通区域的外接矩形框，区域生长成非文字对象的外接矩形框。具体地，首先计算出每个检测到非文字对象连通区域的外接矩形(最小外接矩形，作为该非文字对象连通区域对应的范围)，然后对重叠相交或邻接距离小于设定间距的矩形框进行区域生长，计算最终的外接矩形框。

[0086] 步骤 324，判断矩形框是否合并。具体地，在对文字图层和非文字图层分别进行处理后，可以分别得到一些文字或非文字区域的外接矩形框，这里，通过将这些外接矩形框进行距离上的比较，以确定是否将某些外接矩形框进行合并处理，判断过程包括：

[0087] 如果非文字层的非文字连通对象和文字层的文字类矩形框相交，或者距离小于设定间距，则合并这两个矩形框；

[0088] 如果距离大于字符间距，则不进行合并操作。

[0089] 步骤 326，根据任意两个外接矩形框的合并处理结果(包括进行了合并或没有进行合并)，判断结果是否收敛，若是，则进入步骤 328，否则返回步骤 324，从而确保对所有的矩形框都进行了合并处理，以实现对复合图的准确分割。

[0090] 步骤 328，返回最终矩形框集合，保存文件。具体地，当矩形框没有新的合并操作时，算法收敛，最终返回复合图的外接矩形框信息(确定对应的区域的信息)，将构成复合图所对应的图元 ID 集合保存成 XML 文件。或者，也可以采用将分割出来的复合图保存为图像文件的形式，从而避免对数量众多的图元 ID 进行管理时的效率低下的问题。

[0091] 下面将列举多个实施例，分别具体地对本发明的技术方案进行详细说明。

[0092] 图 4A 至图 4D 示出了根据本发明的一个实施例的对版式文档中的复合图进行提取的示意图。

[0093] 如图 4A 至图 4D 所示，以中文版式文档图书“台灣古厝圖鑑”中的一张双栏页面为例，该图中包括：由文字图元构成的正文文字部分 402A、图注文字部分 402B、页面文字部分 402D 和图中文字部分 402E，以及由非文字图元构成的装饰性复合图 404A、分栏线复合图 404B、正文插图复合图 404C 和正文插图复合图 404D，下面将按照图 3 给出的流程分割出页

面中的复合图对象。

[0094] 首先需要通过解析引擎获取版式文档的各种图元，然后对路径图元进行分组，得到仅包含文字图元的文字图层和包含其余的非文字图元的非文字图层。

[0095] 具体地，可以通过提取文档内嵌的文字图元，并使用提取出来的页面内的文字图元构成文字图层；然后，将文字图元过滤后，利用剩余的非文字图元构成非文字图层。如图 4A 所示，对该页面中的所有文字的外接矩形框进行了可视化显示；将页面内文字图元过滤，重新绘制页面，形成非文字图层，如图 4B 所示。

[0096] 然后需要分别对文字图层和非文字图层进行处理，处理流程如图 3 中的步骤 312 至步骤 316、步骤 318 至步骤 322 所示。

[0097] 1、针对文字图层进行聚类处理，图 4C 给出以页面内文字图元的外接矩形的质心为顶点，采用 Delaunay 三角剖分构建的文字图元邻域关系。以解析后版式文档中包含的文字图元的字体信息为特征，设计基于图的并查集算法，对文字聚类的结果采用不同的颜色来显示，如图 4C 所示，该页面内的文字聚为 4 类，分别属于正文文字部分 402A、图注文字部分 402B、页面文字部分 402D 和图中文字部分 402E。

[0098] 2、对非文字图层进行基于纹理分析和形态学处理的连通域检测，并对得到的连通域进行关联分析和区域生长，以及确定生长完成后的连通域的外接矩形框。

[0099] 3、融合对文字图层和非文字图层的分割结果，该页面的复合图的最终分割结果如图 4D 所示，页面左边的装饰性复合图 404A，内部包括图中文字部分 402E，该图被准确的分割出来；页面下方的正文插图复合图 404C 包含大量的路径操作和环绕其四周文字图元，其分割难度是比较大的，但采用本发明的方法，也被准确的分割出来；对于分栏线复合图 404B 和灰度图（正文插图复合图 404D），都被准确的分割出来。分割结果可直接用于版式文档的流式重排应用。

[0100] 图 5A 至图 5D 示出了根据本发明的另一个实施例的对版式文档中的复合图进行提取的示意图。

[0101] 如图 5A 至图 5D 所示，以英文版式文档图书“Advances in Selected Plant Physiology Aspects”中的一张单栏页面为例，该图中包括：由文字图元构成的正文文字部分 502A 和页眉文字部分 502B，以及由非文字图元构成的正文插图复合图 504A 和分栏线复合图 504B，下面将按照图 3 给出的流程分割出页面中的复合图对象。

[0102] 首先需要通过解析引擎获取版式文档的各种图元，然后对路径图元进行分组，得到仅包含文字图元的文字图层和包含其余的非文字图元的非文字图层。

[0103] 具体地，可以通过提取文档内嵌的文字图元，并使用提取出来的页面内的文字图元构成文字图层；然后，将文字图元过滤后，利用剩余的非文字图元构成非文字图层。如图 5A 所示，对该页面中的所有文字的外接矩形框进行了可视化显示；将页面内文字图元过滤，重新绘制页面，形成非文字图层，如图 5B 所示。

