(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114895785 A
(43)申请公布日 2022.08.12

(21)申请号 202210529511.7

(22)申请日 2018.06.15

(30)优先权数据
17176248.7 2017.06.15 EP
62/519,952 2017.06.15 US
62/680,678 2018.06.05 US

(62)分案原申请数据
201880011958.7 2018.06.15

(71)申请人 杜比国际公司
地址 荷兰阿姆斯特丹
申请人 杜比实验室特许公司

(72)发明人 C·费尔施 N·R·廷哥斯

(74)专利代理机构 北京市汉坤律师事务所 11602
专利代理师 魏小薇 吴丽丽

(51)Int.Cl.
G06F 3/01(2006.01)
G09G 5/36(2006.01)
H04L 67/131(2022.01)
H04S 3/00(2006.01)
H04S 7/00(2006.01)

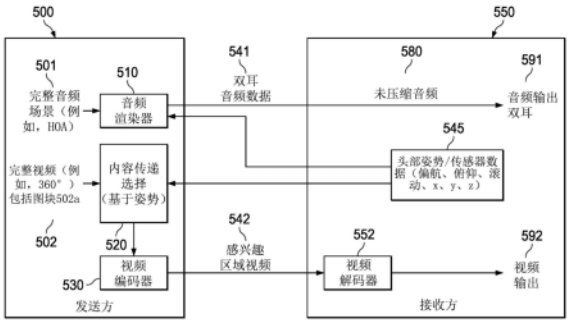
权利要求书4页 说明书25页 附图10页

(54)发明名称

一种包括再现和存储媒体内容的装置的系统及其相关装置

(57)摘要

本公开涉及一种包括再现和存储媒体内容的装置的系统及其相关装置。所述系统包括用于再现媒体内容的第一装置和用于存储所述媒体内容的第二装置，其中，所述第一装置适用于：获得指示用户的位置和/或取向的姿势信息；以及将所述姿势信息传输到所述第二装置。所述第二装置适用于：基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容；以及将所述经渲染的媒体内容传输到所述第一装置以进行再现。



CN 114895785 A

1.一种系统,包括用于再现媒体内容的第一装置和用于存储所述媒体内容的第二装置,

其中,所述第一装置适用于:

-获得指示用户的位置和/或取向的姿势信息;以及

-将所述姿势信息传输到所述第二装置;并且

所述第二装置适用于:

-基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容;以及

-将所述经渲染的媒体内容传输到所述第一装置以进行再现。

2.根据权利要求1所述的系统,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容;和/或

所述媒体内容包括视频内容,并且所述经渲染的媒体内容包括经渲染的视频内容。

3.根据权利要求1所述的系统,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容;并且

所述第一装置进一步适用于生成所述经渲染的音频内容的听觉表示。

4.根据权利要求2所述的系统,其中,所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种,或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

5.根据权利要求2所述的系统,其中,所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种,或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

6.根据权利要求1所述的系统,其中,所述第二装置进一步适用于:

基于所述姿势信息和先前姿势信息获得预测的姿势信息;以及

基于所述预测的姿势信息来渲染所述媒体内容以获得所述经渲染的媒体内容。

7.根据权利要求6所述的系统,其中,所述第二装置进一步适用于:

将所述预测的姿势信息与所述经渲染的媒体内容一起传输到所述第一装置。

8.根据权利要求7所述的系统,其中,所述第一装置进一步适用于:

将所述预测的姿势信息与实际姿势信息进行比较;以及

基于所述比较的结果更新所述经渲染的媒体内容。

9.一种用于进行定时估计的第二装置,所述第二装置适用于:

对编码和解码媒体内容所需的时间进行估计;

对将所述媒体内容传输到第一装置所需的时间进行估计;以及

基于对编码和解码所述媒体内容所需的时间的估计以及对将所述媒体内容传输到所述第一装置所需的时间的估计对预期要由所述第一装置处理所述媒体内容的定时进行估计。

10.根据权利要求8所述的系统,其中,所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计;并且

所述实际姿势信息是在所述第一装置实际处理所述经渲染的媒体内容以进行再现的定时获得的姿势信息。

11.根据权利要求1所述的系统，其中，以未压缩的形式将所述经渲染的媒体内容传输到所述第一装置。

12.根据权利要求1所述的系统，其中，所述第二装置进一步适用于在向所述第一装置传输之前对所述经渲染的媒体内容进行编码；以及所述第一装置进一步适用于在所述第一装置处接收到经编码的所述经渲染的媒体内容之后，对经编码的所述经渲染的媒体内容进行解码。

13.根据权利要求10所述的系统，其中，对预期要由所述第一装置处理所述经渲染的音频内容以进行再现的定时的所述估计包括：对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

14.根据权利要求6所述的系统，其中，所述预测的姿势信息是进一步基于对编码和解码所述经渲染的媒体内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计来获得的。

15.根据权利要求1所述的系统，其中，所述第一装置进一步适用于：

将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较；以及

基于所述比较的结果更新所述经渲染的媒体内容。

16.根据权利要求1所述的系统，其中，所述第二装置进一步适用于：

确定指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息；以及

将所述梯度信息与所述经渲染的媒体内容一起传输到所述第一装置；以及

所述第一装置进一步适用于：

将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较；以及

基于所述梯度信息和所述比较的结果更新所述经渲染的媒体内容。

17.根据权利要求1所述的系统，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容；

所述第一装置进一步适用于将指示所述第一装置所处的环境的声学特性的环境信息传输到所述第二装置；并且

所述渲染所述媒体内容进一步基于所述环境信息。

18.根据权利要求1所述的系统，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容；

所述第一装置进一步适用于将指示所述用户或所述用户的一部分的形貌的形貌信息传输到所述第二装置；并且

所述渲染所述媒体内容进一步基于所述形貌信息。

19.一种用于提供媒体内容以供第一装置再现的第二装置，所述第二装置适用于：

接收指示所述第一装置的用户的位置和/或取向的姿势信息；

基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容；以及

将所述经渲染的媒体内容传输到所述第一装置以进行再现。

20.根据权利要求19所述的第二装置，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容；和/或

所述媒体内容包括视频内容，并且所述经渲染的媒体内容包括经渲染的视频内容。

21.根据权利要求20所述的第二装置,其中,所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种,或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

22.根据权利要求20所述的第二装置,其中,所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种,或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

23.根据权利要求19所述的第二装置,进一步适用于:

基于所述姿势信息和先前姿势信息获得预测的姿势信息;以及

基于所述预测的姿势信息来渲染所述媒体内容以获得所述经渲染的媒体内容。

24.根据权利要求23所述的第二装置,进一步适用于:

将所述预测的姿势信息与所述经渲染的媒体内容一起传输到所述第一装置。

25.根据权利要求23所述的第二装置,其中,所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计。

26.根据权利要求19所述的第二装置,其中,以未压缩的形式将所述经渲染的媒体内容传输到所述第一装置。

27.根据权利要求19所述的第二装置,进一步适用于在向所述第一装置传输之前对所述经渲染的媒体内容进行编码。

28.根据权利要求25所述的第二装置,其中,对预期要由所述第一装置处理所述经渲染的音频内容以进行再现的定时的所述估计包括:对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

29.根据权利要求23所述的第二装置,其中,所述预测的姿势信息是进一步基于对编码和解码所述经渲染的媒体内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计来获得的。

30.根据权利要求19所述的第二装置,进一步适用于:

确定指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息;以及

将所述梯度信息与所述经渲染的媒体内容一起传输到所述第一装置。

31.根据权利要求19所述的第二装置,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容;

所述第二装置进一步适用于从所述第一装置接收指示所述第一装置所处的环境的声学特性的环境信息;并且

所述渲染所述媒体内容进一步基于所述环境信息。

32.根据权利要求19所述的第二装置,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容;

所述第二装置进一步适用于从所述第一装置接收指示所述用户或所述用户的一部分的形貌的形貌信息;并且

所述渲染所述媒体内容进一步基于所述形貌信息。

33.一种用于再现由第二装置提供的媒体内容的第一装置,所述第一装置适用于:

获得指示所述第一装置的用户的位置和/或取向的姿势信息；

将所述姿势信息传输到所述第二装置；

从所述第二装置接收经渲染的媒体内容，其中，所述经渲染的媒体内容已经通过基于所述姿势信息来渲染所述媒体内容而获得；以及

再现所述经渲染的媒体内容。

一种包括再现和存储媒体内容的装置的系统及其相关装置

[0001] 分案声明

[0002] 本申请是申请日为2018年6月15日、申请号为201880011958.7、发明名称为“在计算机介导的现实应用中优化发送方与接收方之间的通信的方法、装置和系统”的发明专利申请的分案申请。

[0003] 相关申请的交叉引用

[0004] 本申请要求2018年6月5日提交的美国临时申请号62/680,678、均于2017年6月15日提交的美国临时申请号62/519,952和欧洲专利申请号17176248.7的优先权,所有这些申请都通过引用以其全文并入本文。

技术领域

[0005] 本公开涉及计算机介导的现实应用,例如虚拟现实(VR)应用、增强现实(AR)应用和混合现实(MR)应用。这些应用可以包括但不限于客户端/接收方双耳化和非双耳化的音频应用和视频应用。

背景技术

[0006] 计算机介导的现实空间(例如,VR、AR和MR空间)中的应用和产品正在快速发展而包括声源和场景的越来越精细的声学模型。并非刻意进行限制,将在本文档的其余部分中提及VR、AR和MR。为了优化计算机介导的现实体验,优选的是使用户移动(例如,头部移动)与对适应于该移动的声音(经渲染的声音)的感知之间的延迟最小化。这种延迟也被称为运动到声音时延或运动到耳朵迟延。另外,还希望最小化对公共接收方设备(如智能电话)的声音进行解码和渲染所需的指令的数量,其中,重要的是优化计算复杂性和功耗。例如对于非通信情况,当传输整个音频场景时,重点在于接收方的渲染时延。例如,线性应用(例如,电影)不会动态地对用户的动作作出反应。然而,对于交互式内容,将必须解决所有累积的往返时延(例如,如果用户触发需要发送回服务器进行渲染的事件)。在消耗内容之前,应该以足够的前置时间对动态变化的内容进行编码,使得用户不会识别到运动与运动产生的效果之间的时延,并且内容的音频与视频之间不存在未校准。在线性应用的情况下,对于运动到声音时延,不考虑编码和解码时延,因为用户移动(位置和/或取向)不影响内容本身。相反,这些移动仅影响观看内容时的视角。因此,对于线性内容,用户移动仅影响渲染,而不影响对输出声音的编码和/或解码。通信情况却不同,因为系统只能在内容(例如,语音)发生时才开始编码、传输和解码媒体。这同样适用于交互式内容(例如,来自游戏引擎)由云中的远程服务器实时渲染和编码的情况。另外,视频和音频系统的整体时延是相同的是非常重要的,因为差异可能导致晕动病。因此,取决于视频系统的时延,需要实现相似水平的音频系统的时延。

发明内容

[0007] 本文档解决了常见AR、VR和MR系统需要太高时延和太高计算复杂性要求来提供引

人注目的体验的技术问题。为了解决该问题，本文档提出了一种处理媒体内容的方法、一种用于处理媒体内容的系统以及相应的装置，所述方法、系统和装置具有相应的独立权利要求的特征。

[0008] 本公开的一方面涉及一种处理媒体内容以供第一装置再现的方法。例如，第一装置可以是接收方、接收方装置或重放装置中的一种。例如，所述第一装置可以对应于、包括AR/VR/MR设备(例如，AR/VR/MR头戴式设备)或结合AR/VR/MR设备进行操作。如此，第一装置可以包括用于再现媒体内容的再现设备(例如，扬声器、头戴式耳机)和耦接到再现设备的处理器。媒体内容可以是或包括音频内容和/或视频内容。处理可以涉及或对应于渲染。再现可以涉及或对应于重放。所述方法可以包括获得指示用户的位置和/或取向的姿势信息。获得姿势信息可以在第一装置处执行。用户可以是第一装置的用户。例如，姿势信息可以与用户的头部有关。所述姿势信息可以通过可被布置成与用户配准的传感器(例如，姿势传感器)获得。因此，姿势信息可以被称为传感器数据。姿势信息可以进一步包括姿势的一个或多个一阶导数和/或姿势的一个或多个二阶导数。例如，用户可以佩戴AR/VR/MR设备。所述方法可以进一步包括将姿势信息传输到提供(例如，存储、中继)媒体内容的第二装置。例如，第二装置可以是发送方装置、服务器装置或内容传递装置中的一种。第二装置可以是用于向第一装置提供媒体内容的装置。第一装置和第二装置可以在空间上彼此分开。所述方法可以进一步包括基于姿势信息来渲染媒体内容以获得经渲染的媒体内容。渲染可以在第二装置处执行。经渲染的媒体内容可以称为预渲染的媒体内容。例如，在音频内容的情况下，渲染可以是渲染到两个或更多个声道。所述方法可以进一步包括将经渲染的媒体内容传输到第一装置以进行再现。传输经渲染的媒体内容可以由第二装置执行。所述方法还可以进一步包括(通过第一装置)再现(例如，重放)经渲染的媒体内容。

[0009] 如果仅需要传输经渲染的媒体内容，则用于无损数据传输的传输比特率可以与完整媒体内容的压缩版本的比特率类似或相当。因此，在所提出的方法的上下文中可能不需要压缩。传输未压缩或无损媒体流将消除或减少因编码和解码而导致的时延。例如，由编码/解码产生的时延可以减少到零，这将导致运动到耳朵时延和/或运动到眼睛时延的整体减少。此外，当不对预渲染的媒体内容进行压缩时，第一装置(接收方)可以输出音频/视频而无需进行解码或渲染。这将导致接收方处的计算复杂性降低，因为不需要执行解码和/或在发送方侧已经完成了渲染。因此，所提出的方法允许减少运动到耳朵时延和/或运动到眼睛时延，并且进一步允许降低接收方侧的计算复杂性。

[0010] 在一些实施例中，媒体内容可以包括音频内容，并且经渲染的媒体内容可以包括经渲染的音频内容。可替代地或另外地，媒体内容可以包括视频内容，并且经渲染的媒体内容可以包括经渲染的视频内容。

[0011] 在一些实施例中，媒体内容可以包括音频内容，并且经渲染的媒体内容可以包括经渲染的音频内容。然后，所述方法可以进一步包括生成经渲染的音频内容的听觉(例如，声学)表示。生成听觉表示可以在第一装置处执行。例如，在音频内容的情况下，可以经由第一装置的两个或更多个扬声器来执行所述生成。

