(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 104142961 A
(43)申请公布日 2014.11.12

(21)申请号 201310172879.3

(22)申请日 2013.05.10

(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.
G06F 17/30(2006.01)
G06F 17/22(2006.01)

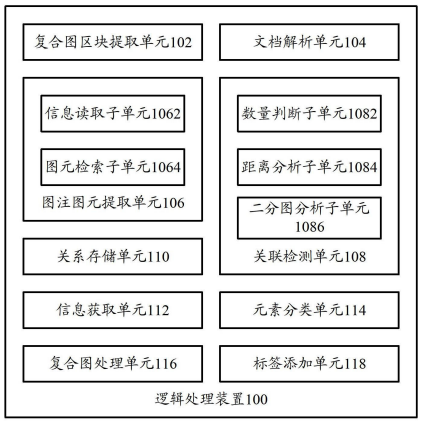
权利要求书2页 说明书10页 附图7页

(54)发明名称

版式文档中复合图的逻辑处理装置和逻辑处理方法

(57)摘要

本发明提供了一种版式文档中复合图的逻辑处理装置，包括：复合图区块提取单元，用于提取所述版式文档中的复合图区块；文档解析单元，用于对所述版式文档进行解析，以获取其中包含的文字图元；图注图元提取单元，用于从所述文字图元中提取出图注图元；关联检测单元，用于检测所述复合图区块与所述图注图元之间的关联关系；关系存储单元，用于存储检测到的所述关联关系。本发明还提出了一种版式文档中复合图的逻辑处理方法。通过本发明的技术方案，可以对从版式文档中分割出来的复合图进行妥善的逻辑处理，从而便于实现对版式文档中图文混排版面的复合图进行版面理解，避免逻辑错误。



CN 104142961 A

1. 一种版式文档中复合图的逻辑处理装置，其特征在于，包括：

复合图区块提取单元，用于提取所述版式文档中的复合图区块；

文档解析单元，用于对所述版式文档进行解析，以获取其中包含的文字图元；

图注图元提取单元，用于从所述文字图元中提取出图注图元；

关联检测单元，用于检测所述复合图区块与所述图注图元之间的关联关系；

关系存储单元，用于存储检测到的所述关联关系。

2. 根据权利要求 1 所述的版式文档中复合图的逻辑处理装置，其特征在于，还包括：

信息获取单元，用于获取所述复合图区块的特征属性信息；

元素分类单元，用于根据所述特征属性信息，对所述复合图区块中包含的元素进行分类；

复合图处理单元，用于根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供所述关联检测单元检测所述正文插图复合图与所述图注图元之间的关联关系。

3. 根据权利要求 2 所述的版式文档中复合图的逻辑处理装置，其特征在于，还包括：

标签添加单元，用于为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及

所述关系存储单元还用于：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

4. 根据权利要求 2 所述的版式文档中复合图的逻辑处理装置，其特征在于，所述关联检测单元包括：

数量判断子单元，用于判断所述复合图区块中包含的正文插图复合图的数量；

距离分析子单元，用于在所述复合图区块中仅包含一个正文插图复合图的情况下，选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；

二分图分析子单元，用于在所述复合图区块中包含多个正文插图复合图的情况下，将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

5. 根据权利要求 1 至 4 中任一项所述的版式文档中复合图的逻辑处理装置，其特征在于，所述图注图元提取单元包括：

信息读取子单元，用于读取预设的所述图注图元的排版特征信息；

图元检索子单元，用于根据读取的所述排版特征信息，在所述文档解析单元解析出的所述文字图元中进行检索，以获取所述图注图元。

6. 一种版式文档中复合图的逻辑处理方法，其特征在于，包括：

步骤 202，提取所述版式文档中的复合图区块；

步骤 204，在从所述版式文档解析出的文字图元中，提取出图注图元；

步骤 206，检测所述复合图区块与所述图注图元之间的关联关系；

步骤 208，存储检测到的所述关联关系。

7. 根据权利要求 6 所述的版式文档中复合图的逻辑处理方法，其特征在于，所述步骤 206 之前，还包括：

获取所述复合图区块的特征属性信息，以对所述复合图区块中包含的元素进行分类；

根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供在所述步骤 206 中，检测所述正文插图复合图与所述图注图元之间的关联关系。

8. 根据权利要求 7 所述的版式文档中复合图的逻辑处理方法，其特征在于，还包括：

为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及

所述步骤 208 还包括：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

9. 根据权利要求 7 所述的版式文档中复合图的逻辑处理方法，其特征在于，所述步骤 206 包括：

若所述复合图区块中仅包含一个正文插图复合图，则选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；

若所述复合图区块中包含多个正文插图复合图，则将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

10. 根据权利要求 6 至 9 中任一项所述的版式文档中复合图的逻辑处理方法，其特征在于，所述步骤 204 包括：

读取预设的所述图注图元的排版特征信息；

根据读取的所述排版特征信息，在解析出的所述文字图元中进行检索，以获取所述图注图元。

版式文档中复合图的逻辑处理装置和逻辑处理方法

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逻辑处理装置和一种版式文档中复合图的逻辑处理方法。

背景技术

[0002] 根据版式文档的生成过程，文档是数据和结构的集合，具体包括内容数据、物理结构和逻辑结构。内容数据指文档中诸如文本、图像、图形等数据。物理结构是对内容数据在页面的布局、组合的描述，包括文本行、文本块、图表等。逻辑结构是对内容数据所反映的信息、信息间关系的描述，不仅包括页面元素的逻辑属性如正文段落、摘要、标题、表格等，也包括文档的层次关系和文档元素之间的逻辑关联关系，如图像和图注的关联等。

[0003] 文档分析是对文档物理结构进行抽取，而文档理解则是在物理结构和逻辑结构之间建立映射关系。对于文档分析任务来说，可得到的输入是文档最后成型的形态，物理和逻辑结构信息都没有显式的体现，文档生成时使用的逻辑模型和物理模型需要反推出来，最大程度地恢复文档的物理和逻辑结构。而在现实应用中，移动设备的可读性需求使物理和逻辑结构的恢复成为重中之重。

[0004] 在对物理和逻辑结构的恢复中，可以从页面层次提取文档的逻辑结构信息，将页面内已抽取的物理结构块根据其逻辑功能进行标注。目前，基于传统图像文档的页面逻辑结构分析得益于人工智能领域的发展。逻辑结构分析的发展正从基于先验规则的方法转向基于机器学习的方法。与传统图像文档方法不同的是，版式文档提供的信息可以辅助版面理解。但固定版式文档中存在大量拼接的图元、相互叠加的图层。这些数据并不能直接用于构造文档的逻辑结构，而需要根据空间关系进行拼接、叠加等操作后才能确定其所要展示的内容。页面内非文本对象的分类及识别和标注是文档理解的重点之一，其中，图文混排版面的复合图的分析和理解皆有挑战性。