[0104] 然后需要分别对文字图层和非文字图层进行处理，处理流程如图 3 中的步骤 312 至步骤 316、步骤 318 至步骤 322 所示。

[0105] 1、针对文字图层进行聚类处理，图 5C 给出以页面内文字图元的外接矩形的质心为顶点，采用 Delaunay 三角剖分构建的文字图元邻域关系。以解析后版式文档中包含的文字图元的字体信息为特征，设计基于图的并查集算法，对文字聚类的结果采用不同的颜色

来显示，如图 5C 所示，该页面内的文字聚为 2 类，分别属于正文文字部分 502A 和页眉文字部分 502B。

[0106] 2、对非文字图层进行基于纹理分析和形态学处理的连通域检测，并对得到的连通域进行关联分析和区域生长，以及确定生长完成后的连通域的外接矩形框。

[0107] 3、融合对文字图层和非文字图层的分割结果，该页面的复合图的最终分割结果如图 5D 所示，页面中间的正文插图复合图 504A，由扫描的 3 个子图构成，图中文字皆属于扫描子图，由这些子图构成的复合图被准确的分割出来；页面上方的分栏线复合图 504B，都被准确的分割出来。分割结果可直接用于版式文档的流式重排应用。

[0108] 以上结合附图详细说明了本发明的技术方案，本发明将基于图像的版面分析技术应用到版式文档复合图的结构信息提取，结合图像文档处理技术和版式文档固有的底层结构信息，为高效可靠的智能文档分析和理解奠定基础，为提高图文及多媒体信息的动态实时混排和跨平台阅读的鲁棒性提供支持。

[0109] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

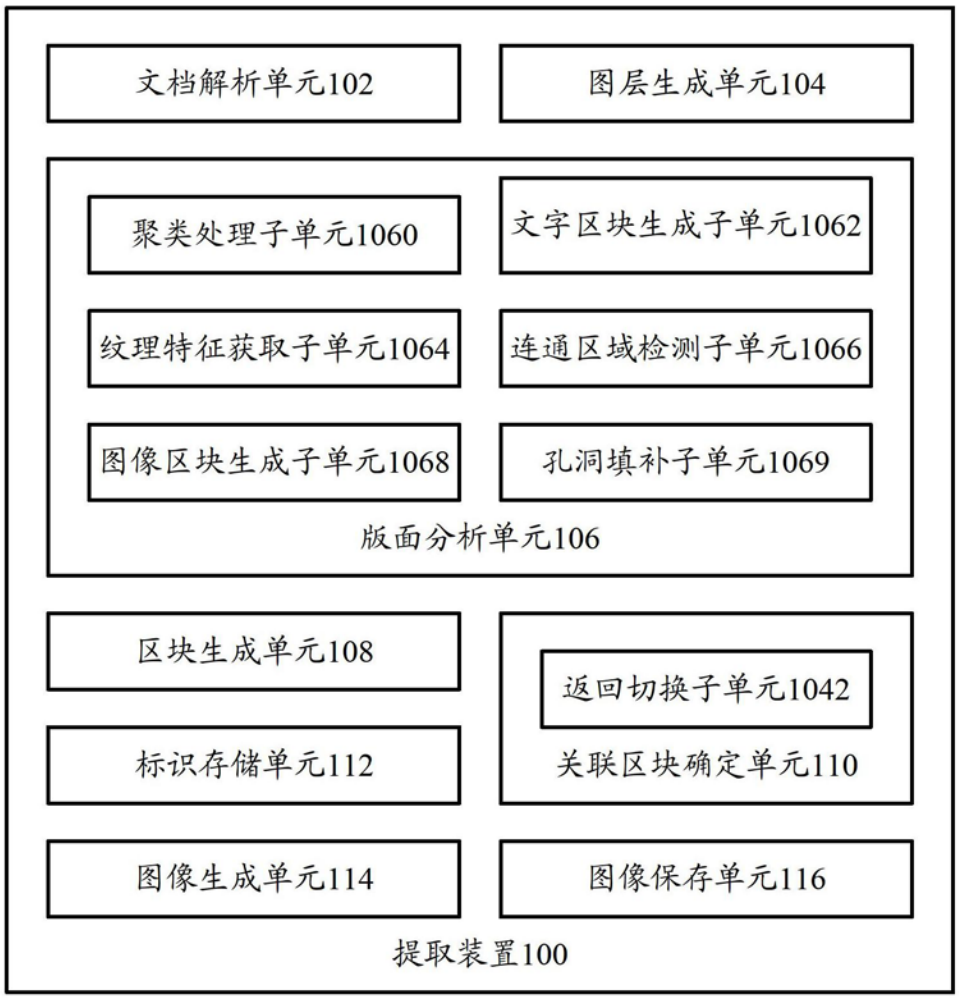

文档解析单元102
图层生成单元104
聚类处理子单元1060
文字区块生成子单元1062
纹理特征获取子单元1064
连通区域检测子单元1066
图像区块生成子单元1068
孔洞填补子单元1069
版面分析单元106
区块生成单元108
返回切换子单元1042
标识存储单元112
关联区块确定单元110
图像生成单元114
图像保存单元116
提取装置100