[0012] 在一些实施例中，音频内容可以是基于一阶高保真度立体声响复制(FOA)的音频内容、基于更高阶高保真度立体声响复制(HOA)的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种，或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音

频内容、或基于声道的音频内容中的两种或更多种的组合。

[0013] 在一些实施例中，经渲染的音频内容可以是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种，或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0014] 在一些实施例中，渲染可以涉及基于姿势信息并且进一步基于先前姿势信息和/或一个或多个一阶导数和/或二阶导数来获得预测的姿势信息。预测的姿势信息可以是未来定时的姿势信息。先前姿势信息可以是在先前定时已经在第一装置处获得或从第一装置接收的姿势信息。预测可以在第二装置处执行。可替代地，预测可以在第一装置处执行。在后一种情况下，第一装置可以将预测的姿势信息传输到第二装置。渲染可以进一步涉及基于预测的姿势信息来渲染媒体内容以获得经渲染的媒体内容。

[0015] 通过考虑预测的姿势信息，可以解决可能由对经渲染的媒体内容进行编码/解码和/或将经渲染的媒体内容传输到第一装置而导致的延迟。换言之，对于适当预测的姿势信息，可以隐藏所述延迟，使得用户不会意识到该延迟并且可能不会察觉到音频、视频与移动之间的任何不匹配。

[0016] 在一些实施例中，所述方法可以进一步包括将预测的姿势信息与经渲染的媒体内容一起传输到第一装置。

[0017] 这使得第一装置能够检查预测的姿势信息(即，在这种情况下是已经用于渲染媒体内容的姿势信息)是否与实际/当前姿势信息(即，当前在第一装置处获得的姿势信息)相同(或基本相同)，并且如果在预测的姿势信息与实际/当前姿势信息之间存在不匹配，则适当地适配经渲染的媒体内容。

[0018] 在一些实施例中，所述方法可以进一步包括将预测的姿势信息与实际姿势信息进行比较。所述方法还可以进一步包括基于比较的结果更新经渲染的媒体内容。所述比较和所述更新可以在第一装置处执行。例如，实际姿势信息可以是在第一装置再现经渲染的媒体内容的定时(例如，在所述定时获得)的姿势信息。更新可以例如基于预测的姿势信息与实际姿势信息之间的差异来执行。所述更新可以涉及例如通过旋转、级别改变和/或盲上混(blind upmixing)对经渲染的媒体内容进行外推。

[0019] 在一些实施例中，预测的姿势信息可以被预测来对预期要由第一装置处理经渲染的媒体内容以进行再现的定时进行估计。第一装置对经渲染的媒体内容的处理可以涉及再现(例如，重放)经渲染的媒体内容。实际姿势信息(例如，当前姿势信息)可以是在第一装置实际处理经渲染的媒体内容以进行再现的定时获得的姿势信息。实际姿势信息可以在第一装置实际处理经渲染的媒体内容的定时获得。

[0020] 因此，可以解决预测的姿势信息与实际姿势信息之间的任何不匹配，从而更好地使经渲染的媒体内容适应用户的姿势(例如，用户头部的姿势)并且避免对于用户而言所感知的音频/视频场景与预期的音频/视频场景之间的任何差异。由于预期预测的姿势信息与实际姿势信息之间的不匹配将很小，因此可以以可管理的计算复杂性将这种适配安全地委托给第一装置。

[0021] 在一些实施例中，可以以未压缩的形式将经渲染的媒体内容传输到第一装置。

[0022] 这使得能够降低第一装置(接收方)处的计算复杂性，并且还减少了姿势的变化与已经根据变化的姿势渲染的媒体内容的再现之间的往返延迟。

[0023] 在一些实施例中，所述方法可以进一步包括在传输到第一装置之前对经渲染的媒体内容进行编码(例如，压缩)。所述方法还可以进一步包括在第一装置处接收之后对经编码的经渲染媒体内容进行解码(例如，解压缩)。编码/解码可以涉及或对应于对经渲染的媒体内容进行压缩/解压缩。编码/解码可以是低延迟编码/解码。

[0024] 在一些实施例中，对预期要由第一装置处理经渲染的媒体内容以进行再现的定时的估计可以包括：对编码和解码经渲染的音频内容所需的时间的估计和/或对将经渲染的媒体内容传输到第一装置所需的时间的估计。

[0025] 在一些实施例中，预测的姿势信息可以进一步基于对编码和解码经渲染的媒体内容所需的时间的估计和/或对将经渲染的媒体内容传输到第一装置所需的时间的估计来获得。

[0026] 因此，可以在用户不会意识到由编码/解码和/或传输导致的延迟的意义上隐藏这些延迟。

[0027] 在一些实施例中，所述方法可以进一步包括将已经用于渲染媒体内容的姿势信息与当前姿势信息进行比较。例如，当前姿势信息可以是在再现经渲染的媒体内容时获得的姿势信息。所述方法还可以进一步包括基于比较的结果更新经渲染的媒体内容。更新可以基于已经用于渲染媒体内容的姿势信息与当前姿势信息之间的差异来执行。所述更新可以涉及例如通过旋转、级别改变和/或盲上混对经渲染的媒体内容进行外推。

[0028] 在一些实施例中，所述方法可以进一步包括在第二装置处确定指示经渲染的媒体内容如何响应于姿势信息的变化(例如，姿势的变化)而变化的梯度信息。梯度信息可以指示(对于音频内容)响应于用户(例如，用户的头部)的平移和/或旋转而引起的(例如，每个声道的)子带能级的变化。所述方法可以进一步包括将梯度信息与经渲染的媒体内容一起传输到第一装置。所述方法可以进一步包括在第一装置处将已经用于渲染媒体内容的姿势信息与当前姿势信息进行比较。已经(由第二装置)用于渲染媒体内容的姿势信息可以与经渲染的媒体内容一起传输到第一装置。在该姿势信息没有与经渲染的媒体内容一起被发送到第一装置的情况下，第一装置可以参考它已经发送到第二装置的姿势信息。例如，当前姿势信息可以是在再现经渲染的媒体内容时获得的姿势信息。所述方法还可以进一步包括基于梯度信息和比较的结果更新经渲染的媒体内容。更新经渲染的媒体内容可以基于已经用于渲染媒体内容的姿势信息与当前姿势信息之间的差异来执行。所述更新可以涉及例如通过旋转、级别改变和/或盲上混对经渲染的媒体内容进行外推。

[0029] 因此，可以校正姿势信息的预测中的小缺陷，并且可以避免姿势与再现的媒体内容之间的任何不匹配。

[0030] 在一些实施例中，媒体内容可以包括音频内容，并且经渲染的媒体内容可以包括经渲染的音频内容。然后，所述方法可以进一步包括将指示第一装置所处的环境的声学特性的环境信息传输到第二装置。在这种情况下，渲染媒体内容可以进一步基于环境信息。环境信息可以包括房间特性和/或双耳房间脉冲响应(BRIR)函数。

[0031] 这使得能够使再现的媒体内容专门适应用户所处的特定环境，从而增强用户对计算机介导的现实的体验。

[0032] 在一些实施例中，媒体内容可以包括音频内容，并且经渲染的媒体内容可以包括经渲染的音频内容。然后，所述方法可以进一步包括将指示用户或用户的一部分的形貌的

形貌信息传输到第二装置。在这种情况下，渲染媒体内容可以进一步基于形貌信息。形貌可以包括或对应于形状或大小，例如，用户头部的形状或大小。形貌信息可以包括头部相关传递函数(HRTF)。渲染可以是双耳渲染。

[0033] 这使得能够使再现的媒体内容专门适应用户或用户的一部分的特定形貌，从而增强用户对计算机介导的现实的体验。

[0034] 本公开的进一步方面涉及根据(例如，实施)上述方面及其实施例的第一装置、第二装置、以及第一装置和第二装置的系统。

[0035] 因此，本公开的另一方面涉及一种系统，所述系统包括用于再现媒体内容的第一装置和存储媒体内容的第二装置。第一装置可以适用于(被配置为)获得指示用户的位置和/或取向的姿势信息。第一装置可以进一步适用于(被配置为)将姿势信息传输到第二装置。第二装置可以适用于(被配置为)基于姿势信息来渲染媒体内容以获得经渲染的媒体内容。第二装置可以进一步适用于(被配置为)将经渲染的媒体内容传输到第一装置以进行再现。例如，第一装置和第二装置可以包括相应的处理器(或相应的处理器组)和耦接到相应处理器(或相应的处理器组)的存储器。处理器可以适用于(被配置为)执行上述操作。

[0036] 本公开的另一方面涉及用于提供媒体内容以供第一装置再现的第二装置。第二装置可以适用于(被配置为)接收指示第一装置的用户的位置和/或取向的姿势信息。第二装置可以进一步适用于(被配置为)基于姿势信息来渲染媒体内容以获得经渲染的媒体内容。第二装置还可以进一步适用于(被配置为)将经渲染的媒体内容传输到第一装置以进行再现。例如，第二装置可以包括处理器(或处理器组)和耦接到处理器(或处理器组)的存储器。处理器(或处理器组)可以适用于(被配置为)执行上述操作。

[0037] 本公开的另一方面涉及用于再现由第二装置提供的媒体内容的第一装置。第一装置可以适用于(被配置为)获得指示第一装置的用户的位置和/或取向的姿势信息。第一装置可以进一步适用于(被配置为)将姿势信息传输到第二装置。第一装置可以进一步适用于(被配置为)从第二装置接收经渲染的媒体内容。可以通过基于姿势信息渲染媒体内容来获得经渲染的媒体内容。第一装置可以进一步适用于(被配置为)再现经渲染的媒体内容。例如，第一装置可以包括处理器(或处理器组)和耦接到处理器(或处理器组)的存储器。处理器(或处理器组)可以适用于(被配置为)执行上述操作。

[0038] 应当注意，关于方法做出的任何陈述同样适用于在这些方法/系统中使用的相应系统和装置，并且反之亦然。

[0039] 本公开的又进一步方面涉及被配置为执行用于渲染音频内容的方法的系统、装置、方法和计算机可读存储介质，所述用于渲染音频内容的方法包括由发送方(S)装置接收用户位置和/或取向数据并发送通常由基于对象或FOA/HOA表示得到的相应的预渲染内容。由发送方生成的预渲染信号可以是双耳、FOA、HOA或任何类型的基于声道的渲染。所述方法可以进一步包括传输未压缩的预渲染内容。所述方法可以进一步包括对预渲染的内容进行编码并传输经编码的预渲染内容。所述方法可以进一步包括由接收方接收预渲染的内容。所述方法可以进一步包括由接收方对预渲染的、预编码的双耳化内容进行解码。用户位置和/或取向数据可以包括指示用户在世界空间中的位置和取向的本地姿势。用户位置数据可以从接收方传输到发送方。所述方法可以进一步包括将用于预渲染的双耳化内容的用户位置数据传输回接收方。所述方法可以进一步包括基于所接收的用户位置数据和本地位置

数据来外推预渲染的内容以确定更新的内容。所述方法可以进一步包括传输关于用户的形貌数据(例如,头部大小)以用于个性化双耳处理。所述方法可以进一步包括传输关于BRIR和房间表征的数据。所述方法可以进一步包括:基于确定内容以收听者不可知的方式(例如,不包括HRTF)传输而在接收方侧执行双耳渲染和个性化。所述方法可以进一步包括在时间点t1提供用户位置和/或取向数据P(t0)。未压缩的预渲染内容可以是双耳化的、未压缩的预渲染内容。

附图说明

[0040] 下文参考附图解释本公开的示例性实施例,在附图中:

[0041] 图1图示了接收方的第一示例;

[0042] 图2图示了接收方的第二示例;

[0043] 图3图示了接收方和服务器系统的第一示例;

[0044] 图4图示了发送方和接收方系统的第二示例;

[0045] 图5图示了发送方和接收方系统的第三示例;

[0046] 图6图示了发送方和接收方系统的第四示例;

[0047] 图7图示了处理媒体内容的方法的第一示例;

[0048] 图8图示了处理媒体内容的方法的第二示例;

[0049] 图9图示了处理媒体内容的方法的第三示例;并且

[0050] 图10图示了处理媒体内容的方法的第四示例。

具体实施方式

[0051] 如本领域技术人员将理解的,完全沉浸在虚拟世界中“欺骗”一个人的大脑相信所感知的内容。当视线受到视野的限制时,声音增加了不可见的维度(例如,从后面冲过来的公牛、在右边的响尾蛇、甚至是从左耳、头部后面移动到右耳的耳语(whisper))。因此,内容创作者可以利用声音来引导用户的注视,并且因此有效地讲述故事。通过基于对象或基于一阶/高阶高保真度立体声响复制(ambisonic)(FOA/HOA)的声音创作、封装和内容回放,如今正在电影院和家庭影院中提供沉浸式音频体验。VR声音需要声音精度才能完全沉浸在虚拟世界中。VR内容的创作者需要在三维空间中创作基于对象和/或基于HOA的声音的能力。此外,需要以允许用户享受内容的精度和效率对这样的内容进行编码、传递、解码和双耳地渲染(在头戴式耳机上或通过扬声器)。

[0052] 接收方可以基于各种参数(例如,带宽和媒体比特率)来选择内容的媒体格式表示,诸如经由MPEG-DASH或MPEG-MMT格式传递的过顶(OTT)内容。接收方还可以接收关于媒体消耗的信息。媒体格式表示的选择可以基于这种媒体消耗。例如,可以基于头戴式耳机或立体声扬声器(例如,具有串扰消除)输出的指示来选择预渲染的双耳化数据。

[0053] 本文描述的示例性实施例描述了适用于处理媒体内容(例如,渲染音频内容)的方法、装置和过程。尽管示例性实施例总体上涉及处理媒体内容(例如,包括音频内容和/或视频内容),但是本文档的其余部分中参考了音频内容,而非刻意进行限制。

[0054] 图1图示了用于进行双耳渲染的接收方/客户端系统100的示例。系统100可以接收音频输入101。音频输入101可以包括包含在来自发送方的编码比特流中的全部场景。接收

方系统100可以接收或检测与用户移动和/或用户头部取向有关的传感器数据(姿势信息)110。传感器数据110可以包括关于取向和位置的信息,如例如偏航、俯仰、滚动和/或(x,y,z)坐标。接收方系统100可以进一步包括解码器102,所述解码器可以将音频输入101解码为未压缩的音频和/或元数据120。接收方系统100可以进一步包括渲染器103,所述渲染器可以将未压缩的音频和/或元数据120渲染到双耳输出150。接收方系统100可以将双耳输出150输出到例如头戴式耳机输出。