[0005] 因此，需要一种新的版式文档中复合图的逻辑处理技术，可以对从版式文档中分割出来的复合图进行妥善的逻辑处理，从而便于实现对版式文档中图文混排版面的复合图进行版面理解，避免逻辑错误。

发明内容

[0006] 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逻辑处理技术，可以对从版式文档中分割出来的复合图进行妥善的逻辑处理，从而便于实现对版式文档中图文混排版面的复合图进行版面理解，避免逻辑错误。

[0007] 有鉴于此，本发明提出了一种版式文档中复合图的逻辑处理装置，包括：复合图区块提取单元，用于提取所述版式文档中的复合图区块；文档解析单元，用于对所述版式文档进行解析，以获取其中包含的文字图元；图注图元提取单元，用于从所述文字图元中提取出图注图元；关联检测单元，用于检测所述复合图区块与所述图注图元之间的关联关系；关系存储单元，用于存储检测到的所述关联关系。

[0008] 在该技术方案中，复合图区块包括版式文档中的图片及图片中或周围的文字等，提取复合图区块是指将这些图片、文字等作为一个整体，将其与版式文档中的正文等部分分割开来，以便于在进行版式文档的流式重排时，能够对复合图进行恰当的排版处理。但由于文档的排版原因，图片与图注（比如位于图片下方，包括图标题或用于描述图片的一段文字等）的间隔较远，而为了能够准确地对复合图区块进行提取，会导致将图片与图注分离开，其中，图片被分割至复合图区块中，而图注被保留在版式文档的其他部分中，则虽然在物理结构上实现了分割，但从逻辑结构上却存在问题。因此，该方案通过将复合图区块与图注图元之间建立关联关系，从而在逻辑结构上完成在复合图区块与图注图元之间的关系建立，使得对于版式文档中的复合图的分割过程更准确、合理。

[0009] 在上述技术方案中，优选地，还包括：信息获取单元，用于获取所述复合图区块的特征属性信息；元素分类单元，用于根据所述特征属性信息，对所述复合图区块中包含的元素进行分类；复合图处理单元，用于根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供所述关联检测单元检测所述正文插图复合图与所述图注图元之间的关联关系。

[0010] 在该技术方案中，复合图区块中包含有正文插图复合图和其他的元素，比如图形商标、公式、分栏线、页眉、页脚、装饰性复合图等等，这些元素可能导致在对复合图区域对应的范围进行确定时，对真正的正文插图复合图对应的范围造成影响。比如正文插图复合图对应的范围是以其最小外接矩形框确定的，而如果不对其他元素进行过滤，可能导致该矩形框比实际范围大，从而可能使得不是图注图元的内容被误认为是图注图元，造成逻辑错误。

[0011] 在上述技术方案中，优选地，还包括：标签添加单元，用于为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及所述关系存储单元还用于：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

[0012] 在该技术方案中，通过对每种元素添加逻辑标签，便于对各种元素对应的逻辑关系进行查看和管理，从而实现对版式文档进行流式转换后，得到更好的显示效果。

[0013] 在上述技术方案中，优选地，所述关联检测单元包括：数量判断子单元，用于判断所述复合图区块中包含的正文插图复合图的数量；距离分析子单元，用于在所述复合图区块中仅包含一个正文插图复合图的情况下，选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；二分图分析子单元，用于在所述复合图区块中包含多个正文插图复合图的情况下，将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

[0014] 在该技术方案中，通过距离靠近原则和二分法最优匹配法，实现对正文插图复合图和图注图元的关联识别，有利于得到更为合理、准确的逻辑关系，以便基于该逻辑关系实现版式文档的流式重排。

[0015] 在上述技术方案中，优选地，所述图注图元提取单元包括：信息读取子单元，用于读取预设的所述图注图元的排版特征信息；图元检索子单元，用于根据读取的所述排版特征信息，在所述文档解析单元解析出的所述文字图元中进行检索，以获取所述图注图元。

[0016] 在该技术方案中，图注图元具有的排版特征信息，比如字体、以“图”等关键字起

始、居中、字数限制、与其他文字之间的位置关系等，通过这些特征信息，既可以找到对应内容的图元，又可以避免将如正文中的“图 1”作为图注图元（具体为图标题，或者也可以为解释性文字等），实现对图注图元的准确获取。

[0017] 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逻辑处理方法，包括：步骤 202，提取所述版式文档中的复合图区块；步骤 204，在从所述版式文档解析出的文字图元中，提取出图注图元；步骤 206，检测所述复合图区块与所述图注图元之间的关联关系；步骤 208，存储检测到的所述关联关系。

[0018] 在该技术方案中，复合图区块包括版式文档中的图片及图片中或周围的文字等，提取复合图区块是指将这些图片、文字等作为一个整体，将其与版式文档中的正文等部分分割开来，以便于在进行版式文档的流式重排时，能够对复合图进行恰当的排版处理。但由于文档的排版原因，图片与图注（比如位于图片下方，包括图标题或用于描述图片的一段文字等）的间隔较远，而为了能够准确地对复合图区块进行提取，会导致将图片与图注分离开，其中，图片被分割至复合图区块中，而图注被保留在版式文档的其他部分中，则虽然在物理结构上实现了分割，但从逻辑结构上却存在问题。因此，该方案通过将复合图区块与图注图元之间建立关联关系，从而在逻辑结构上完成在复合图区块与图注图元之间的关系建立，使得对于版式文档中的复合图的分割过程更准确、合理。

[0019] 在上述技术方案中，优选地，所述步骤 206 之前，还包括：获取所述复合图区块的特征属性信息，以对所述复合图区块中包含的元素进行分类；根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供在所述步骤 206 中，检测所述正文插图复合图与所述图注图元之间的关联关系。

[0020] 在该技术方案中，复合图区块中包含有正文插图复合图和其他的元素，比如图形商标、公式、分栏线、页眉、页脚、装饰性复合图等等，这些元素可能导致在对复合图区域对应的范围进行确定时，对真正的正文插图复合图对应的范围造成影响。比如正文插图复合图对应的范围是以其最小外接矩形框确定的，而如果不对其他元素进行过滤，可能导致该矩形框比实际范围大，从而可能使得不是图注图元的内容被误认为是图注图元，造成逻辑错误。

[0021] 在上述技术方案中，优选地，还包括：为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及所述步骤 208 还包括：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

[0022] 在该技术方案中，通过对每种元素添加逻辑标签，便于对各种元素对应的逻辑关系进行查看和管理，从而实现对版式文档进行流式转换后，得到更好的显示效果。

[0023] 在上述技术方案中，优选地，所述步骤 206 包括：若所述复合图区块中仅包含一个正文插图复合图，则选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；若所述复合图区块中包含多个正文插图复合图，则将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

[0024] 在该技术方案中，通过距离靠近原则和二分法最优匹配法，实现对正文插图复合图和图注图元的关联识别，有利于得到更为合理、准确的逻辑关系，以便基于该逻辑关系实现版式文档的流式重排。