图 1

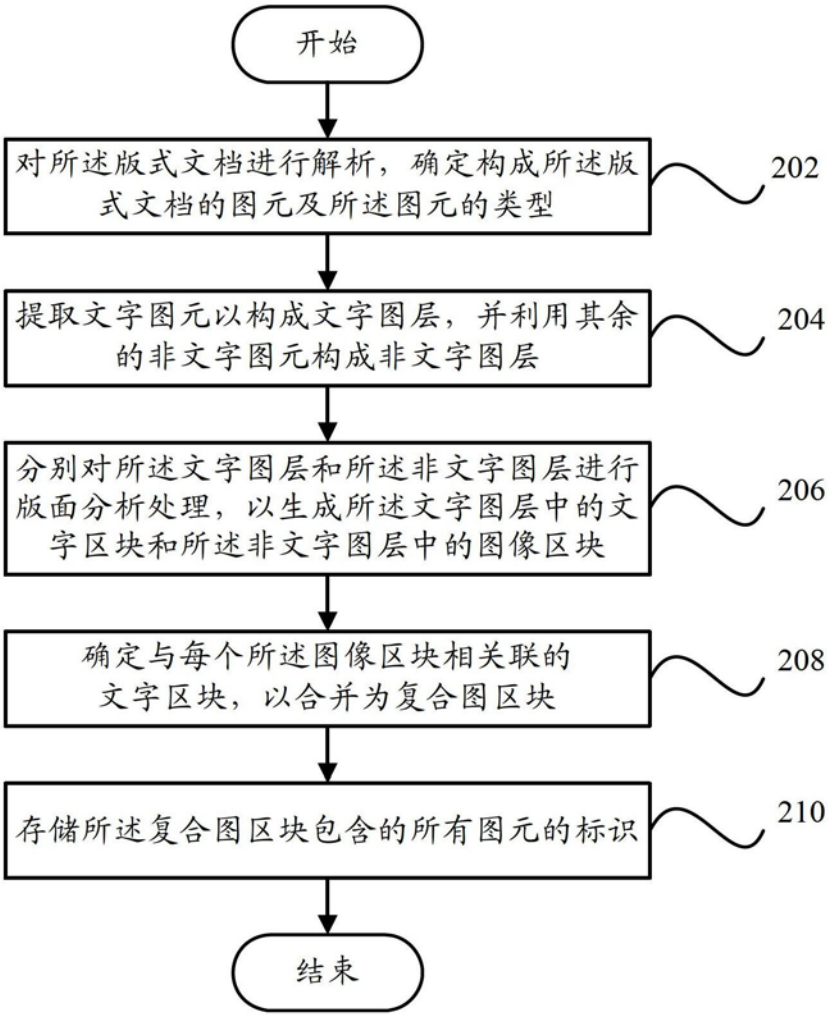

开始
对所述版式文档进行解析，确定构成所述版式文档的图元及所述图元的类型
202
提取文字图元以构成文字图层，并利用其余的非文字图元构成非文字图层
204
分别对所述文字图层和所述非文字图层进行版面分析处理，以生成所述文字图层中的文字区块和所述非文字图层中的图像区块
206
确定与每个所述图像区块相关联的文字区块，以合并为复合图区块
208
存储所述复合图区块包含的所有图元的标识
210
结束