[0055] 图1中图示的接收方/客户端系统100可能存在与本文档开头部分描述的时延和/或计算复杂性有关的问题。

[0056] 为了解决这些问题,本公开在用于处理媒体内容(例如,包括音频和/或视频内容)的系统中提出在接收方处获得用户的姿势信息、将姿势信息传输到发送方、基于姿势信息来渲染媒体内容、并将经渲染的媒体内容传输到接收方。因此,可以显著降低要在接收方侧执行的操作的计算复杂性。进一步地,经渲染的媒体内容可以以未压缩的形式传输,这可以减少姿势的变化(例如,头部移动)与对适应于这种姿势变化的再现媒体内容的感知(例如,声音的感知)之间的延迟。

[0057] 图7是示意性地图示了根据上述考虑因素的处理媒体内容的方法700的示例的流程图。媒体内容可以包括音频内容和/或视频内容。音频内容可以例如是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、基于声道的音频内容、或其组合。对媒体内容的处理可以涉及渲染媒体内容。所述方法可以在包括用于再现媒体内容的第一装置和用于提供媒体内容的第二装置的系统中执行。再现媒体内容可以涉及重放媒体内容。例如,第一装置可以被称为接收方、接收方装置、客户端、客户端装置、或重放装置。例如,第一装置可以包括、对应于计算机介导的现实(例如,VR、AR、MR)设备(如VR/AR/MR头戴式设备(例如,护目镜))或结合计算机介导的现实设备进行操作,并且可以与用户相关联。用户可以佩戴计算机介导的现实设备。第一装置可以包括或(通信地)耦接到用于检测用户或用户的一部分(例如,用户的头部)的姿势(例如,位置和/或取向)的传感器(例如,姿势传感器)。传感器可以进一步检测姿势的变化率((多个)一阶导数,例如速度、角速度/多个角速度、(多个)偏航/滚动/俯仰速率)。传感器还可以进一步检测变化率的变化率((多个)二阶导数,例如,加速度、(多个)角加速度)。由传感器输出的传感器数据可以被称为姿势信息。应当理解,姿势信息通常指示用户或用户的一部分(例如,用户的头部)的位置和/或取向(姿势)。进一步地,姿势信息可以指示姿势的一个或多个变化率(一阶导数)。又进一步地,姿势信息可以指示变化率的一个或多个变化率(二阶导数),例如,姿势的一个或多个变化率的变化率。传感器可以被布置成与用户或用户的相关部分(例如,头部)配准,例如作为计算机介导的现实设备的一部分(例如,VR/AR/MR头戴式设备/护目镜)、或作为由用户携带的移动(计算)设备(例如,智能电话、游戏控制器)的一部分。在这种情况下,传感器可以被称为嵌入式传感器。可替代地,传感器可以设置有保持跟踪用户(或用户的一部分)的姿势的位置服务器(例如,在OptiTrack系统或OptiTrack型系统中)或由所述位置服务器实施。通常,传感器可以是保持跟踪用户(或用户的一部分)的姿势的跟踪系统的一部分或由所述跟踪系统实施。这种位置服务器还可以保持跟踪多于一个用户的姿势。例如,第二装置可以被称为发送方、发送方装置、服务器、服务器装置或内容传递装置。第一装置和第二装置中的每一个都可以包括耦接到相应的存储器并且适用于(被配置为)执行下文阐述的相应操作的处理器(或处理器

组)。例如,所述处理器(或处理器组)可以适用于(被配置为)执行下文描述的方法700的相应步骤。可替代地或另外地,所述处理器(或处理器组)可以适用于(被配置为)执行以下进一步描述的方法800、方法900和方法1000中的任何一个方法的相应步骤。

[0058] 在步骤S710处,获得(例如,确定)指示用户(或用户的一部分,例如用户的头部)的位置和/或取向的姿势信息。该操作可以例如借助于传感器(例如,姿势传感器)来执行。在步骤S720处,将姿势信息传输到第二装置。在步骤S730处,基于姿势信息来渲染媒体内容以获得经渲染的媒体内容。即,基于用户或用户的一部分的位置和/或取向来渲染媒体内容。经渲染的媒体内容可以被称为预渲染的媒体内容(例如,预渲染的音频内容和/或预渲染的视频内容)。如果媒体内容包括音频内容,则音频内容可以例如被渲染为双耳音频内容、B格式音频内容、HOA音频内容、基于声道的音频内容、或其组合。通常,音频内容可以被渲染至两个或更多个声道和/或组件。如果媒体内容包括视频内容,则视频内容可以被图块化,并且整个视频场景的感兴趣区域可以被输出为例如经渲染的视频内容。在步骤S740处,将经渲染的媒体内容传输到第一装置以进行再现。步骤S710和S720可以在第一装置处执行/由第一装置执行,例如,分别通过传感器(例如,姿势传感器)和(第一)传输单元来执行。步骤S730和S740可以在第二装置处执行/由第二装置执行,例如,在渲染器和(第二)传输单元处执行。

[0059] 对于音频内容,方法700可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的两个或更多个扬声器生成经渲染的音频内容的听觉(例如,声学)表示的步骤。例如,所述两个或更多个扬声器可以是计算机介导的现实设备的一部分。对于视频内容,方法700可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的显示设备生成经渲染的视频内容的视觉表示的步骤。例如,所述显示设备可以是计算机介导的现实设备的一部分。通常,生成这样的表示可以在第一装置处执行/由第一装置执行。

[0060] 在图2中示意性地图示了根据上述方法的用于双耳渲染的接收方/客户端系统200的示例。所述系统可以实施方法700中的第一装置。作为经渲染的媒体内容(经渲染的音频内容)的示例,系统200可以接收音频输入201。例如,音频输入201可以采用双耳化的、未压缩的音频的形式。接收方系统200可以输出与用户移动和/或用户头部取向有关的传感器数据(作为姿势信息的示例)。例如,头部姿势(HeadPose)/传感器数据220可以包括关于偏航、俯仰、滚动和/或(x,y,z)坐标的信息。接收方系统200可以将传感器数据输出到发送方/服务器。发送方/服务器可以实施方法700中的第二装置。接收方系统200可以进一步生成音频输入201的听觉表示。例如,接收方系统可以将未压缩的音频输入201输出到头戴式耳机输出。

[0061] 如稍后将更详细描述的,图3、图4、图5和图6中图示的系统中的任何一个都可以实施方法700。

[0062] 为了进一步减少姿势变化与被呈现给用户的媒体内容表示的相应适配之间的延迟,第二装置可以预测姿势信息以预测可能由于传输到第一装置和/或编码/解码而导致的延迟(如下所述)。例如,在方法700中在步骤S730处渲染媒体内容可以涉及获得(例如,确定、计算)预测的姿势信息并基于预测的姿势信息(而不是基于从第一装置接收的姿势信息)来渲染媒体内容。

[0063] 图8是示意性地图示应用姿势信息的预测来处理媒体内容的方法800的示例的流

程图。除非另有说明，否则与上述方法700相关的陈述也适用于此。

[0064] 步骤S810和步骤S820分别对应于方法700中的步骤S710和S720。在步骤S830a处，基于在步骤S820处接收的姿势信息和先前姿势信息来获得(例如，确定、计算)预测的姿势信息。如果姿势信息包括姿势的一阶导数和/或二阶导数，则预测可以基于所述一阶导数和/或二阶导数，作为先前姿势信息的补充或替代。预测的姿势信息可以是用于未来定时的姿势信息，例如，指示在未来定时用户或用户的一部分(例如，头部)的位置和/或取向的姿势信息。在某些实施方式中，预测的姿势信息可以被预测来对预期要由第一装置处理经渲染的媒体内容以进行再现的定时进行估计。对预期第一装置处理经渲染的媒体内容以进行再现的定时的估计可以包括对将经渲染的媒体内容传输到第一装置所需的时间(时延)的估计。可替代地或另外地，如果应用了编码/解码(例如，压缩/解压缩)(下文描述)，则对所述定时的估计可以包括对编码/解码经渲染的媒体内容所需的时间(时延)的估计。即，可以进一步基于对传输经渲染的媒体内容所需的时间和/或对编码/解码经渲染的媒体内容所需的时间的估计来获得预测的姿势信息。先前姿势信息可以是在先前定时已经从第一装置接收的姿势信息。可以使用一个或多个先前姿势信息项(例如，经由外推或基于模型的预测技术)来获得预测的姿势信息。为此，可以存储先前姿势信息项(例如，预定数量的先前姿势信息项)。在步骤S830b处，基于预测的姿势信息来渲染媒体内容以获得经渲染的媒体内容。该操作可以与方法700中的步骤S730不同之处在于使用了预测的姿势信息而不是(在步骤S720或步骤S820接收的)所述姿势信息，但是在其他方面可以以与步骤S730相同的方式执行。在步骤S840处，将经渲染的媒体内容传输到第一装置以进行再现。步骤S810和S820可以在第一装置处/由第一装置执行。步骤S830a、S830b和S840可以在第二装置处/由第二装置执行。例如，步骤S830a可以由姿势预测器执行。

[0065] 对于音频内容，方法800可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的两个或更多个扬声器生成经渲染的音频内容的听觉(例如，声学)表示的步骤。例如，所述两个或更多个扬声器可以是计算机介导的现实设备的一部分。对于视频内容，方法800可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的显示设备生成经渲染的视频内容的视觉表示的步骤。例如，所述显示设备可以是计算机介导的现实设备的一部分。通常，生成这样的表示可以在第一装置处/由第一装置执行。

[0066] 在方法800的修改版中，预测的姿势信息可以在第一装置处被预测。即，第一装置可以执行如上文参考步骤S830a描述的处理，并且随后将预测的姿势信息发送到第二装置。应当理解，在这种情况下可以省略步骤S820。在从第一装置接收到预测的姿势信息之后，第二装置可以以上述方式继续其步骤S830b和后续步骤的处理。

[0067] 如稍后将更详细描述的，图3、图4、图5和图6中图示的系统中的任何一个都可以实施方法800或方法800的修改版。

[0068] 对用于渲染媒体内容的姿势信息的上述预测允许“隐藏”由传输和/或编码/解码引起的延迟，使得可以实现用户移动与经渲染的媒体内容的呈现之间的良好校准。因此，可以降低或完全避免用户受晕动病影响的风险，并且可以改善用户沉浸在计算机介导的现实的体验。在方法800的情况下，通过在服务器/发送方侧执行的过程(即，通过预测姿势信息并使用预测的姿势信息而不是从接收方/重放侧接收的姿势信息来渲染媒体内容)来实现对移动与经渲染的媒体内容的呈现之间的校准的改进。然而，在某些条件下，可能希望通过

在接收方或重放侧执行的措施来实现对移动与经渲染的媒体内容的呈现之间的校准的这种改进。

[0069] 图9是示意性地图示根据上述考虑因素(即,通过在接收方/重放侧执行的措施来改进移动与经渲染的媒体内容的呈现之间的校准)处理媒体内容的方法900的示例的流程图。

[0070] 步骤S910、步骤S920、步骤S930和步骤S940分别对应于方法700中的步骤S710至S740。在步骤S950处,将已经用于渲染媒体内容的姿势信息(例如,已经从第一装置接收的姿势信息)传输到第一装置。所述姿势信息可以与经渲染的媒体内容一起(例如,与经渲染的媒体内容相关联地)传输。在步骤S960处,将已经用于渲染媒体内容的姿势信息与当前姿势信息进行比较。当前姿势信息可以是在再现(例如,重放)经渲染的媒体内容时获得的姿势信息。但当前姿势信息可以在不同(稍后)的定时以上文参考步骤S710描述的方式获得。在步骤S970处,基于比较的结果更新经渲染的媒体内容。例如,经渲染的媒体内容可以基于已经用于渲染媒体内容的姿势信息与当前姿势信息之间的差异来更新。所述更新可以涉及对经渲染的媒体内容的外推。下文将参考图3描述这种更新的非限制性示例。步骤S910、S920、S960和S970可以在第一装置处/由第一装置执行。步骤S930、S940和S950可以在第二装置处/由第二装置执行。

[0071] 在某些实施方式中,可以省略步骤S950,即,可以不将已经用于渲染媒体内容的姿势信息传输到第一装置。在这种情况下,在步骤S960处,可以将已经在步骤S920处发送到第二装置的姿势信息称为已经用于渲染媒体内容的姿势信息。

[0072] 进一步地,在某些实施方式中,方法900可以包括确定经渲染的媒体内容如何响应于姿势信息的变化(例如,响应于用户的姿势或用户的头部的姿势的变化)而变化的梯度信息。然后,方法900还可以进一步包括将梯度信息传输到第一装置。例如,梯度信息可以与经渲染的媒体内容以及可选地已经用于渲染媒体内容的姿势信息一起(例如,相关联地)传输到第一装置。这些附加步骤可以在第二装置处执行。对于音频内容,梯度信息可以指示响应于用户或用户的一部分的平移和/或旋转而引起的(例如,每个声道或每个分量的)子带能级的变化。然后可以在步骤S970处使用梯度信息来更新/调整经渲染的媒体内容。例如,可以基于梯度信息、以及已经用于渲染媒体内容的姿势信息与当前姿势信息之间的差异来调整经渲染的音频内容的子带能级。一般而言,可以基于姿势差异和指示响应于姿势的变化而引起的经渲染的媒体内容的变化的梯度来更新/调整经渲染的媒体内容。

[0073] 对于音频内容,方法900可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的两个或更多个扬声器生成经渲染的音频内容的听觉(例如,声学)表示的步骤。例如,所述两个或更多个扬声器可以是计算机介导的现实设备的一部分。对于视频内容,方法900可以进一步包括例如经由作为第一装置的一部分或耦接到第一装置的显示设备生成经渲染的视频内容的视觉表示的步骤。例如,所述显示设备可以是计算机介导的现实设备的一部分。通常,生成这样的表示可以在第一装置处/由第一装置执行。

[0074] 如稍后将更详细描述的,图3、图4、图5和图6中图示的系统中的任何一个都可以实施方法900。

[0075] 为了进一步改善用户移动与经渲染的媒体内容的呈现之间的校准,可以将在服务器/发送方侧对姿势信息的预测与在接收方/重放侧对经渲染的媒体内容的更新进行组合。