[0025] 在上述技术方案中，优选地，所述步骤 204 包括：读取预设的所述图注图元的排版特征信息；根据读取的所述排版特征信息，在解析出的所述文字图元中进行检索，以获取所述图注图元。

[0026] 在该技术方案中，图注图元具有的排版特征信息，比如字体、以“图”等关键字起始、居中、字数限制、与其他文字之间的位置关系等，通过这些特征信息，既可以找到对应内容的图元，又可以避免将如正文中的“图 1”作为图注图元（具体为图标题，或者也可以为解释性文字等），实现对图注图元的准确获取。

[0027] 通过以上技术方案，可以对从版式文档中分割出来的复合图进行妥善的逻辑处理，从而便于实现对版式文档中图文混排版面的复合图进行版面理解，避免逻辑错误。

附图说明

[0028] 图 1 示出了根据本发明的实施例的版式文档中复合图的逻辑处理装置的框图；

[0029] 图 2 示出了根据本发明的实施例的版式文档中复合图的逻辑处理方法的流程图；

[0030] 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逻辑处理的具体流程图；

[0031] 图 4A 和图 4B 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逻辑处理的示意图；

[0032] 图 5A 和图 5B 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逻辑处理的示意图。

具体实施方式

[0033] 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。

[0034] 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。

[0035] 图 1 示出了根据本发明的实施例的版式文档中复合图的逻辑处理装置的框图。

[0036] 如图 1 所示，根据本发明的实施例的版式文档中复合图的逻辑处理装置 100，包括：复合图区块提取单元 102，用于提取所述版式文档中的复合图区块；文档解析单元 104，用于对所述版式文档进行解析，以获取其中包含的文字图元；图注图元提取单元 106，用于从所述文字图元中提取出图注图元；关联检测单元 108，用于检测所述复合图区块与所述图注图元之间的关联关系；关系存储单元 110，用于存储检测到的所述关联关系。

[0037] 在该技术方案中，复合图区块包括版式文档中的图片及图片中或周围的文字等，提取复合图区块是指将这些图片、文字等作为一个整体，将其与版式文档中的正文等部分分割开来，以便于在进行版式文档的流式重排时，能够对复合图进行恰当的排版处理。但由于文档的排版原因，图片与图注（比如位于图片下方，包括图标题或用于描述图片的一段文字等）的间隔较远，而为了能够准确地对复合图区块进行提取，会导致将图片与图注分离开，其中，图片被分割至复合图区块中，而图注被保留在版式文档的其他部分中，则虽然在

物理结构上实现了分割，但从逻辑结构上却存在问题。因此，该方案通过将复合图区块与图注图元之间建立关联关系，从而在逻辑结构上完成在复合图区块与图注图元之间的关系建立，使得对于版式文档中的复合图的分割过程更准确、合理。

[0038] 在上述技术方案中，优选地，还包括：信息获取单元 112，用于获取所述复合图区块的特征属性信息；元素分类单元 114，用于根据所述特征属性信息，对所述复合图区块中包含的元素进行分类；复合图处理单元 116，用于根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供所述关联检测单元 108 检测所述正文插图复合图与所述图注图元之间的关联关系。

[0039] 在该技术方案中，复合图区块中包含有正文插图复合图和其他的元素，比如图形商标、公式、分栏线、页眉、页脚、装饰性复合图等等，这些元素可能导致在对复合图区域对应的范围进行确定时，对真正的正文插图复合图对应的范围造成影响。比如正文插图复合图对应的范围是以其最小外接矩形框确定的，而如果不对其他元素进行过滤，可能导致该矩形框比实际范围大，从而可能使得不是图注图元的内容被误认为是图注图元，造成逻辑错误。

[0040] 在上述技术方案中，优选地，还包括：标签添加单元 118，用于为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及所述关系存储单元 110 还用于：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

[0041] 在该技术方案中，通过对每种元素添加逻辑标签，便于对各种元素对应的逻辑关系进行查看和管理，从而实现对版式文档进行流式转换后，得到更好的显示效果。

[0042] 在上述技术方案中，优选地，所述关联检测单元 108 包括：数量判断子单元 1082，用于判断所述复合图区块中包含的正文插图复合图的数量；距离分析子单元 1084，用于在所述复合图区块中仅包含一个正文插图复合图的情况下，选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；二分图分析子单元 1086，用于在所述复合图区块中包含多个正文插图复合图的情况下，将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

[0043] 在该技术方案中，通过距离靠近原则和二分法最优匹配法，实现对正文插图复合图和图注图元的关联识别，有利于得到更为合理、准确的逻辑关系，以便基于该逻辑关系实现版式文档的流式重排。

[0044] 在上述技术方案中，优选地，所述图注图元提取单元 106 包括：信息读取子单元 1062，用于读取预设的所述图注图元的排版特征信息；图元检索子单元 1064，用于根据读取的所述排版特征信息，在所述文档解析单元 104 解析出的所述文字图元中进行检索，以获取所述图注图元。

[0045] 在该技术方案中，图注图元具有的排版特征信息，比如字体、以“图”等关键字起始、居中、字数限制、与其他文字之间的位置关系等，通过这些特征信息，既可以找到对应内容的图元，又可以避免将如正文中的“图 1”作为图注图元（具体为图标题，或者也可以为解释性文字等），实现对图注图元的准确获取。

[0046] 图 2 示出了根据本发明的实施例的版式文档中复合图的逻辑处理方法的流程图。

[0047] 如图 2 所示，根据本发明的实施例的版式文档中复合图的逻辑处理方法，包括：步

骤 202，提取所述版式文档中的复合图区块；步骤 204，在从所述版式文档解析出的文字图元中，提取出图注图元；步骤 206，检测所述复合图区块与所述图注图元之间的关联关系；步骤 208，存储检测到的所述关联关系。

[0048] 在该技术方案中，复合图区块包括版式文档中的图片及图片中或周围的文字等，提取复合图区块是指将这些图片、文字等作为一个整体，将其与版式文档中的正文等部分分割开来，以便于在进行版式文档的流式重排时，能够对复合图进行恰当的排版处理。但由于文档的排版原因，图片与图注（比如位于图片下方，包括图标题或用于描述图片的一段文字等）的间隔较远，而为了能够准确地对复合图区块进行提取，会导致将图片与图注分离开，其中，图片被分割至复合图区块中，而图注被保留在版式文档的其他部分中，则虽然在物理结构上实现了分割，但从逻辑结构上却存在问题。因此，该方案通过将复合图区块与图注图元之间建立关联关系，从而在逻辑结构上完成在复合图区块与图注图元之间的关系建立，使得对于版式文档中的复合图的分割过程更准确、合理。

[0049] 在上述技术方案中，优选地，所述步骤 206 之前，还包括：获取所述复合图区块的特征属性信息，以对所述复合图区块中包含的元素进行分类；根据分类结果，保留所述复合图区块中的正文插图复合图，并过滤其他元素，以供在所述步骤 206 中，检测所述正文插图复合图与所述图注图元之间的关联关系。