图 2

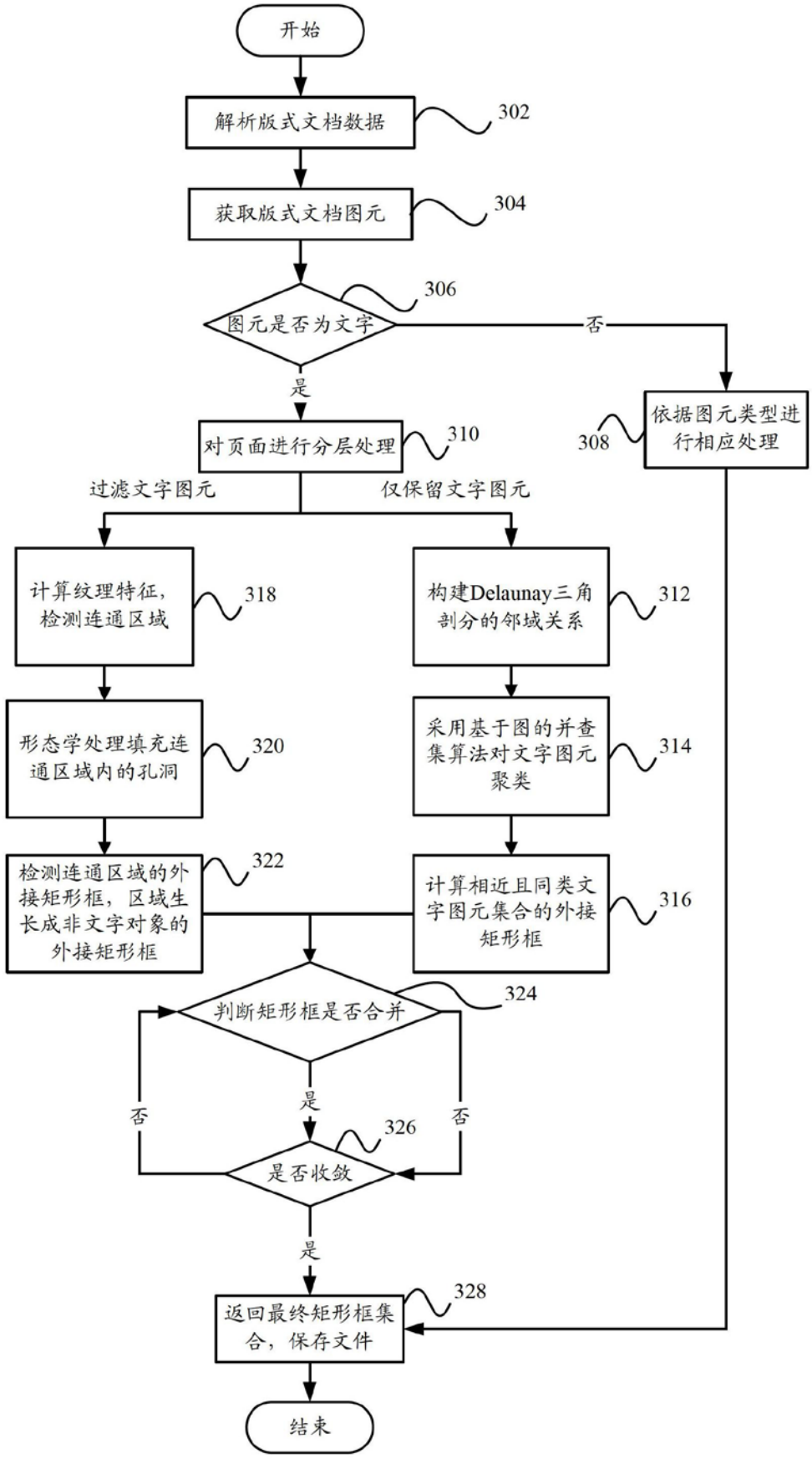

开始
解析版式文档数据
302
获取版式文档图元
304
图元是否为文字
306
否
是
依据图元类型进行相应处理
308
对页面进行分层处理
310
过滤文字图元
仅保留文字图元
计算纹理特征，检测连通区域
318
构建Delaunay三角剖分的邻域关系
312
形态学处理填充连通区域内的孔洞
320
采用基于图的并查集算法对文字图元聚类
314
检测连通区域的外接矩形框，区域生长成非文字对象的外接矩形框
322
计算相近且同类文字图元集合的外接矩形框
316
判断矩形框是否合并
324
否
是
否
是否收敛
326
是
返回最终矩形框集合，保存文件
328
结束

图 3

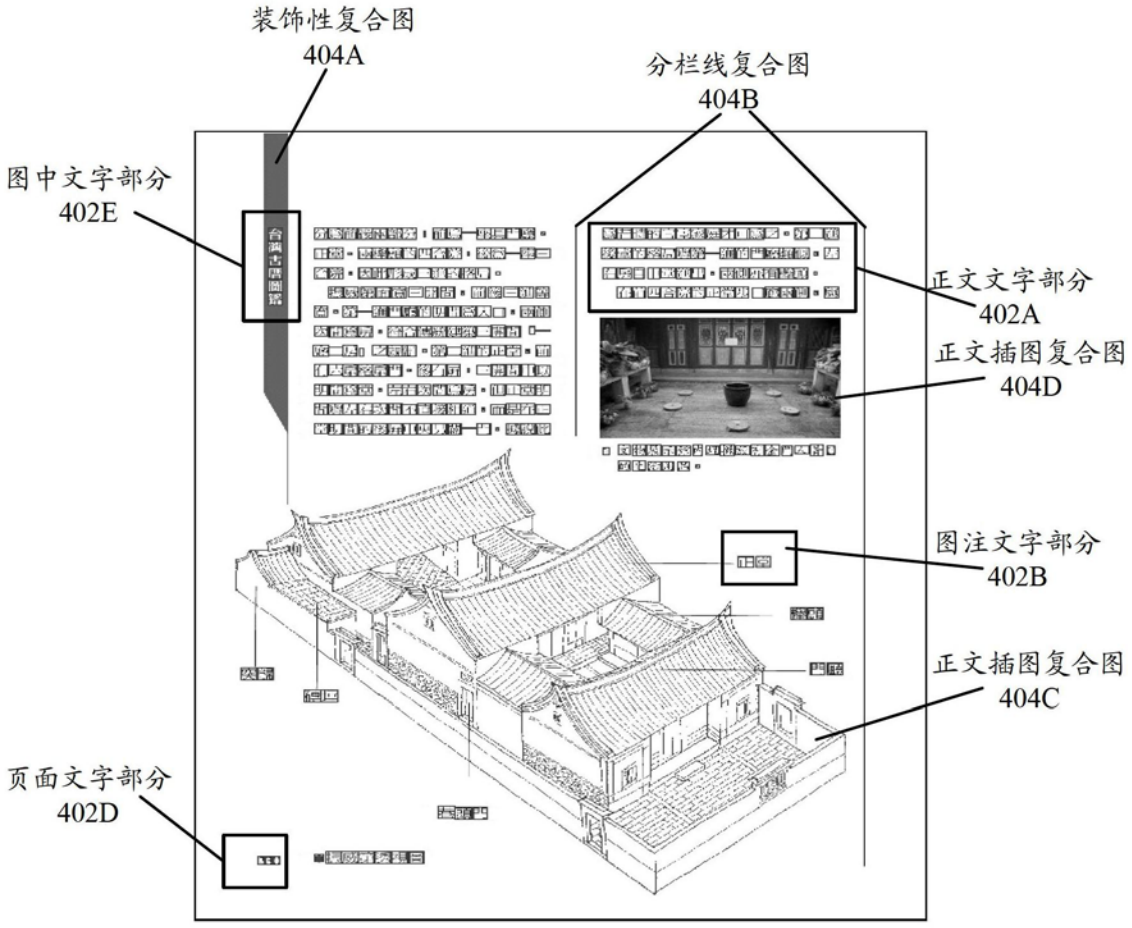

装饰性复合图
404A
分栏线复合图
404B
图中文字部分
402E
正文文字部分
402A
正文插图复合图
404D
图注文字部分
402B
正文插图复合图
404C
页面文字部分
402D

