



(12)发明专利

(10)授权公告号 CN 104995681 B

(45)授权公告日 2017. 10. 31

(21)申请号 201480008260.1

(22)申请日 2014.02.12

(65)同一申请的已公布的文献号
申请公布号 CN 104995681 A

(43)申请公布日 2015.10.21

(30)优先权数据
61/765,556 2013.02.15 US
13/831,018 2013.03.14 US

(85)PCT国际申请进入国家阶段日
2015.08.11

(86)PCT国际申请的申请数据
PCT/US2014/016059 2014.02.12

(87)PCT国际申请的公布数据
W02014/127019 EN 2014.08.21

(73)专利权人 高通股份有限公司
地址 美国加利福尼亚州

(72)发明人 向佩

(74)专利代理机构 北京律盟知识产权代理有限公司 11287
代理人 宋献涛

(51)Int.Cl.
G11B 27/031(2006.01)
G11B 27/28(2006.01)
G10L 19/008(2006.01)
H04S 3/00(2006.01)
H04S 7/00(2006.01)

(56)对比文件
US 2012062700 A1, 2012.03.15,
WO 2010140254 A1, 2010.12.09,
US 2003053680 A1, 2003.03.20,
审查员 樊云

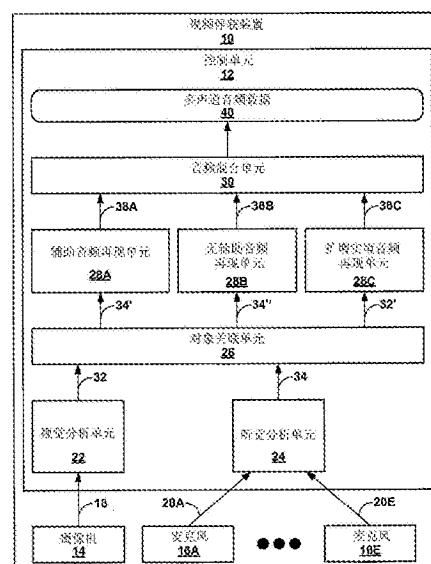
权利要求书4页 说明书20页 附图11页

(54)发明名称

多声道音频数据的视频分析辅助产生

(57)摘要

一般来说,本发明描述用于俘获多声道音频数据的技术。一种包括一个或多个处理器的装置可经配置以实施所述技术。所述处理器可分析所俘获音频数据来识别音频对象并分析在俘获所述音频数据的同时所俘获的视频数据来识别视频对象。所述处理器随后可使所述音频对象中的至少一者与所述视频对象中的至少一者相关联并且基于音频对象中的所述至少一者与所述视频对象中的所述至少一者的所述关联从所述音频数据产生多声道音频数据。



1. 一种产生多声道音频数据的方法,该方法包括:分析用装置俘获的音频数据,所述分析包括

执行所述音频数据的听觉场景分析来识别一或多个音频对象,并识别描述所述一或多个音频对象的音频元数据;

分析在所述俘获所述音频数据的同时用所述装置俘获的视频数据,所述分析包括执行所述视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据,所述音频元数据及所述视频元数据使用共同格式;

用所述装置使所述一或多个音频对象中的至少一者与所述一或多个视频对象中的至少一者相关联;及

通过基于描述所述音频对象中的所述至少一者的所述音频元数据与描述所述视频对象中的所述相关联的至少一者的所述视频元数据之间的相关度,将所述音频对象中的所述至少一者再现为扩散音频对象,用所述装置基于所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者的所述关联从所述音频数据产生所述多声道音频数据。

2. 根据权利要求1所述的方法,

其中所述音频元数据包括所述对应音频对象的位置、形状、速度及位置可信程度中的一或多者,并且

其中所述视频元数据包括所述对应音频对象的位置、形状、速度及位置可信程度中的一或多者。

3. 根据权利要求2所述的方法,

其中使所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者相关联包括:

将所述一或多个音频对象中的每一者分类为一种类型的音频对象;

将所述一或多个视频对象中的每一者分类为一种类型的视频对象;

确定所述音频对象中的所述至少一者的所述类型为与所述视频对象中的所述至少一者相同的类型;及

响应于所述确定所述一或多个音频对象中的所述至少一者的所述类型与所述一或多个视频对象中的所述至少一者的所述类型相同,使所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者相关联。