[0076] 图10是示意性地图示根据上述考虑因素(即,通过在服务器/发送方侧执行的措施以及在接收方/重放侧执行的措施来改进移动与经渲染的媒体内容的呈现之间的校准)处理媒体内容的方法1000的示例的流程图。

[0077] 步骤S1010、步骤S1020和步骤S1040分别对应于方法700中的步骤S710、S720和S740。步骤S1030a和步骤S1030b分别对应于方法800中的步骤S830和S830b。在步骤S1050处,将预测的姿势信息(即,已经用于渲染媒体内容的姿势信息)传输到第一装置。预测的姿势信息可以与经渲染的媒体内容一起(例如,与经渲染的媒体内容相关联地)传输。在步骤S1060处,将预测的姿势信息与实际/当前姿势信息进行比较。实际姿势信息可以是在再现(例如,重放)经渲染的媒体内容时获得的姿势信息。但实际姿势信息可以在不同(稍后)的定时以上文参考步骤S710描述的方式获得。在步骤S1070处,基于比较的结果更新经渲染的媒体内容。例如,经渲染的媒体内容可以基于预测的姿势信息与实际姿势信息之间的差异来更新。通常,可以以与方法900中的步骤S970相同的方式执行更新。步骤S1010、S1020、S1060和S1070可以在第一装置处/由第一装置执行。步骤S1030a、S1030b和、S1040和S1050可以在第二装置处/由第二装置执行。

[0078] 在某些实施方式中,方法1000可以包括确定经渲染的媒体内容如何响应于姿势信息的变化(例如,响应于用户的姿势或用户的头部的姿势的变化)而变化的梯度信息。然后,方法1000还可以进一步包括将梯度信息传输到第一装置。例如,梯度信息可以与经渲染的媒体内容以及可选地已经用于渲染媒体内容的姿势信息一起(例如,相关联地)传输到第一装置。这些附加步骤可以在第二装置处执行。对于音频内容,梯度信息可以指示响应于用户或用户的一部分的平移和/或旋转而引起的(例如,每个声道或每个分量的)子带能级的变化。然后可以在步骤S1070处使用梯度信息来更新/调整经渲染的媒体内容。例如,可以基于梯度信息、以及已经用于渲染媒体内容的姿势信息与当前姿势信息之间的差异来调整经渲染的音频内容的子带能级。一般而言,可以基于姿势差异和指示响应于姿势的变化而引起的经渲染的媒体内容的变化的梯度来更新/调整经渲染的媒体内容。

[0079] 对于音频内容,方法1000可以进一步包括生成经渲染的音频内容的听觉(例如,声学)表示(例如,经由作为第一装置的一部分或耦接到第一装置的两个或更多个扬声器)的步骤。例如,所述两个或更多个扬声器可以是计算机介导的现实设备的一部分。对于视频内容,方法1000可以进一步包括生成经渲染的视频内容的视觉表示(例如,经由作为第一装置的一部分或耦接到第一装置的显示设备)的步骤。例如,所述显示设备可以是计算机介导的现实设备的一部分。通常,生成这样的表示可以在第一装置处/由第一装置执行。

[0080] 在方法1000的修改版中,预测的姿势信息可以在第一装置处预测。即,第一装置可以执行如上文参考步骤S1030a描述的处理,并且随后将预测的姿势信息发送到第二装置。应当理解,在这种情况下可以省略步骤S1020。在从第一装置接收到预测的姿势信息之后,第二装置可以使用预测的姿势信息以上文参考步骤S1030b描述的方式渲染媒体内容,并且以上文参考步骤S1040描述的方式将经渲染的媒体内容传输到第一装置。在这种情况下可以省略步骤S1050。在接收到经渲染的媒体内容之后,第一装置可以以上述方式执行步骤S1060和S1070。值得注意的是,由于在这种情况下对姿势信息的预测是在第一装置处执行的,因此第一装置不需要从第二装置接收预测的姿势信息。

[0081] 如稍后将更详细描述的,图3、图4、图5和图6中图示的系统中的任何一个都可以实

施方法1000或方法1000的修改版。

[0082] 在任何上述方法中，经渲染的媒体内容都可以以未压缩的形式传输到第一装置。这通过第二装置处的预渲染来实现，从而使得不需要传输全部媒体内容(例如，音频/视频场景的完整表示)。以未压缩形式传输经渲染的媒体内容有助于减少往返延迟，因为可以节省通常用于压缩/解压缩的时间。另一方面，如果带宽限制需要的话，可以在向第一装置传输之前对经渲染的媒体内容进行编码(压缩)。在这种情况下，如上所述，在获得预测的姿势信息时可以将进行编码/解码(例如，压缩/解压缩)所需的时间考虑在内。

[0083] 进一步地，对于音频内容，任何上述方法可以进一步包括：将指示第一装置所处的环境的声学特性的环境信息传输到第二装置。环境信息可以包括房间(room)特性和/或双耳房间脉冲响应(Binaural Room Impulse Response)(BRIR)函数。该步骤可以在第一装置处/由第一装置执行(例如，在设置时)。然后，可以进一步基于环境信息来渲染音频内容。可替代地或另外地，任何上述方法可以进一步包括：将指示用户或用户的一部分的形貌的形貌信息传输到第二装置。形貌可以包括或对应于形状或大小，例如，用户头部的形状或大小。形貌信息可以包括头部相关传递函数(HRTF)。渲染可以是双耳渲染。该步骤可以在第一装置处/由第一装置例如在设置时执行。然后，可以进一步基于形貌信息来渲染音频内容。

[0084] 图3图示了包括服务器/发送方300和客户端/接收方350的示例性系统的进一步细节。如上所述，该系统可以实施方法700、800、900和1000中的任何一个。例如实施第二装置的服务器/发送方300可以包括渲染器320(例如，音频渲染器)和编码器330。例如实施第一装置的客户端/接收方350可以在时间点t0将当前姿势(例如，头部姿势)P(t0)发送到服务器/发送方300。当前姿势P(t0)还可以包括指定当前姿势P(t0)被创建的时间的时间戳t0本身。姿势P(t0)可以由姿势350框确定并发送。

[0085] 例如实施第二装置的服务器/发送方300可以进一步包括位置预测器310。服务器/发送方300可以在时间点t1接收用户位置和当前姿势P(t0)(对应于头部取向)，其中，t1>t0。位置预测器310可以使用所接收的当前姿势P(t0)和t0本身来预测位置P(t1)。位置预测器310可以考虑先前接收的姿势P(tn)和tn来预测位置P(t1)，其中，n可以是0到负无穷大(来自较早时间点的姿势和时间戳值)。位置P(t1)可以类似于姿势P(t0)。位置P(t1)可以由音频渲染器320用来在时间点t1渲染音频场景，并且因此确定经渲染的音频数据R(t1)340。可以使用音频编码器330对经渲染的音频数据R(t1)340进行编码以确定音频数据A(t1)。服务器/发送方300可以将音频数据A(t1)和位置P(t1)发送到客户端/接收方350。位置P(t1)可以被编码为音频比特流的一部分。客户端/接收方350可以在时间点t2从服务器/发送方300接收音频数据A(t1)和位置P(t1)(例如，以元数据的形式)，其中，t2>t1。客户端/接收方350可以在音频解码器351处接收音频数据A(t1)和位置P(t1)，所述音频解码器可以确定未压缩的音频U(t1)。头部姿势/传感器数据352框可以在时间点t2确定姿势P(t2)。音频外推器353可以在时间点t2使用所接收的P(t1)通过从姿势P(t2)中减去姿势P(t1)来计算姿势差DeltaP。音频外推器353可以使用DeltaP来在输出390之前适配/外推未压缩音频U(t1)。在音频内容是FOA并且运动被限制为偏航、俯仰和/或滚动移动的情况下，客户端/接收方350可以应用局部旋转作为外推的一部分。在音频内容是预渲染的双耳内容或预渲染的基于声道的内容的情况下，客户端/接收方350可以进一步应用盲上混(blind upmixing)作为外推的一部分。

[0086] 代替预测位置P(t1),可以针对预期客户端/接收方350要接收或处理音频数据的时间点t2'来预测位置P(t2')。考虑到传输和/或编码/解码音频数据所需的时间(时延),可以从时间点t1开始估计时间点t2'。然后,上文中的P(t1)、R(t1)、A(t1)和U(t1)将必须被分别替换为P(t2')、R(t2')、A(t2')和U(t2')。上述元件中的任何一个都可以由相应装置的处理器(或处理器组)实施。

[0087] 可以使用MPEG-H 3D音频(ISO/IEC 23008-3)和/或MPEG标准的未来版本中的以下语法来传输P(t)3自由度(P(t)3Degrees of Freedom)(3DoF)数据:

[0088]

语法	比特数	助记符
mpegh3daSceneDisplacementData()		
{		
sd_yaw;	9	uimsbf
sd_pitch;	9	uimsbf
sd_roll;	9	uimsbf
}		

[0089] 表1

[0090] 可以根据MPEG-H 3D音频(ISO/IEC 23008-3)和/或MPEG标准的未来版本来定义语义。

[0091] 用于传输6DoF数据和时间戳的完整语法可以如下所示:

[0092]

语法	比特数

head_pose()
{
if (b_yaw) {
yaw;
}
if (b_pitch) {
pitch;
}
if (b_roll) {
roll;
}
if (b_x_pos) {
x_pos;
}
if (b_y_pos) {
y_pos;
}
if (b_z_pos) {
z_pos;
}
if (b_timestamp) {
timestamp;
}
if (PosDerivatives) {
d_x_pos
d_y_pos
d_z_pos
dd_x_pos
dd_y_pos
dd_z_pos
}
if (OrientDerivatives) {
d_roll
d_pitch
d_yaw
dd_roll
dd_pitch
dd_yaw
}
	
	1
	9
	
	1
	9
	
	1
	9
	
	1
	9
	
	1
	9
	
	1
	可变
	
	1
	9
	9
	9
	9
	9
	9
	
	1
	9
	9
	9
	9
	9
	9

}	

[0093] 表2

[0094] 可以根据MPEG-H 3D音频(ISO/IEC 23008-3)和/或MPEG标准的未来版本来定义语义。

[0095] 图4图示了包括发送方400和接收方450的示例性系统。图4中描述的系统可以包括图1至图3中描述的系统的全部或部分方面。具体地,所述系统可以实施上述方法700、800、900和1000中的任何一个或全部。例如实施第二装置的发送方/服务器400可以接收完整音频场景(例如,基于FOA/HOA或基于对象的场景)401和完整视频场景(例如,360°视频)402(作为媒体内容的示例)。完整音频场景401可以由音频编码器410处理以确定完整的音频场

景传递441。完整音频场景401可以由整个音频场景和/或伴随元数据(如音频对象位置、方向等)构成。完整视频402可以由内容传递选择420处理。完整视频420可以被分割成不同的部分(诸如,感兴趣区域),并且通过内容选择420相应地“被图块化”(360°视频可以被分割成图块)以确定图块402a。内容传递和选择420可以使用如被描述为来自图3中的位置预测器310的输出的预测位置P(t1)(或预测的位置P(t2’)),或者所述内容传递和选择420可以使用未改变的头部姿势/传感器数据454。例如,可以基于从接收方450接收的传感器数据454在内容传递选择420中选择完整360°视频402中的图块402a。该选择可以被称为对视频内容的渲染。视频编码器430对图块402a进行编码以输出可被传输到(例如,实施第一装置的)客户端/接收方450的感兴趣区域视频442。接收方450可以包括可以接收感兴趣区域视频442的视频解码器452。视频解码器452可以使用感兴趣区域442来对视频进行解码并将其输出到视频输出492。完整音频场景401可以由音频解码器451接收,所述音频解码器可以对内容进行解码并将经解码的音频场景提供给音频渲染器453。音频解码器451可以向音频渲染器453提供未压缩的音频和元数据455(它们可以对应于经解码的音频场景)。音频渲染器453可以基于传感器数据454来渲染经解码的音频,并且可以输出音频输出491。可以从能够检测用户的移动和/或用户的头部取向的传感器(例如,基于陀螺仪的传感器)接收传感器数据454。然后可以进一步将所述传感器数据454提供给音频渲染器453以使完整音频场景401适应用户的当前头部取向和/或位置,并且将所述传感器数据454提供给内容传递选择420以使完整视频场景402适应用户的当前头部取向和/或位置。值得注意的是,在图4的示例性系统中,在服务器/发送方侧渲染视频内容(即,在服务器/发送方侧生成准备好在接收方/重放侧重放的视频内容),而在接收方/重放侧渲染音频内容。上述元件中的任何一个都可以由相应装置的处理器(或处理器组)实施。

[0096] 图5图示了包括发送方500和接收方550的示例性系统。所述系统可以实施上述方法700、800、900和1000中的任何一个或全部。例如实施第二装置的发送方/服务器500可以接收完整音频场景(例如,基于HOA或基于对象的场景)501和完整视频场景(例如,360°视频)502(作为媒体内容的示例)。完整音频场景501可以由音频渲染器510处理以确定双耳音频数据541。音频渲染器510可以考虑传感器数据545来确定双耳音频数据541。传感器数据545可以包括偏航、俯仰、滚动、x、y、z信息。双耳音频数据541可以是未压缩的、无损压缩的、或有损低时延压缩的。例如,双耳音频数据551可以是未压缩音频580,所述未压缩音频可以由(例如,实施第一装置的)接收方550接收并提供给双耳音频输出591。完整视频502可以由内容传递选择520处理。完整视频502可以被分割成诸如感兴趣区域等不同的部分,并且在内容传递选择520中相应地“被图块化”(360°视频可以被分割成图块)以确定图块502a。可以基于从接收方550接收的传感器数据545在内容传递选择520中选择完整360°视频502中的图块502a。该选择可以被称为对视频内容的渲染。视频编码器530对图块502a进行编码以输出可以被传输到客户端/接收方550的感兴趣区域视频542。接收方550可以包括可以接收感兴趣区域视频542的视频解码器552。视频解码器552可以使用感兴趣区域542来对视频进行解码并将其输出到视频输出592。传感器数据545可以从能够检测用户的移动和/或用户的头部取向的传感器(例如,基于陀螺仪的传感器)接收。然后可以进一步将所述传感器数据545提供给内容传递选择520,以使完整视频场景502适应用户的当前头部取向和/或位置。然后可以进一步将所述传感器数据545提供给内容音频渲染器510,以使完整音频场景