[0050] 在该技术方案中，复合图区块中包含有正文插图复合图和其他的元素，比如图形商标、公式、分栏线、页眉、页脚、装饰性复合图等等，这些元素可能导致在对复合图区域对应的范围进行确定时，对真正的正文插图复合图对应的范围造成影响。比如正文插图复合图对应的范围是以其最小外接矩形框确定的，而如果不对其他元素进行过滤，可能导致该矩形框比实际范围大，从而可能使得不是图注图元的内容被误认为是图注图元，造成逻辑错误。

[0051] 在上述技术方案中，优选地，还包括：为所述复合图区块中的每种元素添加对应的逻辑标签，以标定其所属分类；以及所述步骤 208 还包括：存储所述逻辑标签和 / 或所述逻辑标签对应的图元的标识。

[0052] 在该技术方案中，通过对每种元素添加逻辑标签，便于对各种元素对应的逻辑关系进行查看和管理，从而实现对版式文档进行流式转换后，得到更好的显示效果。

[0053] 在上述技术方案中，优选地，所述步骤 206 包括：若所述复合图区块中仅包含一个正文插图复合图，则选取与该正文插图复合图的距离小于预设距离的图注图元，以作为与该正文插图复合图相关联的图注图元；若所述复合图区块中包含多个正文插图复合图，则将所述多个正文插图复合图和解析出的所有图注图元分别作为二分图的顶点，以利用所述二分图判断所述多个正文插图复合图与所述图注图元之间的关联关系。

[0054] 在该技术方案中，通过距离靠近原则和二分法最优匹配法，实现对正文插图复合图和图注图元的关联识别，有利于得到更为合理、准确的逻辑关系，以便基于该逻辑关系实现版式文档的流式重排。

[0055] 在上述技术方案中，优选地，所述步骤 204 包括：读取预设的所述图注图元的排版特征信息；根据读取的所述排版特征信息，在解析出的所述文字图元中进行检索，以获取所述图注图元。

[0056] 在该技术方案中，图注图元具有的排版特征信息，比如字体、以“图”等关键字起

始、居中、字数限制、与其他文字之间的位置关系等，通过这些特征信息，既可以找到对应内容的图元，又可以避免将如正文中的“图 1”作为图注图元(具体为图标题，或者也可以为解释性文字等)，实现对图注图元的准确获取。

[0057] 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逻辑处理的具体流程图。

[0058] 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逻辑处理的具体流程包括：

[0059] 步骤 302，对版式文档中的复合图进行分割，具体地，分割出来的复合图中可能包含有插图复合图，还可能包含装饰性复合图、分栏线等其他的元素。

[0060] 在完成分割后，可以将分割出来的复合图中所有图元的 ID 进行存储，比如存储在 XML 文件中，以便在对该复合图进行调用或处理时，根据存储的图元 ID 查找到该复合图。

[0061] 实际上，上述对复合图的分割过程，仅是从物理结构上，将对应于复合图的区块从版式文档中分割出来，但并不包含对其逻辑结构上的分析，因此，在正常的版式文档结构下进行分割时，往往是根据图像与文字间的距离等物理特性进行关联的，从而会导致分割出来的复合图中不包含图注。

[0062] 在下面的步骤中，将会完成复合图与“遗留”在版式文档中的图注进行准确地关联等，从而实现对复合图的逻辑处理。

[0063] 步骤 304 至步骤 308 是对复合图的处理：

[0064] 步骤 304，获取复合图的特征属性信息。具体地，涉及提取复合图在页面空间的布局、样式信息和内容图像的纹理等特征，具体的主要特征如表 1 所示：

[0065]

1	Height	复合图的高度
2	Length	复合图的宽度
3	Area	复合图的面积
4	Eccentricity	复合图的离心率
5	BlkPix	黑色像素在复合图中所占百分比
6	Std	复合图各像素灰度值的标准方差
7	Entropy	图像熵度量图像信息量
8	Contrast	惯性矩反映图像纹理清晰程度
9	Correlation	相关性衡量某一方向的纹理相关性
10	Energy	角二阶矩度量灰度分布均匀性
11	Homogeneity	逆差矩反映图像局部均匀性

[0066] 表 1

[0067] 同时，根据实际复合图纹理的特点，选取距离和方向，计算出灰度共生矩阵及特征系数，将特征系数组成纹理特征矢量，作为统计分类器的输入。

[0068] 步骤 306，对复合图中包含的元素进行分类。具体地，可以使用 SVM（Support Vector Machine，支持向量机）为分类器，选择 RBF（Radial Basis Function，径向基核函数），对分割出来的复合图中包含的插图复合图、图形商标、公式、分栏线、页眉、页脚、装饰性复合图等等各种元素进行分类，根据分类结果对每个元素进行标定，以得到其在版面中的逻辑标签。

[0069] 步骤 308，过滤干扰元素，保留插图复合图。具体地，是指过滤图形商标、公式、分栏线、页眉、页脚、装饰性复合图对象，这些复合图的大量存在，影响正文中的插图复合图和图注的关联。

[0070] 步骤 310 和步骤 312 是对文字图元的处理：

[0071] 步骤 310，对版式文档进行解析，得到解析出来的文字图元。

[0072] 步骤 312，提取文字图元中的图注图元。具体地，可以根据图注图元的文字特征属性，将其与正文文字等区别开来，比如以图标题为例，其字体小于正文主要字体，以关键字起始，如“图 /Figure/Fig”、“图 /Figure/Fig1”、“图 /Figure/Fig1-1”等等，可用正则表达式来表示。

[0073] 同时，所提取的图标题也可能是该图在正文中的引用，可以根据图注文本的排版特点，比如居中设置、每段的字数限制等等，从而过滤待选图标题在正文中的引用。

[0074] 步骤 314，判断当前复合图区块中的插图复合图的数量，若为单个，则进入步骤 316，否则进入步骤 318。

[0075] 步骤 316，根据距离选择与插图复合图相关联的图注图元。具体地，以图标题为例，当页面含有单个插图复合图和单个（或多个）图标题时，即 1 对 1（或 1 对多）的模式，采用距离靠近原则，选取距离插图复合图最近的图标题为其标题。

[0076] 步骤 318，利用二分图的方法选择与插图复合图相关联的图注图元。具体地，当页面上含有多个插图复合图和多个图标题时，不能单靠图标题的距离和样式，采用二分图的方法，将插图复合图和图标题分别表示为二分图的顶点，根据图标题和插图复合图的距离定义顶点间的关联权值，然后通过查找二分图的最大权匹配，寻找最可能的插图复合图和图标题的关联方案，取得全局上的关联匹配最优。

[0077] 步骤 320，保存插图复合图和图标题的关联关系。此外，还可以保存步骤 306 中的分类结果得到复合图中的各个元素在版面中的逻辑标签，以及每个逻辑类别所对应的元素的图元 ID 集合。具体地，可以存储为 XML 的形式。