图 4A

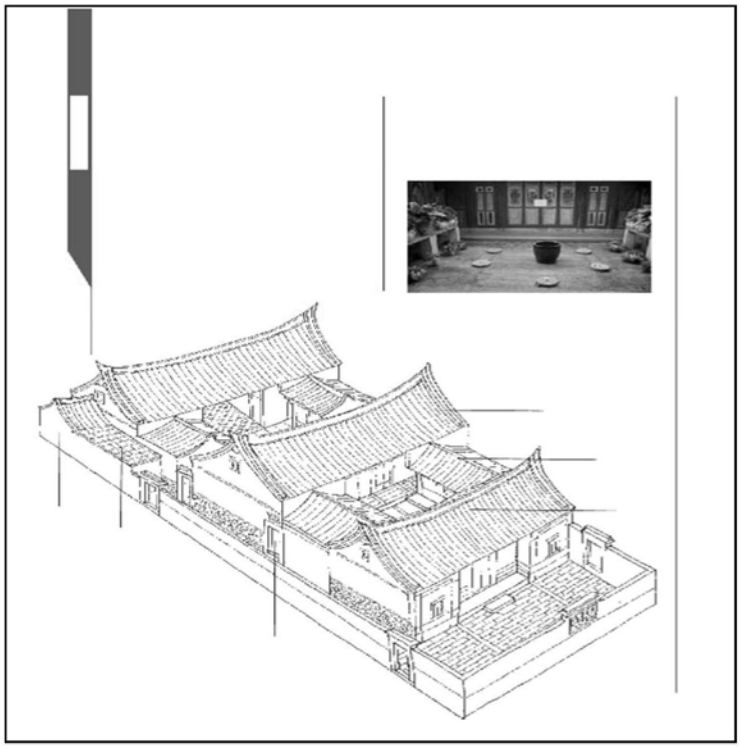


图 4B

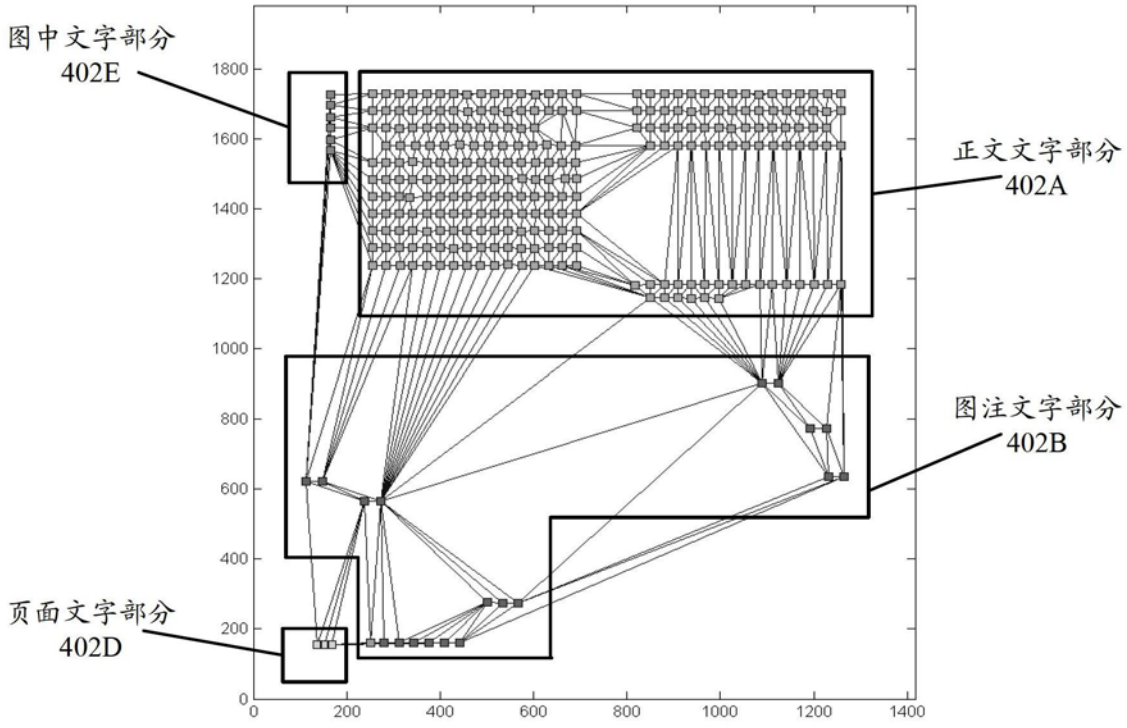

图中文字部分
402E
正文文字部分
402A
图注文字部分
402B
页面文字部分
402D
0
200
400
600
800
1000
1200
1400
1600
1800