501适应用户的当前头部取向和/或位置。上述元件中的任何一个都可以由相应装置的处理器(或处理器组)实施。

[0097] 图6图示了包括发送方600和接收方650的示例性系统。所述系统可以实施上述方法700、800、900和1000中的任何一个或全部。例如,实施第二装置的发送方/服务器600可以接收完整音频场景(例如,基于HOA或基于对象的场景)601和完整视频场景(例如,360°视频)602(作为媒体内容的示例)。完整音频场景601可以由音频渲染器610处理,并且音频渲染器610的输出然后可以由低延迟(LowDelay)音频编码器660处理。音频渲染器610可以考虑传感器数据645。低延迟音频编码器660可以输出双耳音频数据641,然后可以将所述双耳音频数据645发送到例如实施第一装置的接收方650。双声道音频数据641可以在接收方650处由低延迟音频解码器670接收,所述低延迟音频解码器670将双耳音频数据641转换为未压缩音频680。然后可以将未压缩音频680提供给双耳音频输出691。完整视频602可以由内容传递选择620处理。完整视频602可以被分割成诸如感兴趣区域等不同的部分,并且在内容传递选择620中相应地“被图块化”(360°视频可以被分割成图块)以确定可以基于从接收方650接收的传感器数据645而在内容传递选择620中被选择的图块。该选择可以被称为对视频内容的渲染。视频编码器630对图块和/或视频进行编码以输出可以被传输到客户端/接收方650的感兴趣区域视频642。接收方650可以包括可以接收感兴趣区域视频642的视频解码器652。视频解码器652可以使用感兴趣区域642来对视频进行解码并将其输出到视频输出692。传感器数据645可以从能够检测用户的移动和/或用户的头部取向的传感器(例如,基于陀螺仪的传感器)接收。然后可以进一步将所述传感器数据645提供给内容传递选择620,以使完整视频场景602适应用户的当前头部取向和/或位置。然后可以进一步将所述传感器数据645提供给内容音频渲染器610,以使完整音频场景601适应用户的当前头部取向和/或位置。上述元件中的任何一个都可以由相应装置的处理器(或处理器组)实施。

[0098] 传统上,如图1和图4所示,从发送方(S)传输到接收方(R)的音频(作为媒体内容的非限制性示例)在接收方处被渲染。为了最大化接收方侧的灵活性,可以发送音频场景的复杂表示(如可以在接收侧自适应地渲染的对象或HOA),例如以匹配本地收听者的视点/姿势。然而,编码这样的表示可能需要大的时延,这将阻止这些方法用于通信或交互式应用。

[0099] 本公开提供了用于减少所提到的时延和/或用于降低接收方中的计算复杂性的方法、系统和装置。从接收方传输到发送方的用户位置和取向允许服务器/发送方计算出与接收方的当前姿势/视点紧密匹配的内容的更紧凑、预渲染的版本。然而,从发送方到接收方的传输时延将在本地接收姿势与已在服务器上计算渲染的姿势之间引入可能的不匹配。本公开提出:发送方用信号通知已执行渲染的位置,从而允许接收方将经渲染的信号外推到其当前的本地姿势。另外,发送方可以发送音频场景的预渲染、未压缩或有损压缩表示,以便消除系统中的编码和解码时延。发送方执行例如针对双耳立体声、FOA或HOA表示的渲染算法。渲染算法可以将诸如音频对象等音频数据渲染至两个声道(例如,预渲染的双耳化内容)以输出声道。然后可以对声道进行编码,特别是如果需要压缩(例如,取决于系统的带宽)以输出经编码的音频数据比特流的话。可以将信号传输到客户端或接收方,并且可以经由头戴式耳机或立体声扬声器系统来输出信号。

[0100] 当双耳化回放需要适应于用户头部的物理现象时,接收方可以传输与用户头部的属性相对应的头部相关传递函数(HRTF)。接收方可以进一步传输与旨在用于再现的房间相

对应的双耳房间脉冲响应(BRIR)函数。可以在传输设置期间传输该信息。

[0101] 本公开的实施例可以提供至少以下优点：

[0102] • 如果仅传输双耳化(立体声)数据，则用于无损音频数据传输的传输比特率可以与完整压缩音频场景的比特率相似或相当。

[0103] • 传输未压缩或无损音频流将消除或减少因编码和解码而导致的时延。例如，由编码/解码所产生的时延可以减少到零，这将导致运动到耳朵时延的整体减少。

[0104] • 当不对音频数据进行压缩时，接收方将仅输出音频而不进行解码或渲染。这将导致接收方处的计算复杂性降低，因为不需要执行解码和/或在发送方侧已经完成了渲染。

[0105] • 可以在以较高比特率的最小时延和最小接收方计算复杂性与以较高时延和较高接收方计算复杂性的最小比特率之间进行不同的折衷，例如：

[0106] o传输未压缩数据以实现最小时延和计算复杂性，但需要足够的带宽来传输未压缩数据，

[0107] o如果带宽不足以传输未压缩数据，则传输无损压缩数据以实现最小时延和稍高的计算复杂性，

[0108] o如果带宽受限，则传输有损但低延迟的压缩数据，以实现低时延和较高的计算复杂性。

[0109] 当R和S同时作为接收方和发送方时，上述内容也适用于R与S之间的双向通信。

[0110] 表3示出了说明这种折衷的系统比较的示例。

[0111]

编解码器	比特率(传输质量)	编码时延	解码时延	传输时延	整体时延	接收方计算复杂性	输出灵活性
3D 音频编解码器	800 kbps-1200 kbps	~120 ms	~40 ms	~5 ms	~165 ms	高	高
预渲染无损编码	768 kbps	~5 ms	~0 ms	2*~5 ms	~15 ms	低	低

[0112] 表3

[0113] 在某些上下文中，本公开的实施例可以涉及基于内容外推来隐藏传输时延。当总时延(例如，传输时延)太高(通常高于20毫秒)时，期望编码格式和/或回放系统在等待传递下一个更新的内容帧的同时提供用于对内容进行外推以匹配接收方的本地姿势(位置和取向)的均值。可以基于音频数据的往返中的所有时延的总和来确定总时延。例如，总时延可以基于往返时延、编码时延、解码时延和渲染时延。

[0114] 隐藏该时延可以通过以下方式来实现：将本地姿势从接收方传输到发送方/服务器以进行渲染(例如，如上文参考步骤S920和S1020所描述的)，并且使发送方/服务器发送回哪个姿势已被用于每个经渲染的内容帧(例如，如上文参考步骤S950和S1050所描述的)。发送方/发送方可以预测用户的移动，以便补偿在发送方渲染内容之时与在接收方接收到内容之时之间引入的附加时延，包括将先前接收的位置考虑在内。

[0115] 然后，在给定了用于在发送方侧渲染内容的姿势与接收方R的本地姿势(例如，当前姿势或实际姿势)之间的增量的情况下，接收方可以对从服务器接收的预渲染音频进行外推(例如，如上文参考步骤S970和S1070所描述的)。

[0116] 可以基于经渲染的内容的灵活性以多种方式实施这种外推。在一个示例中，当内

容是预渲染的高保真度立体声响复制B格式并且运动是三自由度运动时，外推可以基于在回放之前FOA或B格式内容的客户端侧局部旋转。在另一示例中，对于预渲染的双耳内容，可以通过盲上混(参见附录A)或通过向双耳流添加元数据(参见附录B)来实现外推。在另一示例中，对于预渲染的基于声道的内容，可以在接收端应用低时延盲上混音器(upmixer)。

[0117] 如果渲染和编码紧密集成在发送方侧，则可以通过添加元数据编码(例如不同子带的方向/距离或基于当前渲染位置P的能量梯度$\nabla E(P)$)来增加预渲染内容的灵活性。

[0118] 如果要渲染的原始内容是基于对象的，则可以围绕期望位置计算多个渲染并对水平梯度进行编码。该水平梯度G通常由3D向量(三个轴x、y、z中的每一个都有一个值)构成。接收方然后可以基于预渲染位置P与当前接收方位置P'之间的差异简单地调整接收信号中的子带能量E(P)，如$E(P') = E(P) \cdot (P'-P) \cdot \nabla E(P)$。

[0119] 接收方可以使用该额外信息来进一步外推预渲染的流(即，预渲染的媒体内容)，例如(使用距离信息)解决视差效应或(使用水平梯度信息)调整渲染的级别。

[0120] 在一个示例中，如果接收方在计算能力方面受到约束，则可以在发送方侧的编码期间执行上混。例如，可以将B格式或声道转换为对象。这可能增加编码路径时延，但是得到的内容可以更灵活并且可以在接收方端进行外推。

[0121] 对于游戏用途，其中，用户动作(例如，按钮触发)可以影响游戏可玩性，整体系统时延仍然需要小于20毫秒，这可能会阻止运行复杂的上混操作。因此，诸如B格式等灵活格式可能是使用低时延无损或有损编解码器进行渲染/传输的最佳候选项，因为这种格式也可以在接收端以低时延进行渲染和旋转。

[0122] 各种音频编解码器可以包含上述数据传输模式。可以针对以下内容来适配编解码器：(i)传输无损编码(零时延编码)的立体声音频数据或低时延有损数据的可能性；(ii)在需要关闭“常规”渲染(例如，设备中的双耳化)的情况下，用于发信号通知该内容已经被预渲染的手段(例如，比特流语法字段，Dolby AC-4和MPEG-H第3部分(3D音频)都已经包括这种比特字段，如Dolby AC-4中的b_pre_virtualized)；以及(iii)用于在必要时传输HRTF和BRIR的手段。

[0123] 因此，在本公开的上下文中，发送方还可以向接收方提供发送方提供了预渲染音频内容的指示(例如，标志、比特字段、语法字段/元素、参数)。如果接收方接收到这样的指示，则接收方可以放弃对音频内容进行任何(接收方侧)渲染。例如，对于双耳预渲染音频内容，接收方可以直接将如从发送方接收的预渲染音频内容路由到头戴式耳机(的扬声器)以进行再现，而无需任何进一步的渲染。这种指示可以是参数directHeadphone的形式，所述参数在比特流中用信号通知给接收方。如果双耳输出被渲染，则directHeadphone参数可以定义(类型)声道的相应信号组直接进入头戴式耳机输出。所述信号可以被路由到左右头戴式耳机声道。

[0124] 表4中再现了该参数的语法的可能示例。

[0125]

语法	比特数	助记符
prodMetadataConfig() { /*高分辨率物距*/ **hasObjectDistance;**	**1**	**bslbf**

/*直接到耳机*/
for (gp = 0; gp < numChannelGroups;
gp++) {		

directHeadphone [gp];	**1**	**bslbf**
} }		

[0126] 表4

[0127] 可以根据MPEG-H 3D音频(ISO/IEC 23008-3)和/或MPEG标准的未来版本来定义语义。

[0128] 在下文列出的枚举示例性实施例(EEE)中概述了本公开的其他示例性实施例。

[0129] 第一EEE涉及一种处理媒体内容以供第一装置再现的方法,所述方法包括:获得指示用户的位置和/或取向的姿势信息;将所述姿势信息传输到提供所述媒体内容的第二装置;基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容;以及将所述经渲染的媒体内容传输到所述第一装置以进行再现。

[0130] 第二EEE涉及根据第一EEE所述的方法,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容;和/或所述媒体内容包括视频内容,并且所述经渲染的媒体内容包括经渲染的视频内容。

[0131] 第三EEE涉及根据第一EEE所述的方法,其中,所述媒体内容包括音频内容,并且所述经渲染的媒体内容包括经渲染的音频内容,并且所述方法进一步包括生成所述经渲染的音频内容的听觉表示。

[0132] 第四EEE涉及根据第二EEE所述的方法,其中,所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种,或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0133] 第五EEE涉及根据第二或第三EEE所述的方法,其中,所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种,或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0134] 第六EEE涉及根据第一至第五EEE中任一项所述的方法,其中,所述渲染涉及:基于所述姿势信息和先前姿势信息获得预测的姿势信息,并基于所述预测的姿势信息来渲染所述媒体内容以获得所述经渲染的媒体内容。

[0135] 第七EEE涉及根据第六EEE所述的方法,进一步包括:将所述预测的姿势信息与所述经渲染的媒体内容一起传输到所述第一装置。

[0136] 第八EEE涉及根据第七EEE所述的方法，进一步包括：将所述预测的姿势信息与实际姿势信息进行比较，并基于所述比较的结果更新所述经渲染的媒体内容。

[0137] 第九EEE涉及根据第八EEE所述的方法，其中，所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计，并且所述实际姿势信息是在所述第一装置实际处理所述经渲染的媒体内容以进行再现的定时获得的姿势信息。

[0138] 第十EEE涉及根据第一至第九EEE中任一项所述的方法，其中，所述经渲染的媒体内容以未压缩的形式传输到所述第一装置。

[0139] 第十一EEE涉及根据第一至第十EEE中任一项所述的方法，进一步包括：

[0140] 在向所述第一装置传输之前对所述经渲染的媒体内容进行编码；以及在所述第一装置处接收到经编码的所述经渲染的媒体内容之后，对经编码的所述经渲染的媒体内容进行解码。

[0141] 第十二EEE涉及根据第九EEE或包括第九EEE的特征的任何EEE所述的方法，其中，对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时的所述估计包括：对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

[0142] 第十三EEE涉及根据第六EEE或包括第六EEE的特征的任何EEE所述的方法，其中，所述预测的姿势信息是进一步基于对编码和解码所述经渲染的媒体内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计来获得的。

[0143] 第十四项涉及根据第一至第十三EEE中任一项所述的方法，进一步包括：将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较，并基于所述比较结果更新所述经渲染的媒体内容。

[0144] 第十五EEE涉及根据第一至第十四EEE中任一项所述的方法，进一步包括：在所述第二装置处确定指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息；将所述梯度信息与所述经渲染的媒体内容一起传输到所述第一装置；在所述第一装置处将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较；以及基于所述梯度信息和所述比较的结果更新所述经渲染的媒体内容。

[0145] 第十六EEE涉及根据第一至第十五EEE中任一项所述的方法，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述方法进一步包括将指示所述第一装置所处的环境的声学特性的环境信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述环境信息。

[0146] 第十七EEE涉及根据第一至第十六EEE中任一项所述的方法，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述方法进一步包括将指示所述用户或所述用户的一部分的形貌的形貌信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述形貌信息。

[0147] 第十八EEE涉及一种系统，所述系统包括用于再现媒体内容的第一装置和存储媒体内容的第二装置，其中，所述第一装置适用于：获得指示用户的位置和/或取向的姿势信息，并将所述姿势信息传输到所述第二装置，并且所述第二装置适用于：基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容，并将所述经渲染的媒体内容传输到所述第