[0078] 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。

[0079] 图 4A 和图 4B 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逻辑处理的示意图。

[0080] 如图所示，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，经过对该图的分割处理，从中提取出复合图区块包括插图复合图 402A、分栏线复合图 402B 和装饰性复合图 402C。可以将复合图区块中的所有图元 ID 存储在 XML 文件中，以便于对该复合图区块的处理。下面将按照图 3 给出的流程对页面中的复合图对象进行逻辑处理。

[0081] 首先，通过解析引擎获取版式文档的各种图元后，对文档进行版面分析，将版面分析中复合图区块的分割结果从 XML 文件中读取，包括读入其外接矩形框和组合该复合图的图元 ID 集合。具体地，将外接矩形框绘制在页面图的效果如图 4A 所示。

[0082] 然后提取页面内所有复合图的布局、样式信息和内容图像的纹理等特征属性信息，具体地，主要的特征属性信息如表 1 所示。将特征属性信息作为已经训练好的统计分类器 SVM 的输入，对该页面内的 5 个复合图进行分类，并根据分类结果进行逻辑标签的标定。具体地，分类结果如图 4B 所示，该页面包含三类复合图逻辑标签，其中，正文中 2 个插图复合图 402A、2 个分栏线复合图 402B 和左边页边的 1 个装饰性复合图 402C。正文页面下方的插图复合图 402A 和页面左边的装饰性复合图 402C，包括文字图元和大量的路径操作，不仅分割难度大，且识别率低，但采用本发明的方法，该页面的复合图皆被准确的标注了逻辑类别标签。逻辑标定结果可直接用于版式文档的流式重排应用。

[0083] 在上述实施例中，主要描述了对于版式文档中分割出来的复合图的逻辑标签进行标定的过程，下面通过另一个实施例来说明将复合图与图注进行关联的方案。

[0084] 图 5A 和图 5B 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逻辑处理的示意图。

[0085] 如图所示，以英文版式文档论文“TOASTER and KROONDE:High-Resolution and High-Speed Real-time Sensor Interfaces”中的一张双栏页面为例经过对该图的分割处理，从中提取出复合图区块包括插图复合图 502A1、插图复合图 502A2、插图复合图 502A3、插图复合图 502A4 和分栏线复合图 502B。可以将复合图区块中的所有图元 ID 存储在 XML 文件中，以便于对该复合图区块的处理。下面将按照图 3 给出的流程对页面中的复合图对象进行逻辑处理。

[0086] 首先，通过解析引擎获取版式文档的各种图元后，对文档进行版面分析，将版面分析中复合图区块的分割结果从 XML 文件中读取，包括读入其外接矩形框和组合该复合图的图元 ID 集合。具体地，将外接矩形框绘制在页面图的效果如图 5A 所示。

[0087] 然后，对复合图区块进行处理。具体地，对复合图区块包含的所有元素进行类别分析，并根据分析结果保留插图复合图，而将页面内图形商标、公式、分栏线、页眉、页脚、装饰性复合图过滤，这些复合图的存在，影响正文中的插图复合图和图标题及图注的关联和识别。

[0088] 同时，还包括对图注信息的获取，这里以图标题的获取为例。从解析后的版式文档文字元素中，可以根据图标题的文字特征属性(如在该页面中以关键字 Figure 起始)和排版特征属性(如居中设置)，提取关于图标题的信息，并且过滤待选图标题在正文中的引用。具体地，分析得到如图所示的图标题 504A、图标题 504B、图标题 504C 等。

[0089] 最后，对插图复合图和图标题进行关联设置。具体地，由于该页面中包含多个插图复合图，因而采用二分图的方法，将插图复合图和图标题分别表示为二分图的顶点，根据图标题和图的距离定义顶点间的关联权值，查找二分图的最大权匹配，寻找最可能的图表和其标题的关联。该页面的输入有 6 个复合图，如图 5B 所示，页面右下方的分栏线复合图被过滤，左栏的中间的 2 个插图复合图合并后，页面的 4 个插图复合图和 4 个图标题得到关联。该结果可直接用于版式文档的流式重排应用。

[0090] 以上结合附图详细说明了本发明的技术方案，本发明通过对版式文档(如 PDF 文

档）内嵌的元数据信息进行解析和分析，在分割页面所包含的复合图后，对页面内所有的复合图，提取其页面空间的布局、样式信息和内容图像的纹理等特征，作为 SVM 分类器的输入，依据分类的类型对复合图进行逻辑标注。同时，从解析后的版式文档文字元素中，提取待选图标题，采用距离靠近原则和二分法最优匹配法对插图复合图和其图标题进行关联识别。保证版式文档中的图像转化为流式文档后，图注能和图像保持同步即保持相连，从而最终实现版式固定文档按阅读顺序重排成连贯的流式文档。

[0091] 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。

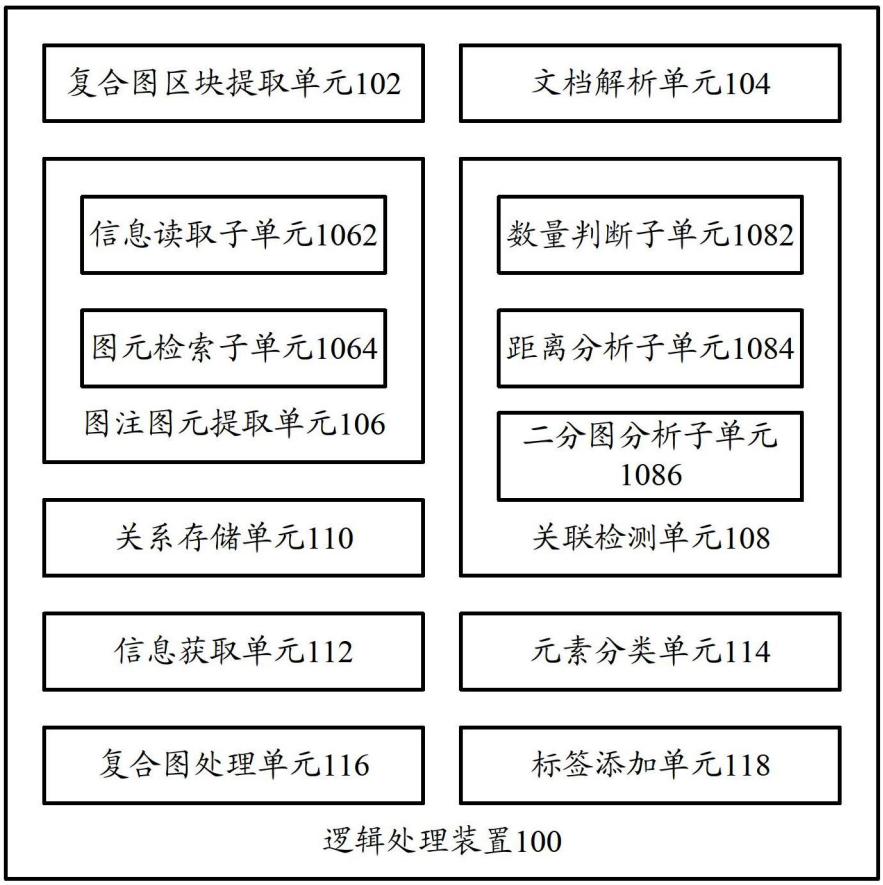

复合图区块提取单元102
文档解析单元104
信息读取子单元1062
图元检索子单元1064
图注图元提取单元106
数量判断子单元1082
距离分析子单元1084
二分图分析子单元1086
关联检测单元108
关系存储单元110
信息获取单元112
元素分类单元114
复合图处理单元116
标签添加单元118
逻辑处理装置100