图 4C

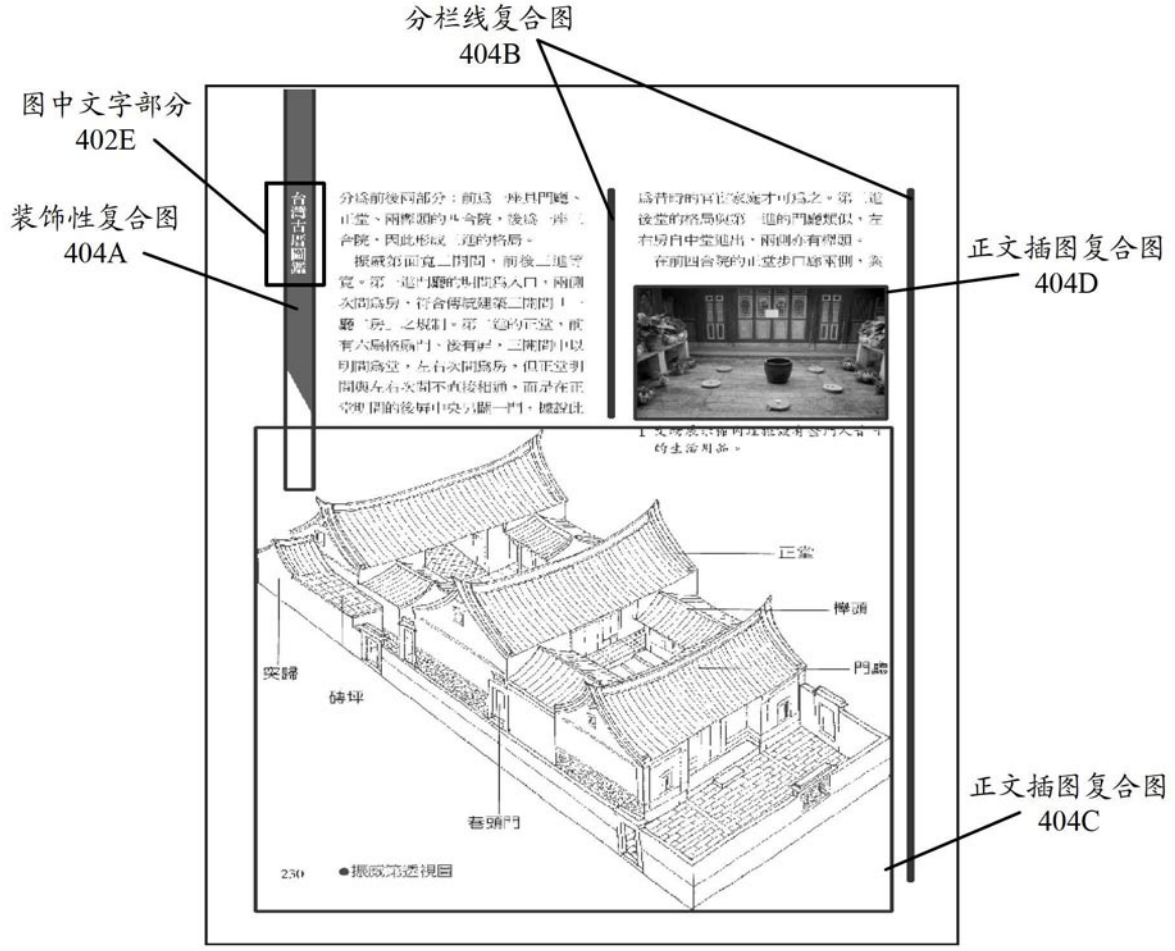

分栏线复合图
404B
图中文字部分
402E
装饰性复合图
404A
正文插图复合图
404D
正文插图复合图
404C
台灣古厝圖鑑
正堂
樑頭
門廳
龍埕
護龍
春頭門
230
●振威第透視圖