一装置以进行再现。

[0148] 第十九EEE涉及根据第十八EEE所述的系统，其中，所述媒体内容包括音频内容并且所述经渲染的媒体内容包括经渲染的音频内容，和/或所述媒体内容包括视频内容并且所述经渲染的媒体内容包括经渲染的视频内容。

[0149] 第二十EEE涉及根据第十八EEE所述的系统，其中，所述媒体内容包括音频内容并且所述经渲染的媒体内容包括经渲染的音频内容，并且所述第一装置进一步适用于生成所述经渲染的音频内容的听觉表示。

[0150] 第二十一EEE涉及根据第十九EEE所述的系统，其中，所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种，或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0151] 第二十二EEE涉及根据第十九至第二十一EEE中任一项所述的系统，其中，所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种，或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0152] 第二十三EEE涉及根据第十八至第二十二EEE中任一项所述的系统，其中，所述第二装置进一步适用于：基于所述姿势信息和先前姿势信息获得预测的姿势信息，并基于所述预测的姿势信息来渲染所述媒体内容以获得所述经渲染的媒体内容。

[0153] 第二十四EEE涉及根据第二十三EEE所述的系统，其中，所述第二装置进一步适用于：将所述预测的姿势信息与所述经渲染的媒体内容一起传输到所述第一装置。

[0154] 第二十五EEE涉及根据第二十四EEE所述的系统，其中，所述第一装置进一步适用于：将所述预测的姿势信息与实际姿势信息进行比较，并基于所述比较的结果更新所述经渲染的媒体内容。

[0155] 第二十六EEE涉及根据第二十五EEE所述的系统，其中，所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计，并且所述实际姿势信息是在所述第一装置实际处理所述经渲染的媒体内容以进行再现的定时获得的姿势信息。

[0156] 第二十七EEE涉及根据第十八至第二十六EEE中任一项所述的系统，其中，所述经渲染的媒体内容以未压缩的形式传输到所述第一装置。

[0157] 第二十八EEE涉及根据第十八至第二十七EEE中任一项所述的系统，其中，所述第二装置进一步适用于在向所述第一装置传输之前对所述经渲染的媒体内容进行编码，并且所述第一装置进一步适用于在所述第一装置处接收到经编码的所述经渲染的媒体内容之后，对经编码的所述经渲染媒体内容进行解码。

[0158] 第二十九EEE涉及根据第二十六EEE或包括第二十六EEE的特征的任何EEE所述的系统，其中，对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时的所述估计包括：对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

[0159] 第三十EEE涉及根据第二十三EEE或包括第二十三EEE的特征的任何EEE所述的系统，其中，所述预测的姿势信息是进一步基于对编码和解码所述经渲染的媒体内容所需的

时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计来获得的。

[0160] 第三十一EEE涉及根据第十八至第三十EEE中任一项所述的系统，其中，所述第一装置进一步适用于：将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较，并基于所述比较结果更新所述经渲染的媒体内容。

[0161] 第三十二EEE涉及根据第十八至第三十一EEE中任一项所述的系统，其中，所述第二装置进一步适用于：确定指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息；以及将所述梯度信息与所述经渲染的媒体内容一起传输到所述第一装置，并且所述第一装置进一步适用于：将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较，并基于所述梯度信息和所述比较结果更新所述经渲染的媒体内容。

[0162] 第三十三EEE涉及根据第十八至第三十二EEE中任一项所述的系统，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第一装置进一步适用于将指示所述第一装置所处的环境的声学特性的环境信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述环境信息。

[0163] 第三十四EEE涉及根据第十八至第三十三EEE中任一项所述的系统，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第一装置进一步适用于将指示所述用户或所述用户的一部分的形貌的形貌信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述形貌信息。

[0164] 第三十五EEE涉及一种用于提供媒体内容以供第一装置再现的第二装置，所述第二装置适用于：接收指示所述第一装置的用户的位置和/或取向的姿势信息；基于所述姿势信息来渲染所述媒体内容以获得经渲染的媒体内容；以及将所述经渲染的媒体内容传输到所述第一装置以进行再现。

[0165] 第三十六EEE涉及根据第三十五EEE所述的第二装置，其中，所述媒体内容包括音频内容并且所述经渲染的媒体内容包括经渲染的音频内容，和/或所述媒体内容包括视频内容并且所述经渲染的媒体内容包括经渲染的视频内容。

[0166] 第三十七EEE涉及根据第三十六EEE所述的第二装置，其中，所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种，或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0167] 第三十八EEE涉及根据第三十六EEE所述的第二装置，其中，所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种，或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0168] 第三十九EEE涉及根据第三十五至第三十八EEE中任一项所述的第二装置，进一步适用于：基于所述姿势信息和先前姿势信息获得预测的姿势信息，并基于所述预测的姿势信息来渲染所述媒体内容以获得所述经渲染的媒体内容。

[0169] 第四十EEE涉及根据第三十九EEE所述的第二装置，进一步适用于：将所述预测的姿势信息与所述经渲染的媒体内容一起传输到所述第一装置。

[0170] 第四十一EEE涉及根据第三十九或第四十EEE所述的第二装置，其中，所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计。

[0171] 第四十二EEE涉及根据第三十五至第四十一EEE中任一项所述的第二装置，其中，所述经渲染的媒体内容以未压缩的形式传输到所述第一装置。

[0172] 第四十三EEE涉及根据第三十五至第四十二EEE中任一项所述的第二装置，进一步适用于在向所述第一装置传输之前对所述经渲染的媒体内容进行编码。

[0173] 第四十四EEE涉及根据第四十一EEE或包括第四十一EEE的特征的任何EEE所述的第二装置，其中，对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时的所述估计包括：对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

[0174] 第四十五EEE涉及根据第三十九EEE或包括第三十九EEE的特征的任何EEE所述的第二装置，其中，所述预测的姿势信息是进一步基于对编码和解码所述经渲染的媒体内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计来获得的。

[0175] 第四十六EEE涉及根据第三十五至第四十五EEE中任一项所述的第二装置，进一步适用于：确定指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息；以及将所述梯度信息与所述经渲染的媒体内容一起传输到所述第一装置。

[0176] 第四十七EEE涉及根据第三十五至第四十六EEE中任一项所述的第二装置，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第二装置进一步适用于从所述第一装置接收指示所述第一装置所处的环境的声学特性的环境信息，并且所述渲染所述媒体内容进一步基于所述环境信息。

[0177] 第四十八EEE涉及根据第三十五至第四十七EEE中任一项所述的第二装置，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第二装置进一步适用于从所述第一装置接收指示所述用户或所述用户的一部分的形貌的形貌信息，并且所述渲染所述媒体内容进一步基于所述形貌信息。

[0178] 第四十九EEE涉及一种用于再现由第二装置提供的媒体内容的第一装置，所述第一装置适用于：获得指示所述第一装置的用户的位置和/或取向的姿势信息；将所述姿势信息传输到所述第二装置；从所述第二装置接收经渲染的媒体内容，其中，所述经渲染的媒体内容已经通过基于所述姿势信息来渲染所述媒体内容而获得；以及再现所述经渲染的媒体内容。

[0179] 第五十EEE涉及根据第四十九EEE所述的第一装置，其中，所述媒体内容包括音频内容并且所述经渲染的媒体内容包括经渲染的音频内容，和/或所述媒体内容包括视频内容并且所述经渲染的媒体内容包括经渲染的视频内容。

[0180] 第五十一EEE涉及根据第四十九EEE所述的第一装置，其中，所述媒体内容包括音频内容并且所述经渲染的媒体内容包括经渲染的音频内容，并且所述第一装置进一步适用于生成所述经渲染的音频内容的听觉表示。

[0181] 第五十二EEE涉及根据第五十或第五十一EEE所述的第一装置，其中，所述音频内容是基于一阶高保真度立体声响复制FOA的音频内容、基于更高阶高保真度立体声响复制

HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的一种，或者是基于FOA的音频内容、基于HOA的音频内容、基于对象的音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0182] 第五十三EEE涉及根据第五十至第五十二EEE中任一项所述的第一装置，其中，所述经渲染的音频内容是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的一种，或者是双耳音频内容、FOA音频内容、HOA音频内容、或基于声道的音频内容中的两种或更多种的组合。

[0183] 第五十四EEE涉及根据第四十九至第五十三EEE中任一项所述的第一装置，进一步适用于：接收已经用于渲染所述媒体内容的姿势信息以及来自所述第二装置的所述经渲染的媒体内容；将已经用于渲染所述媒体的所述姿势信息与实际姿势信息进行比较；以及基于所述比较的结果更新所述经渲染的媒体内容。

[0184] 第五十五EEE涉及根据第五十四EEE所述的第一装置，其中，所述实际姿势信息是在所述第一装置处理所述经渲染的媒体内容以进行再现的定时获得的姿势信息。

[0185] 第五十六EEE涉及根据第四十九至第五十五EEE中任一项所述的第一装置，进一步适用于基于所述姿势信息和先前姿势信息获得预测的姿势信息，并将所述预测的姿势信息传输到所述第二装置。

[0186] 第五十七EEE涉及根据第五十六EEE所述的第一装置，其中，所述预测的姿势信息被预测来对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时进行估计。

[0187] 第五十八EEE涉及根据第四十九至第五十七EEE中任一项所述的第一装置，其中，所述经渲染的媒体内容是以未压缩的形式从所述第二装置接收的。

[0188] 第五十九EEE涉及根据第四十九至第五十八EEE中任一项所述的第一装置，其中，所述第一装置进一步适用于对经编码的经渲染媒体内容进行解码。

[0189] 第六十EEE涉及根据第五十七EEE或包括第五十七EEE的特征的任何EEE所述的第一装置，其中，对预期要由所述第一装置处理所述经渲染的媒体内容以进行再现的定时的所述估计包括：对编码和解码所述经渲染的音频内容所需的时间的估计和/或对将所述经渲染的媒体内容传输到所述第一装置所需的时间的估计。

[0190] 第六十一EEE涉及根据第四十九至第六十EEE中任一项所述的第一装置，进一步适用于：将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较，并基于所述比较结果更新所述经渲染的媒体内容。

[0191] 第六十二EEE涉及根据第四十九至第六十一EEE中任一项所述的第一装置，进一步适用于：接收指示所述经渲染的媒体内容如何响应于所述姿势信息的变化而变化的梯度信息、以及来自所述第二装置的所述经渲染的媒体内容；将已经用于渲染所述媒体内容的所述姿势信息与当前姿势信息进行比较；以及基于所述梯度信息和所述比较结果更新所述经渲染的媒体内容。

[0192] 第六十三EEE涉及根据第四十九至第六十二EEE中任一项所述的第一装置，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第一装置进一步适用于将指示所述第一装置所处的环境的声学特性的环境信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述环境信息。

[0193] 第六十四EEE涉及根据第四十九至第六十三EEE中任一项所述的第一装置，其中，所述媒体内容包括音频内容，并且所述经渲染的媒体内容包括经渲染的音频内容，所述第一装置进一步适用于将指示所述用户或所述用户的一部分的形貌的形貌信息传输到所述第二装置，并且所述渲染所述媒体内容进一步基于所述形貌信息。

[0194] 第六十五EEE涉及一种用于渲染音频内容的方法。所述方法包括：由发送方(S)装置接收用户位置和/或取向数据，并发送通常由基于对象-5或HOA表示得到的相应的预渲染内容。

[0195] 第六十六EEE涉及根据第六十五EEE所述的方法，其中，由所述发送方生成的所述预渲染信号可以是双耳、FOA/B格式、HOA、或任何类型的基于声道的渲染。

[0196] 第六十七EEE涉及根据第六十五或第六十六EEE所述的方法，进一步包括传输未压缩的预渲染内容。

[0197] 第六十八EEE涉及根据第六十五或第六十六EEE所述的方法，进一步包括对所述预渲染的内容进行编码并传输经编码的预渲染内容。

[0198] 第六十九EEE涉及根据第六十五至第六十八EEE中任一项所述的方法，进一步包括由接收方接收所述预渲染内容。

[0199] 第七十EEE涉及根据第六十五至第六十九EEE中任一项所述的方法，进一步包括由接收方对预渲染的、预编码的双耳化内容进行解码。

[0200] 第七十一EEE涉及根据第六十五至第七十EEE中任一项所述的方法，其中，所述用户位置和/或取向数据包括指示所述用户在世界空间中的位置和取向的本地姿势。

[0201] 第七十二EEE涉及根据第六十五至第七十一EEE中任一项所述的方法，其中，所述用户位置数据从接收方传输到所述发送方。

[0202] 第七十三EEE涉及根据第六十五到第七十二EEE中任一项所述的方法，进一步包括将用于预渲染的双耳化内容的所述用户位置数据传输回所述接收方。

[0203] 第七十四EEE涉及根据第六十五至第七十三EEE中任一项所述的方法，进一步包括基于所接收的用户位置数据和所述本地位置数据来外推所述预渲染的内容以确定更新的内容。

[0204] 第七十五EEE涉及根据第六十五至第七十四EEE中任一项所述的方法，进一步包括传输关于所述用户的形貌数据(例如，头部大小、头部形状)以用于个性化双耳处理。

[0205] 第七十六EEE涉及根据第六十五至第七十五EEE中任一项所述的方法，进一步包括传输关于BRIR函数和/或房间表征的数据。

[0206] 第七十七EEE涉及根据第六十五至第七十六EEE中任一项所述的方法，进一步包括基于确定所述内容以收听者不可知的方式(例如，不包括HRTF)传输，然后在所述接收方侧执行双耳渲染和个性化。