图 1

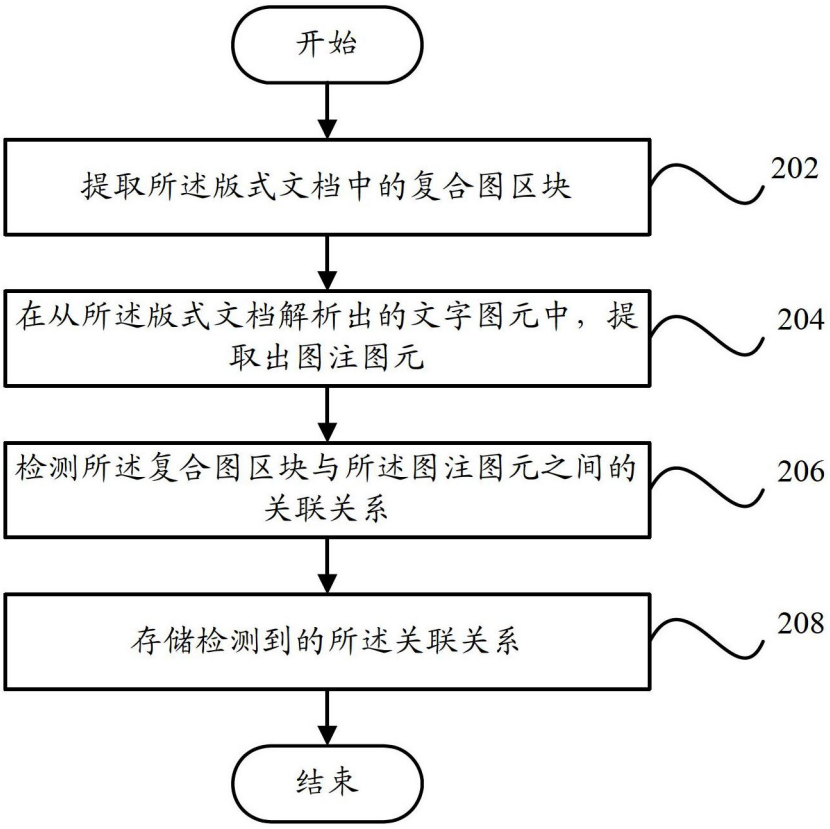

开始
提取所述版式文档中的复合图区块
202
在从所述版式文档解析出的文字图元中，提取出图注图元
204
检测所述复合图区块与所述图注图元之间的关联关系
206
存储检测到的所述关联关系
208
结束

图 2

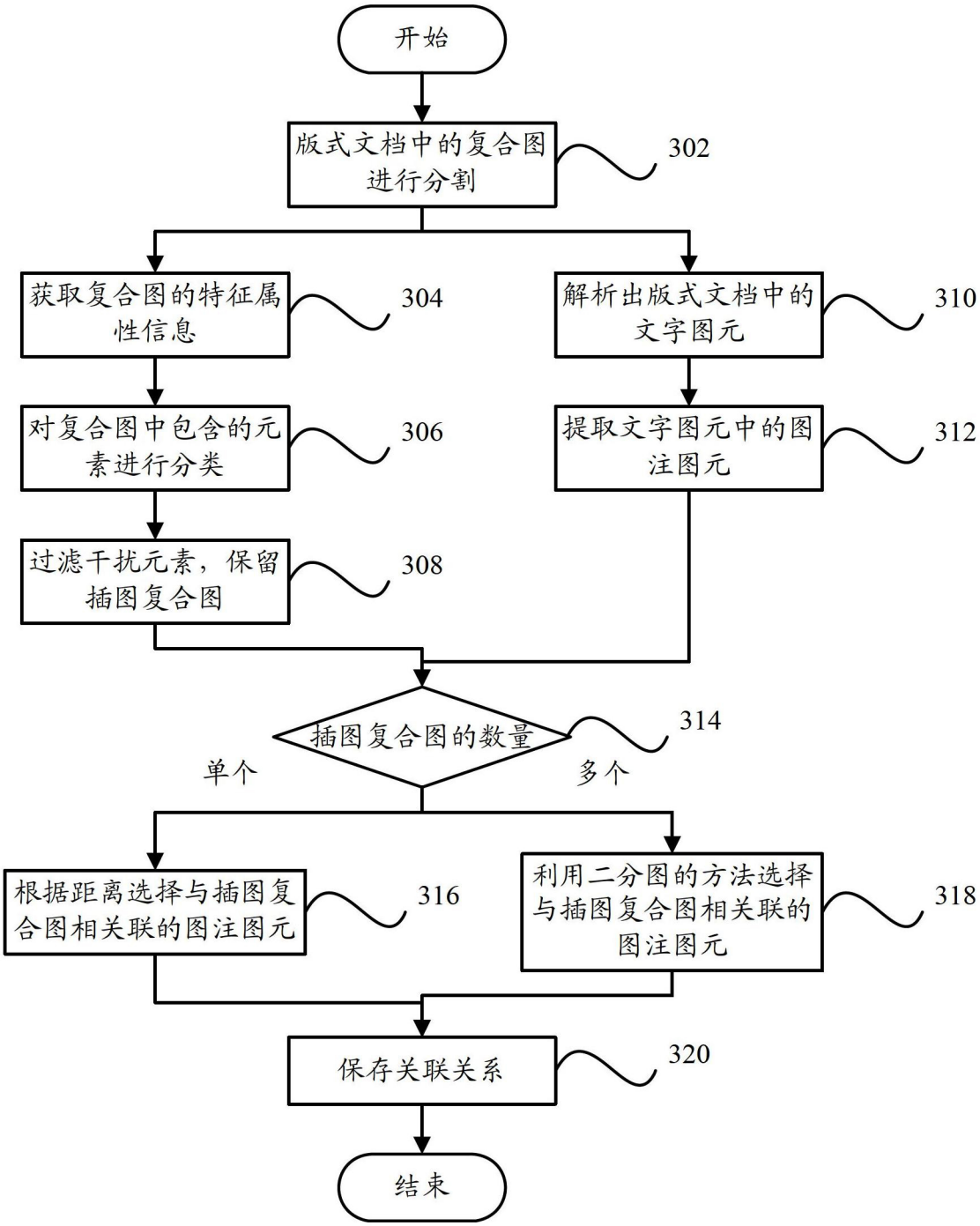

开始
版式文档中的复合图进行分割
302
获取复合图的特征属性信息
304
对复合图中包含的元素进行分类
306
过滤干扰元素，保留插图复合图
308
解析出版式文档中的文字图元
310
提取文字图元中的图注图元
312
插图复合图的数量
314
单个
多个
根据距离选择与插图复合图相关联的图注图元
316
利用二分图的方法选择与插图复合图相关联的图注图元
318
保存关联关系
320
结束