图 4D

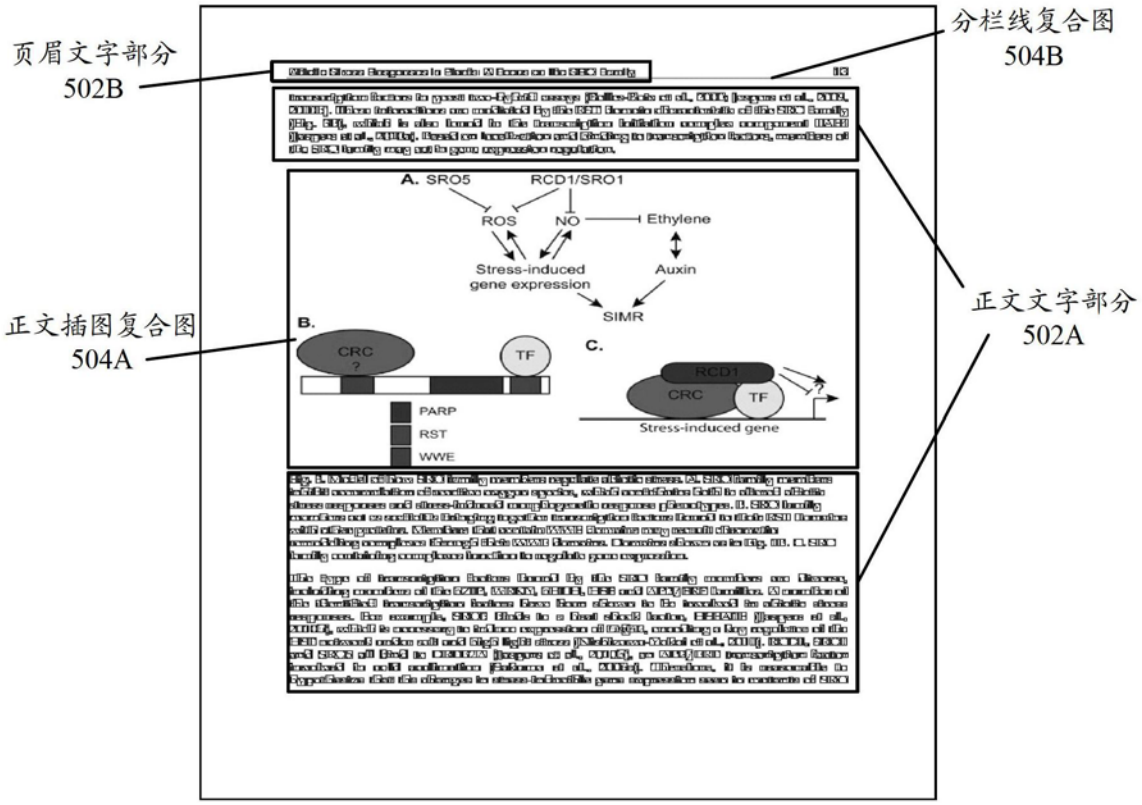

页眉文字部分
502B
分栏线复合图
504B
正文插图复合图
504A
正文文字部分
502A
A. SRO5
RCD1/SRO1
ROS
NO
Ethylene
Stress-induced gene expression
Auxin
SIMR
B.
CRC
?
TF
PARP
RST
WWE
C.
RCD1
CRC
TF
?
Stress-induced gene

图 5A

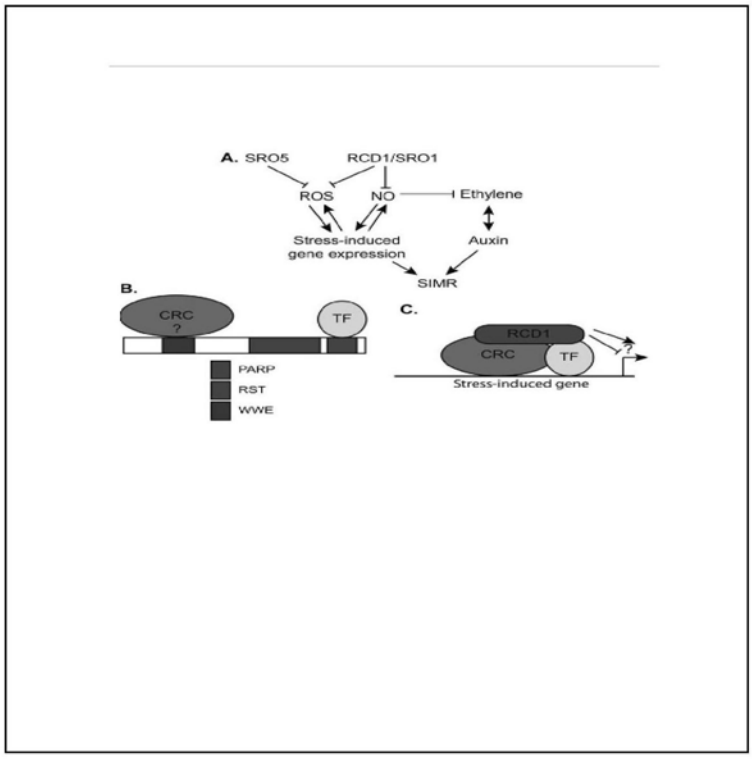

A. SRO5
RCD1/SRO1
ROS
NO
Ethylene
Stress-induced gene expression
Auxin
SIMR
B.
CRC
?
TF
PARP
RST
WWE
C.
RCD1
CRC
TF
?
Stress-induced gene