[0207] 第七十八EEE涉及根据第六十五至第七十七EEE中任一项所述的方法，进一步包括在时间点t1提供用户位置和/或取向数据P(t0)。

[0208] 第七十九EEE涉及根据第六十七EEE所述的方法，其中，所述未压缩的预渲染内容是双耳化的未压缩的预渲染内容。

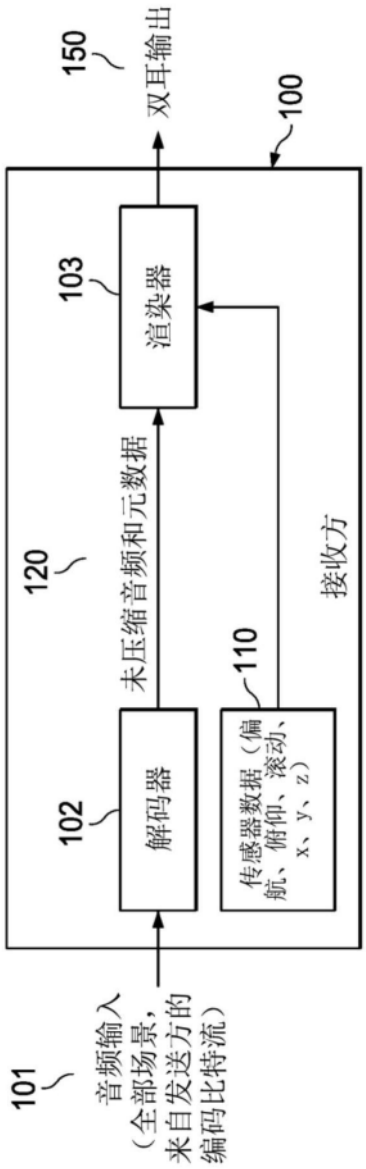

150
双耳输出
100
103
渲染器
未压缩音频和元数据
120
接收方
110
102
解码器
传感器数据（偏航、俯仰、滚动、x、y、z）
101
音频输入
（全部场景，来自发送方的编码比特流）

图1

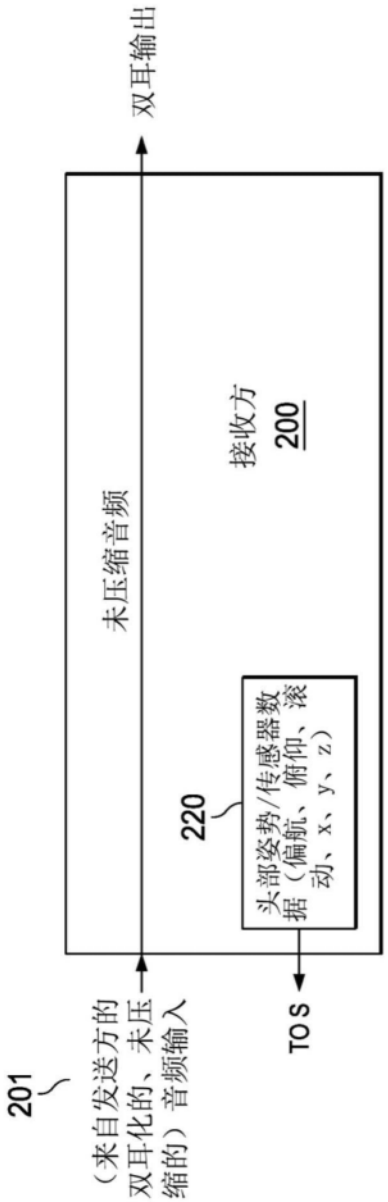

双耳输出
未压缩音频
接收方
200
220
头部姿势/传感器数据（偏航、俯仰、滚动、x、y、z）
TOS
201
（来自发送方的双耳化的、未压缩的）音频输入

图2

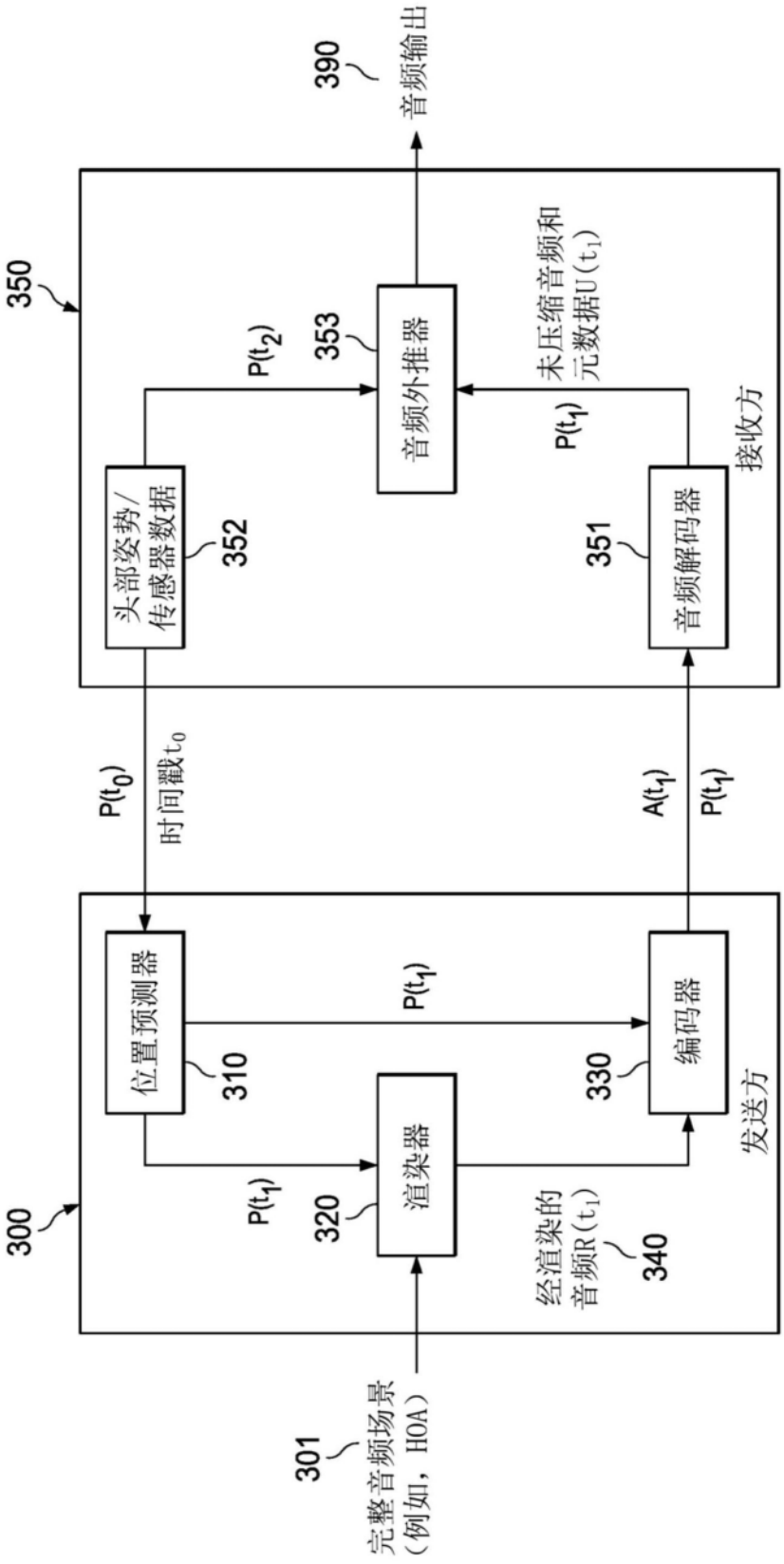

390
音频输出
350
P(t2)
353
音频外推器
未压缩音频和
元数据U(t1)
P(t1)
头部姿势/
传感器数据
352
351
音频解码器
接收方
P(t0)
时间戳t0
A(t1)
P(t1)
位置预测器
310
P(t1)
330
编码器
渲染器
320
P(t1)
300
经渲染的
音频R(t1)
340
发送方
301
完整音频场景
（例如，HOA）

图3

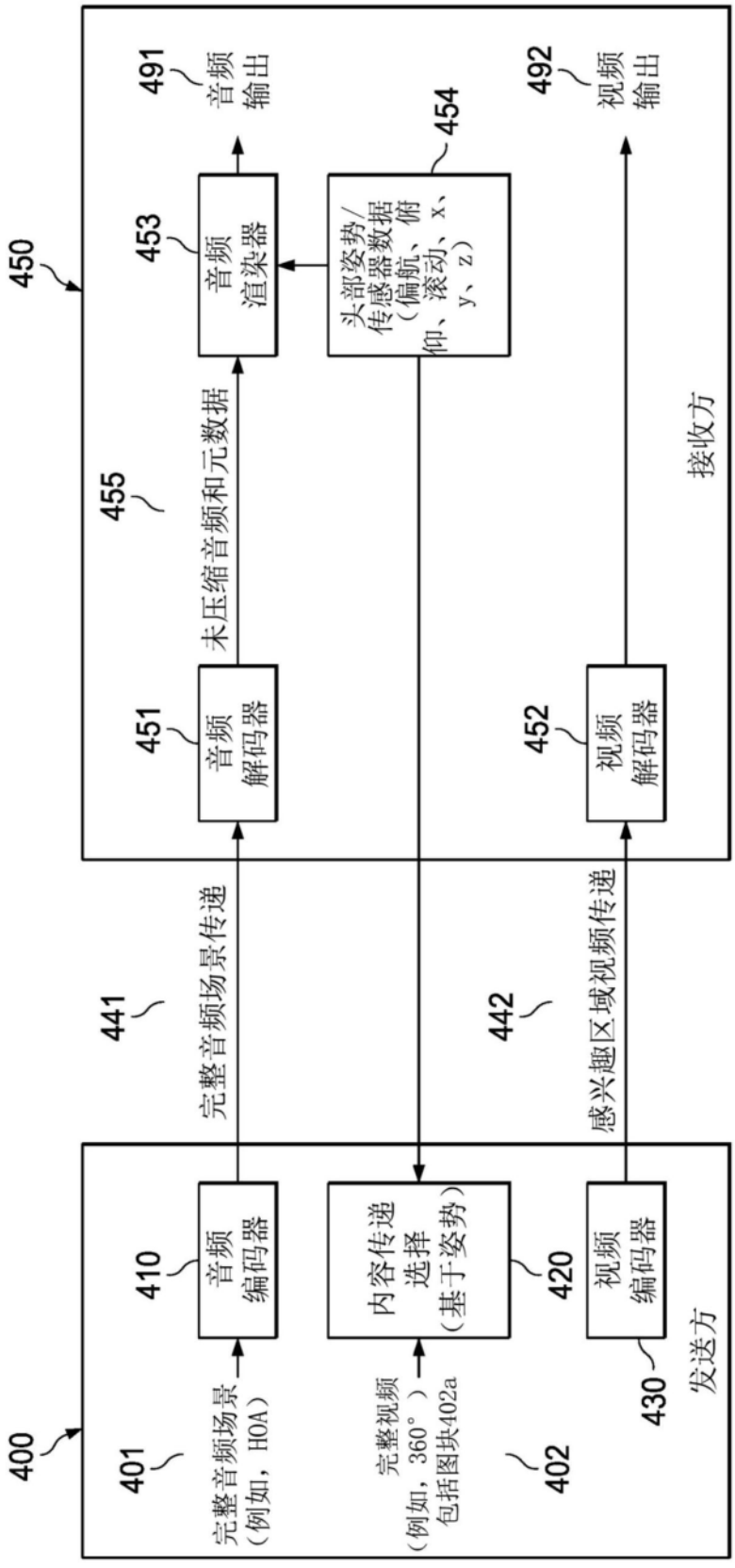

400
401
完整音频场景
（例如，HOA）
410
音频
编码器
402
完整视频
（例如，360°）
包括图块402a
420
内容传递
选择
（基于姿势）
430
视频
编码器
发送方
441
完整音频场景传递
442
感兴趣区域视频传递
450
451
音频
解码器
455
未压缩音频和元数据
453
音频
渲染器
454
头部姿势/
传感器数据
（偏航、俯
仰、滚动、x、
y、z）
452
视频
解码器
491
音频
输出
492
视频
输出
接收方

图4

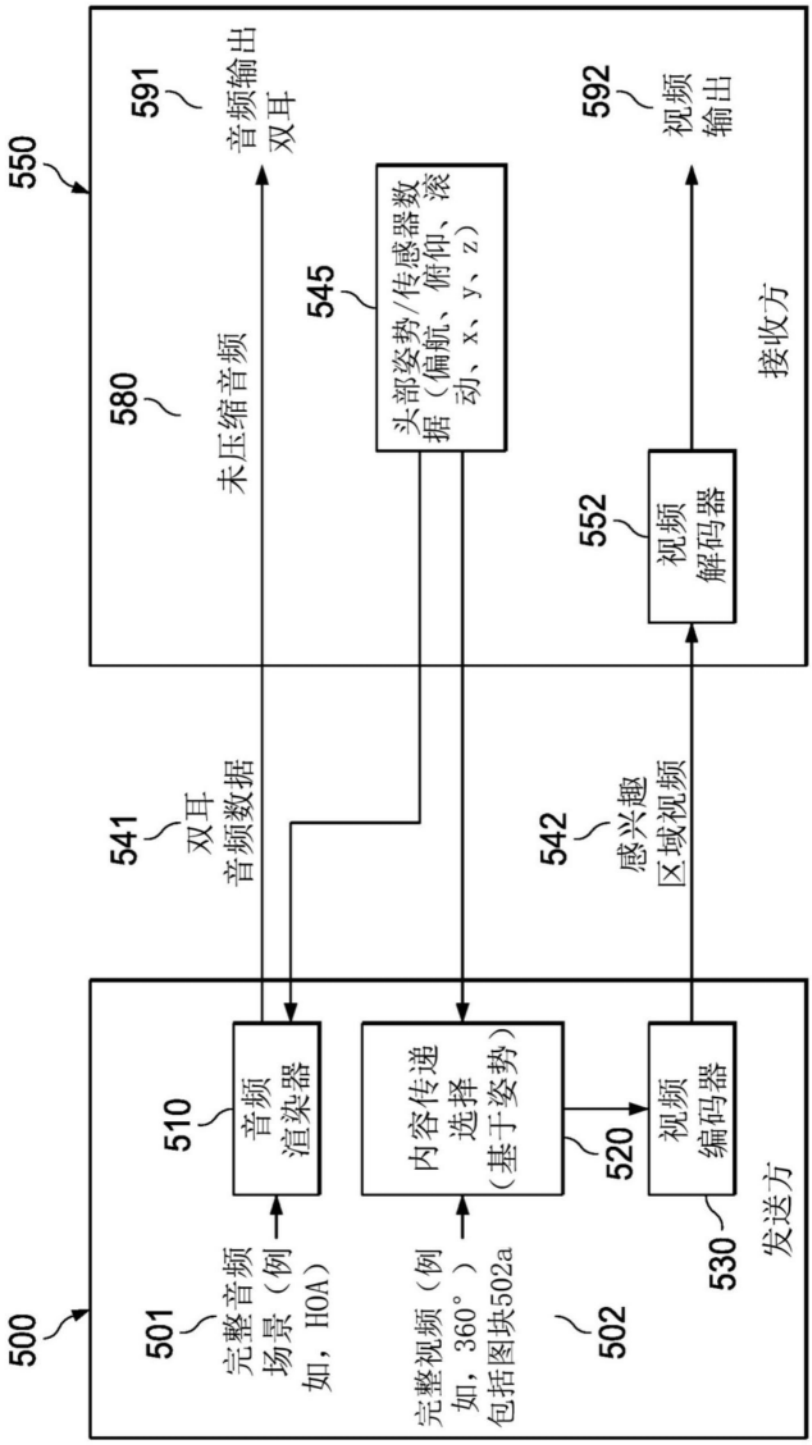

500
发送方
501
完整音频场景（例如，HOA）
510
音频渲染器
完整视频（例如，360°）包括图块502a
502
内容传递选择（基于姿势）
520
视频编码器
530
541
双耳音频数据
542
感兴趣区域视频
550
接收方
580
未压缩音频
591
音频输出双耳
545
头部姿势/传感器数据（偏航、俯仰、滚动、x、y、z）
552
视频解码器
592
视频输出