图 3

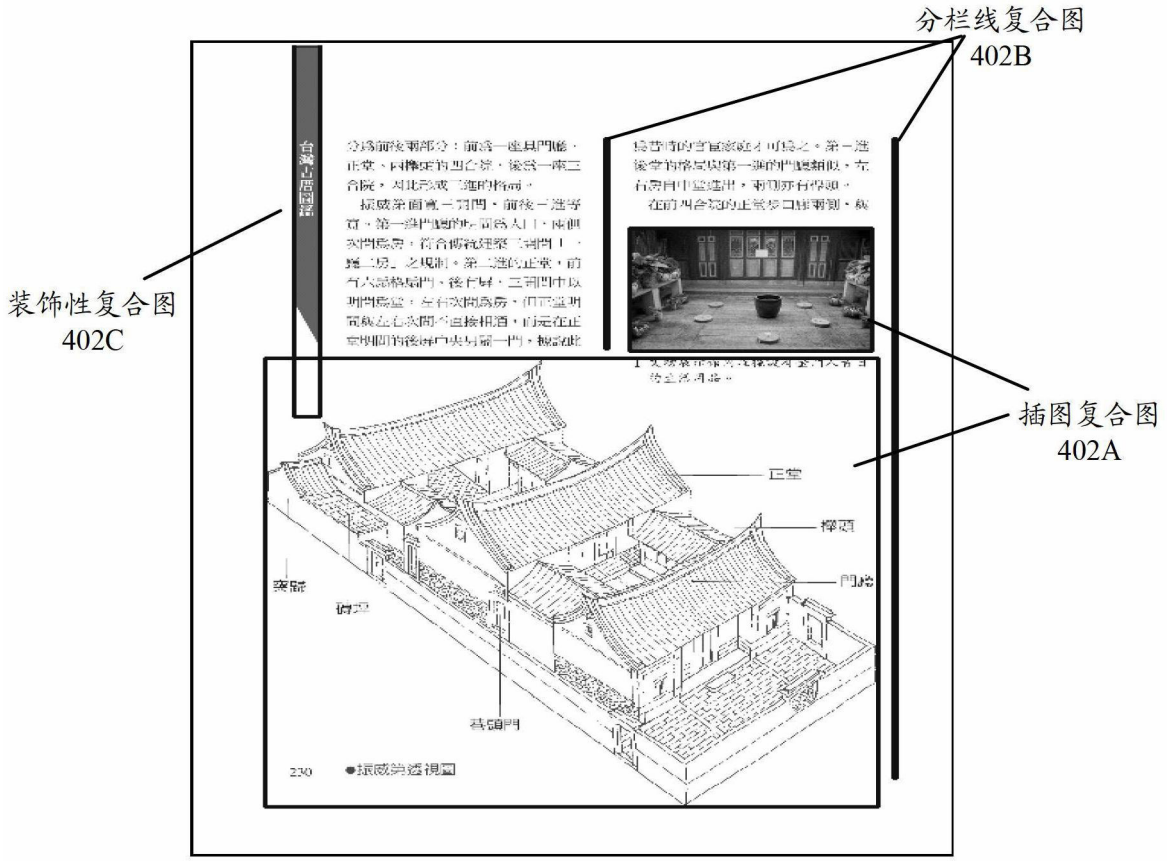

分栏线复合图
402B
装饰性复合图
402C
插图复合图
402A
台灣古厝圖說
正堂
護龍
門廊
230
●振威第透視圖

图 4A

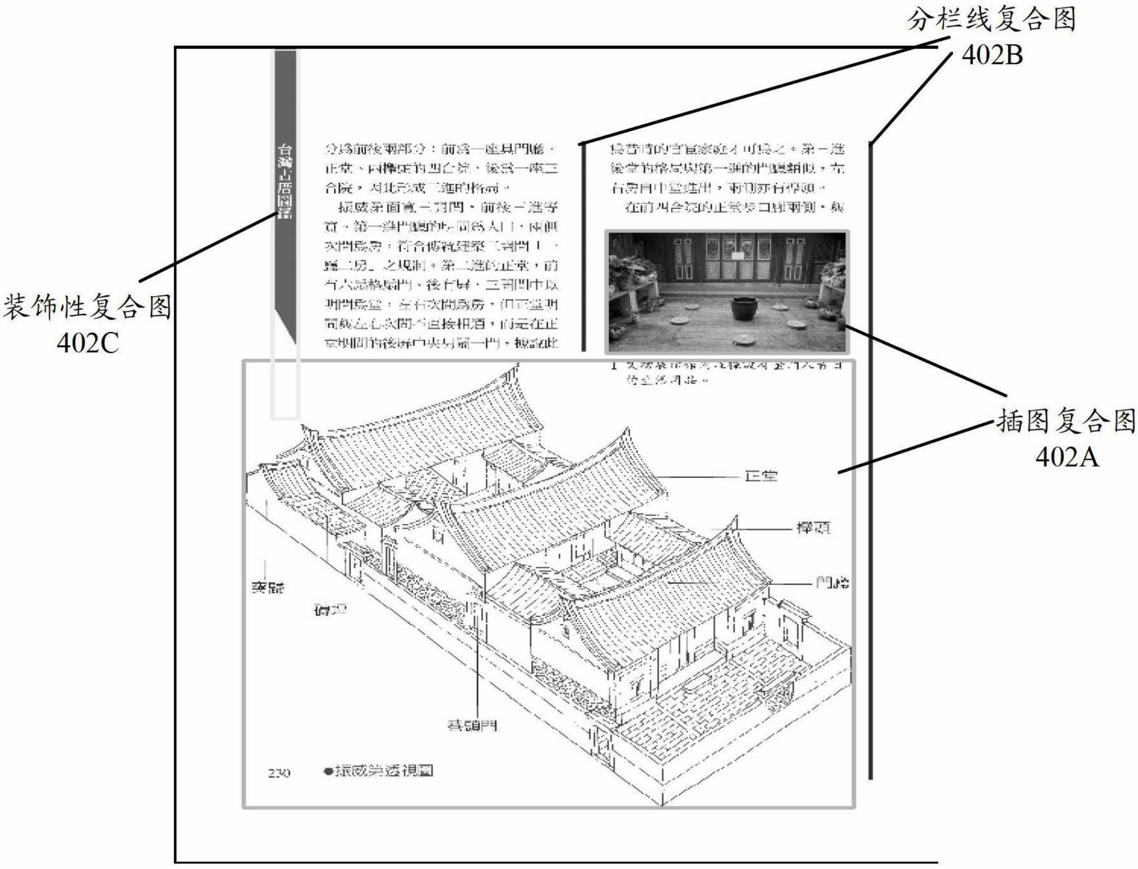


图 4B

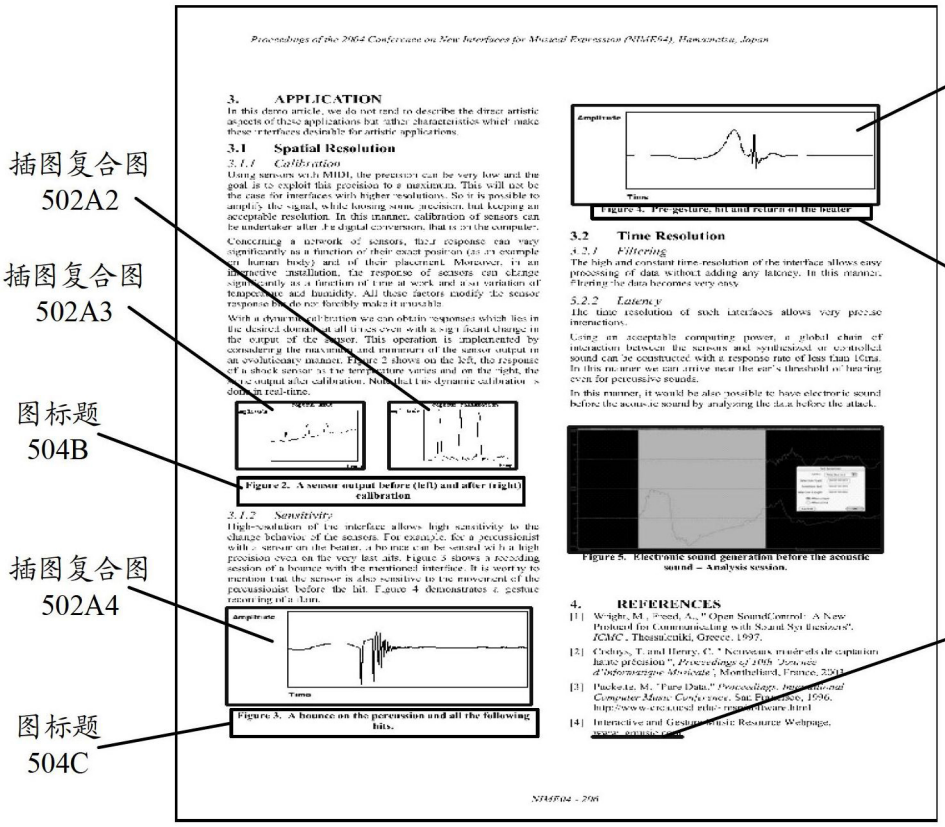

插图复合图
502A1
插图复合图
502A2
插图复合图
502A3
图标题
504B
插图复合图
502A4
图标题
504C
图标题
504A
分栏线复合图
502B

图 5A

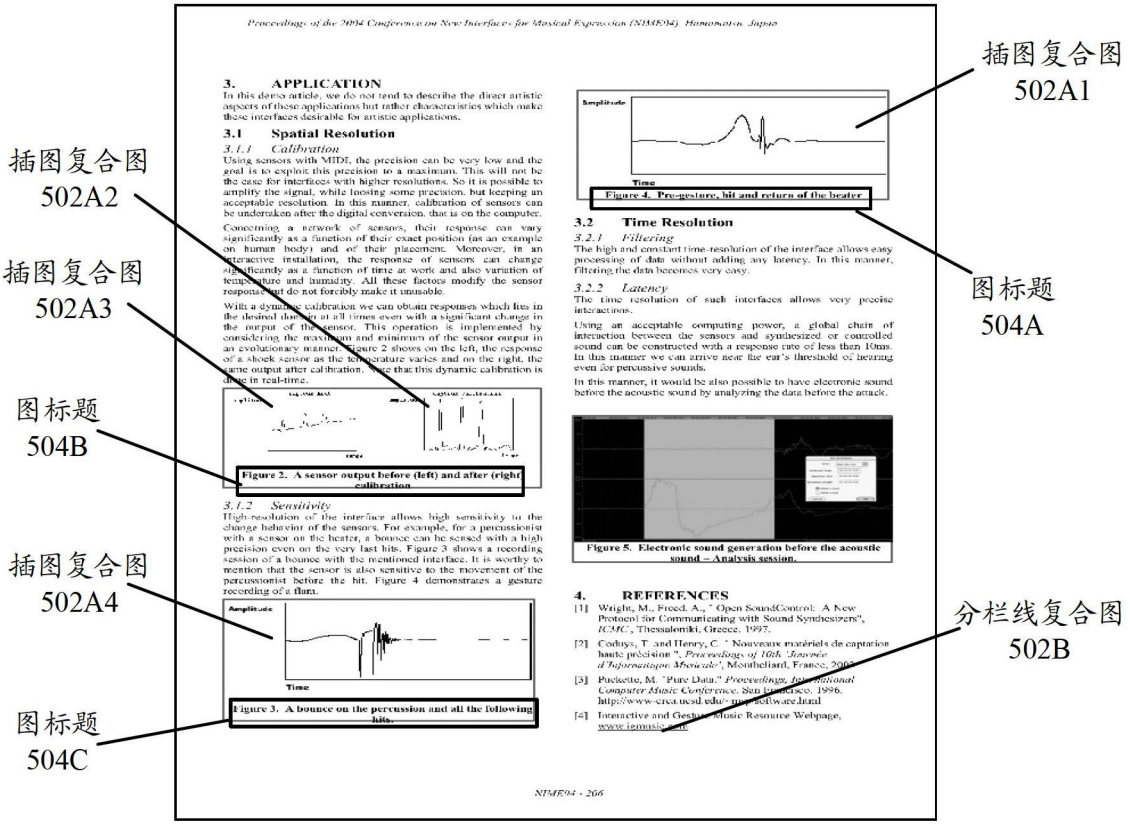

插图复合图
502A1
插图复合图
502A2
插图复合图
502A3
图标题
504B
插图复合图
502A4
图标题
504C
图标题
504A
分栏线复合图
502B
3. APPLICATION
3.1 Spatial Resolution
3.1.1 Calibration
Figure 2. A sensor output before (left) and after (right) calibration.
3.1.2 Sensitivity
Figure 3. A bounce on the percussion and all the following hits.
Figure 4. Pre-gesture, hit and return of the beater
3.2 Time Resolution
3.2.1 Filtering
3.2.2 Latency
Figure 5. Electronic sound generation before the acoustic sound – Analysis session.
4. REFERENCES
NIME04 - 266

图 5B