图 5B

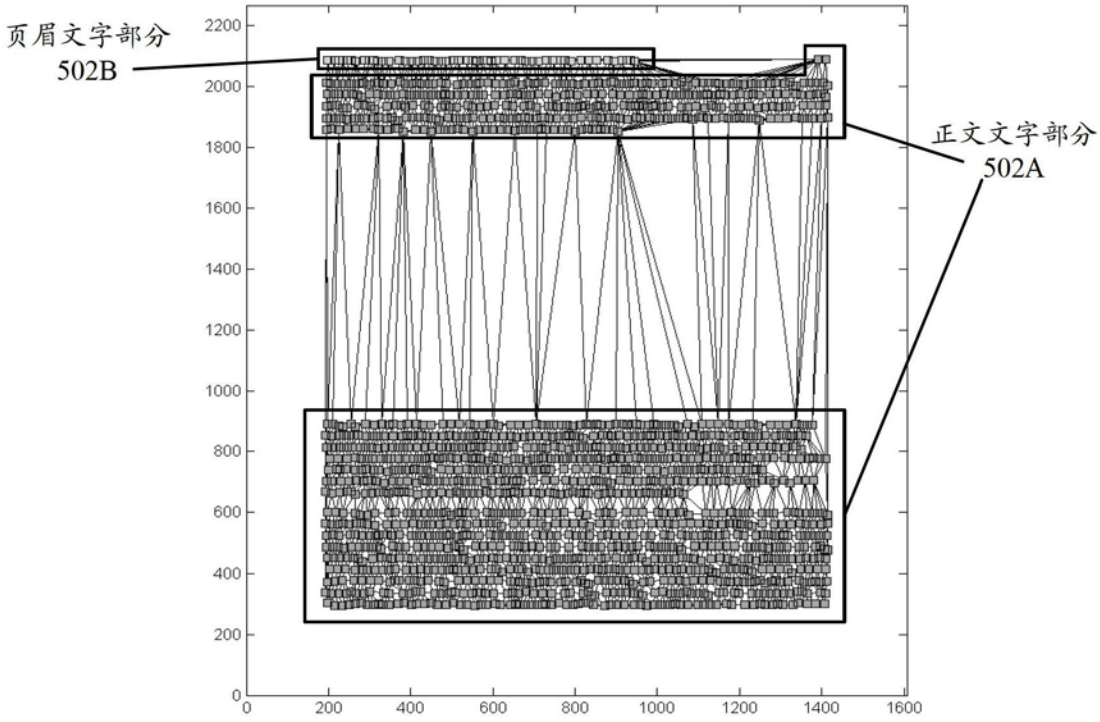

页眉文字部分
502B
正文文字部分
502A
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600

图 5C

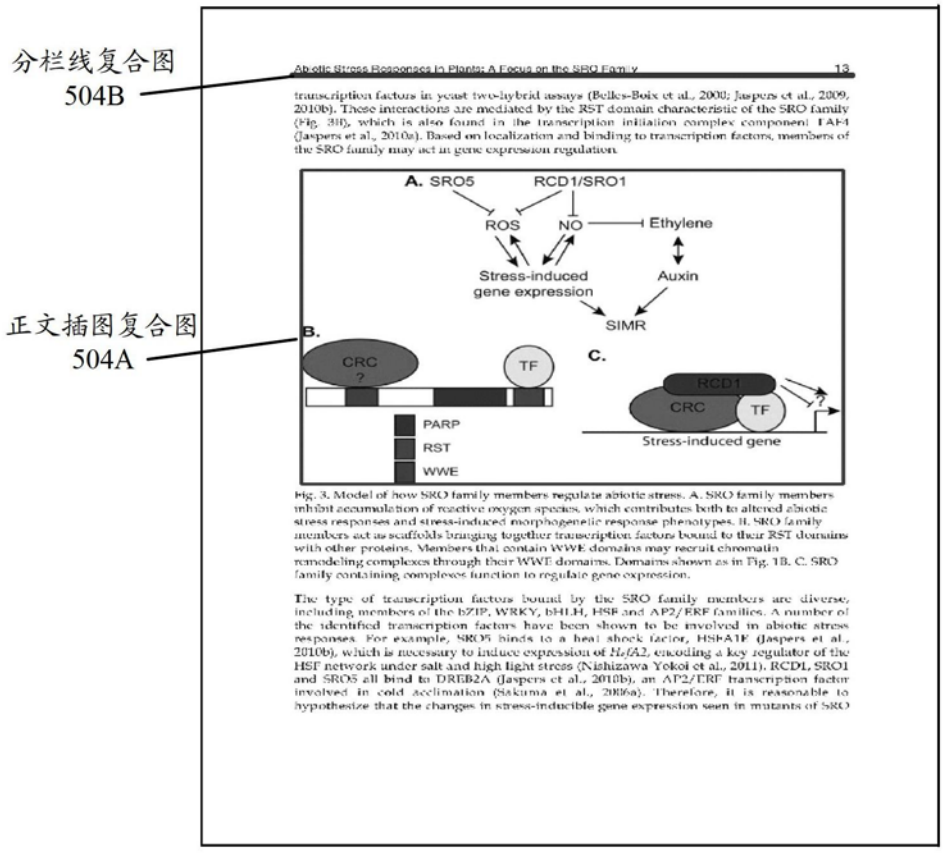

分栏线复合图
504B
正文插图复合图
504A
Abiotic Stress Responses in Plants: A Focus on the SRO Family 13
transcription factors in yeast two-hybrid assays (Belles-Boix et al., 2000; Jaspers et al., 2009, 2010b). These interactions are mediated by the RST domain characteristic of the SRO family (Fig. 3B), which is also found in the transcription initiation complex component TAF4 (Jaspers et al., 2010a). Based on localization and binding to transcription factors, members of the SRO family may act in gene expression regulation.
A. SRO5
RCD1/SRO1
ROS
NO
Ethylene
Stress-induced gene expression
Auxin
SIMR
B.
CRC
?
TF
PARP
RST
WWE
C.
RCD1
CRC
TF
?
Stress-induced gene
Fig. 3. Model of how SRO family members regulate abiotic stress. A. SRO family members inhibit accumulation of reactive oxygen species, which contributes both to altered abiotic stress responses and stress-induced morphogenetic response phenotypes. B. SRO family members act as scaffolds bringing together transcription factors bound to their RST domains with other proteins. Members that contain WWE domains may recruit chromatin remodeling complexes through their WWE domains. Domains shown as in Fig. 1B. C. SRO family containing complexes function to regulate gene expression.
The type of transcription factors bound by the SRO family members are diverse, including members of the bZIP, WRKY, bHLH, HSF and AP2/ERF families. A number of the identified transcription factors have been shown to be involved in abiotic stress responses. For example, SRO5 binds to a heat shock factor, HSFA1F (Jaspers et al., 2010b), which is necessary to induce expression of HsfA2, encoding a key regulator of the HSF network under salt and high light stress (Nishizawa Yokoi et al., 2011). RCDL SRO1 and SRO5 all bind to DREB2A (Jaspers et al., 2010b), an AP2/ERF transcription factor involved in cold acclimation (Sakuma et al., 2006a). Therefore, it is reasonable to hypothesize that the changes in stress-inducible gene expression seen in mutants of SRO

图 5D