图5

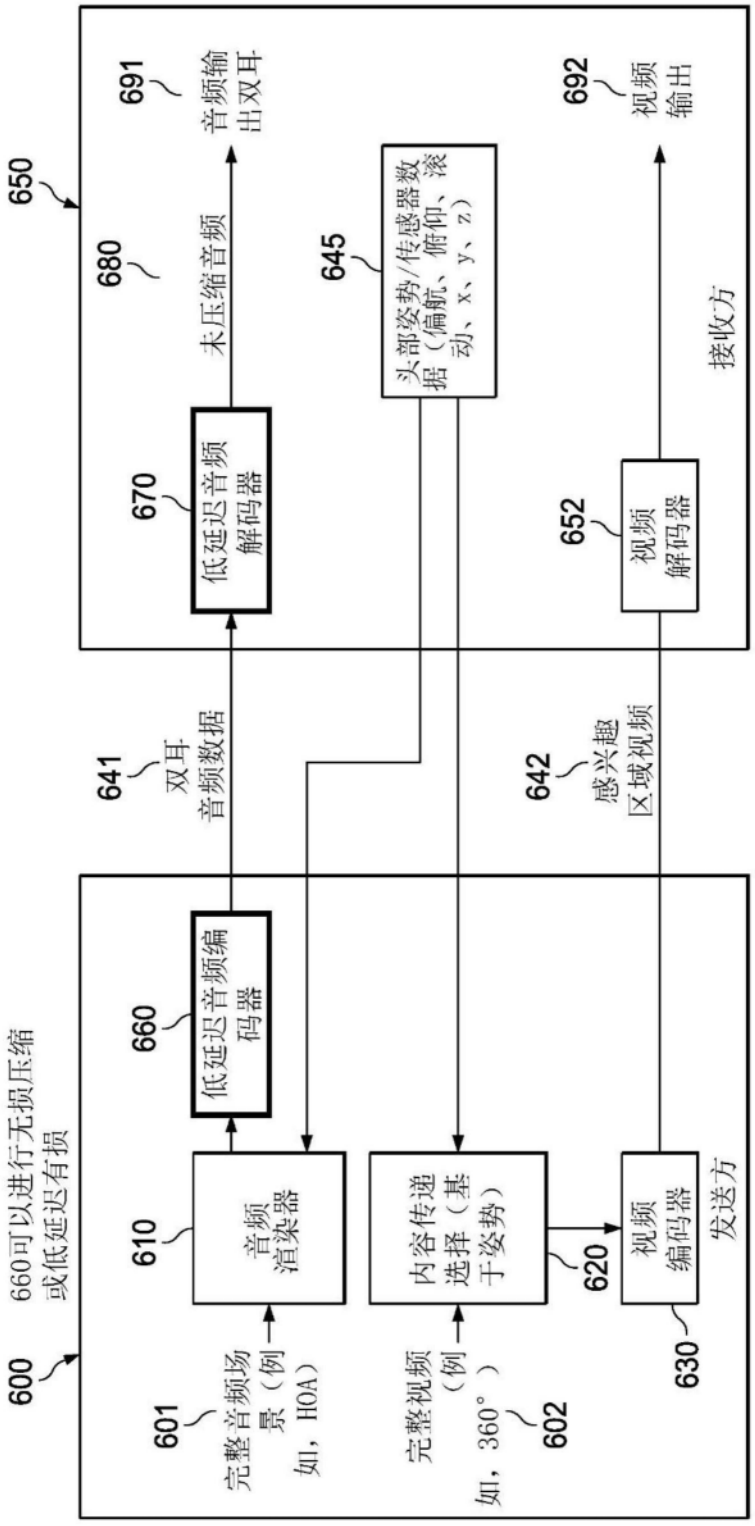

600
660可以进行无损压缩
或低延迟有损
601
完整音频场景（例如，HOA）
602
完整视频（例如，360°）
610
音频渲染器
660
低延迟音频编码器
620
内容传递选择（基于姿势）
630
视频编码器
发送方
641
双耳音频数据
642
感兴趣区域视频
650
670
低延迟音频解码器
680
未压缩音频
691
音频输出双耳
645
头部姿势/传感器数据（偏航、俯仰、滚动、x、y、z）
652
视频解码器
692
视频输出
接收方

图6

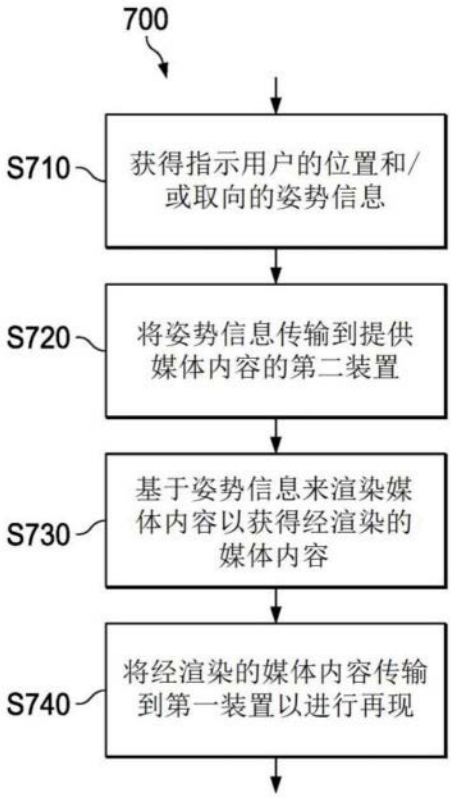

700
S710
获得指示用户的位置和/或取向的姿势信息
S720
将姿势信息传输到提供媒体内容的第二装置
S730
基于姿势信息来渲染媒体内容以获得经渲染的媒体内容
S740
将经渲染的媒体内容传输到第一装置以进行再现

图7

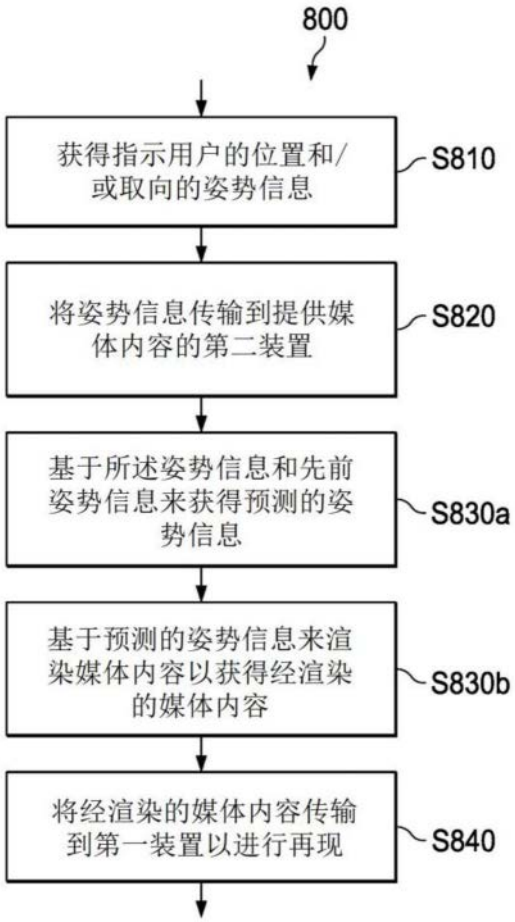

800
获得指示用户的位置和/或取向的姿势信息
S810
将姿势信息传输到提供媒体内容的第二装置
S820
基于所述姿势信息和先前姿势信息来获得预测的姿势信息
S830a
基于预测的姿势信息来渲染媒体内容以获得经渲染的媒体内容
S830b
将经渲染的媒体内容传输到第一装置以进行再现
S840

图8

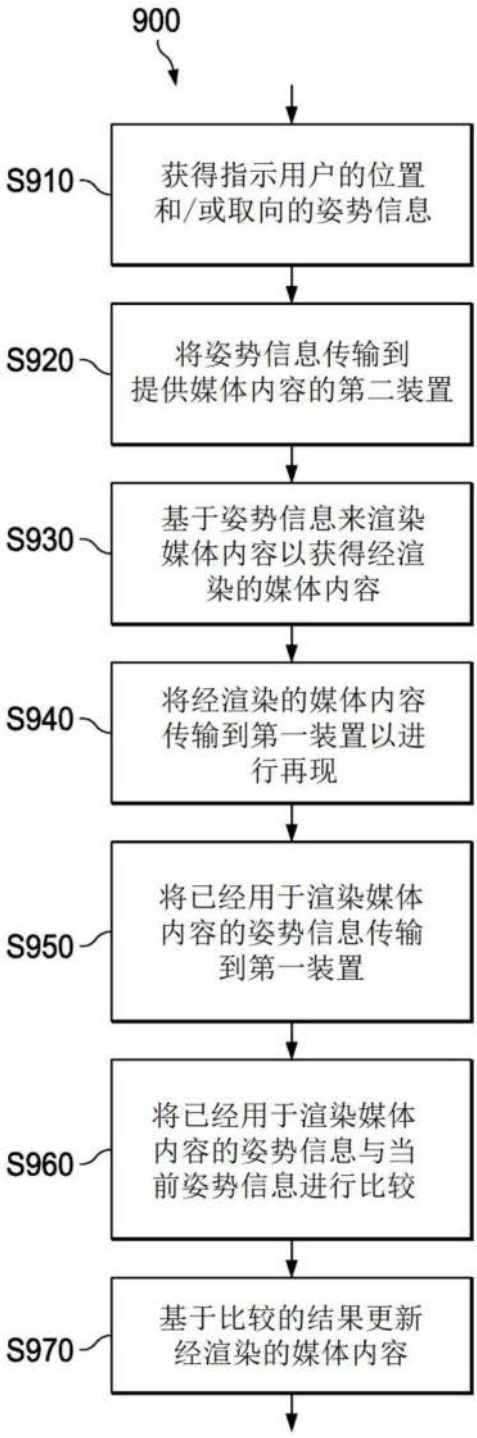

900
S910
获得指示用户的位置和/或取向的姿势信息
S920
将姿势信息传输到提供媒体内容的第二装置
S930
基于姿势信息来渲染媒体内容以获得经渲染的媒体内容
S940
将经渲染的媒体内容传输到第一装置以进行再现
S950
将已经用于渲染媒体内容的姿势信息传输到第一装置
S960
将已经用于渲染媒体内容的姿势信息与当前姿势信息进行比较
S970
基于比较的结果更新经渲染的媒体内容

图9

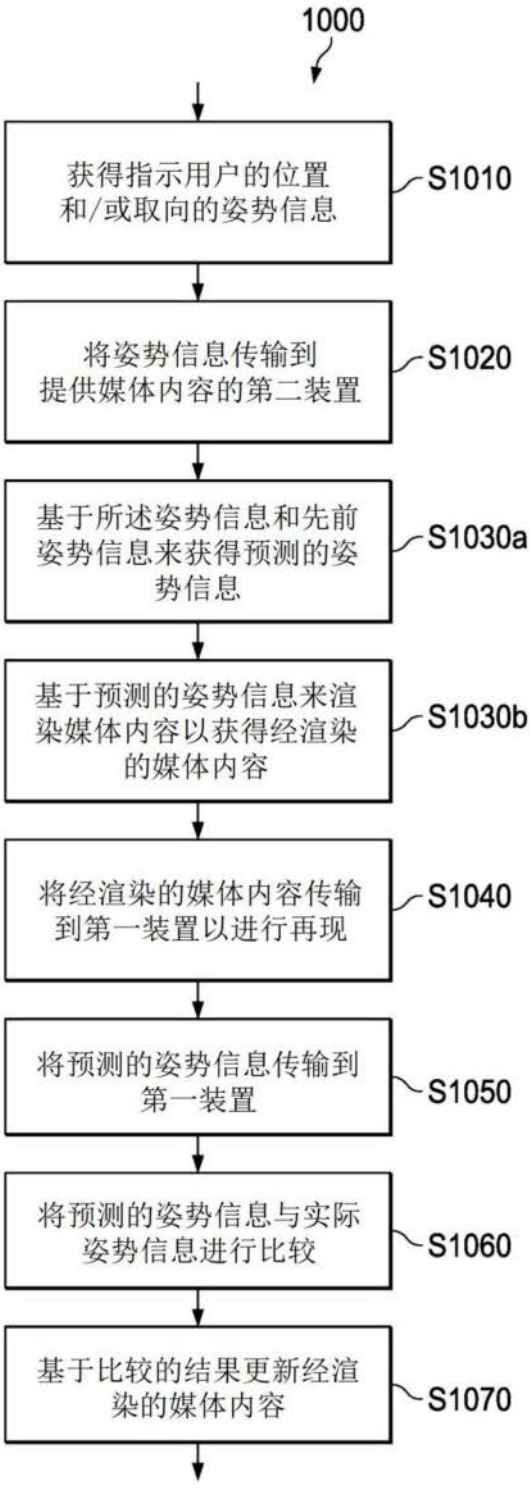

1000
获得指示用户的位置和/或取向的姿势信息
S1010
将姿势信息传输到提供媒体内容的第二装置
S1020
基于所述姿势信息和先前姿势信息来获得预测的姿势信息
S1030a
基于预测的姿势信息来渲染媒体内容以获得经渲染的媒体内容
S1030b
将经渲染的媒体内容传输到第一装置以进行再现
S1040
将预测的姿势信息传输到第一装置
S1050
将预测的姿势信息与实际姿势信息进行比较
S1060
基于比较的结果更新经渲染的媒体内容
S1070

图10