4. 根据权利要求2所述的方法,其中产生所述多声道音频数据包括:

确定所述一或多个音频对象中的所述至少一者的所述音频元数据与所述一或多个视频对象中的所述至少一者的所述视频元数据之间的所述相关度,所述一或多个视频对象中的所述至少一者与所述一或多个音频对象中的所述至少一者相关联;

基于所述确定的相关度产生用于所述一或多个音频对象中的所述至少一者的组合元数据,所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者相关联;及

基于针对所述一或多个音频对象中的所述至少一者产生的所述组合元数据将所述一或多个音频对象中的所述至少一者再现在所述多声道音频数据的一或多个前景声道中。

5. 根据权利要求1所述的方法,

其中所述一或多个音频对象中的所述至少一者包括所述一或多个音频对象中的第一者，

其中所述方法进一步包括确定所述一或多个音频对象中的第二者不与所述一或多个视频对象中的任一者相关联，并且

其中产生所述多声道音频数据包括产生所述多声道音频数据以使得所述音频对象中的所述第二者起源于所述多声道音频数据的一或多个背景声道。

6. 根据权利要求5所述的方法，其中产生所述多声道音频数据包括产生所述多声道音频数据以使得所述音频对象中的所述第二者作为扩散音频对象起源于所述多声道音频数据的所述一或多个背景声道。

7. 根据权利要求1所述的方法，

其中所述一或多个视频对象中的所述至少一者包括所述一或多个视频对象中的第一者，

其中所述方法进一步包括：

确定所述视频对象中的所述一或多者中的第二者不与所述一或多个音频对象中的任一者相关联；

响应于确定所述视频对象中的所述一或多者中的所述第二者不与所述一或多个音频对象中的任一者相关联，从音频库检索原本与所述一或多个视频对象中的所述第二者相关联的参考音频对象；及

基于所述一或多个视频对象中的所述第二者再现所述参考音频对象来产生所述多声道音频数据的至少一部分。

8. 根据权利要求1所述的方法，

其中所述音频元数据以与用于定义所述视频元数据的文本格式共同的所述文本格式来加以定义。

9. 根据权利要求1所述的方法，

其中产生所述多声道音频数据包括：

确定描述所述音频对象中的所述至少一者的所述音频元数据与描述所述视频对象中的所述关联者的所述视频元数据之间的所述相关度；及

基于在产生所述多声道音频数据时所确定的所述相关度将所述音频对象中的所述至少一者再现为扩散音频对象。

10. 根据权利要求1所述的方法，其在移动电话上加以执行。

11. 一种经配置以产生多声道音频数据的装置，所述装置包括：

一或多个处理器，其经配置以：至少部分通过执行所述音频数据的听觉场景分析以识别音频对象及描述所述音频对象的音频元数据来分析音频数据，至少部分通过执行视频数据的视觉场景分析以识别视频对象及描述所述视频对象的视频元数据来分析所述视频数据，使所述音频对象与所述视频对象相关联，及通过基于描述所述音频对象的所述音频元数据与描述所述视频对象中的所述相关联的至少一者的所述视频元数据之间的相关度，将所述音频对象中的所述至少一者再现为扩散音频对象，来基于所述音频对象与所述视频对象的所述关联从所述音频数据产生所述多声道音频数据。

12. 根据权利要求11所述的装置，其中所述音频元数据包括所述音频对象的大小及所

述音频对象的位置。

13. 根据权利要求11所述的装置, 其中所述视频元数据包括所述视频对象的大小及所述视频对象的位置。

14. 根据权利要求11所述的装置, 其中所述一或多个处理器经进一步配置以至少部分产生包括大小及位置中的一或多者的组合元数据。

15. 根据权利要求14所述的装置,

其中所述音频元数据包含位置元数据,

其中所述视频元数据包含位置元数据,

其中产生所述组合元数据包括:

比较所述音频元数据的所述位置元数据与所述视频元数据的所述位置元数据来确定所述相关度;

基于对所述相关度是否超过可信度阈值的确定产生所述组合元数据的位置元数据。

16. 一种经配置以产生多声道音频数据的装置, 所述装置包括:

用于分析用装置俘获的音频数据的装置, 所述分析包括执行所述音频数据的听觉场景分析来识别一或多个音频对象, 并识别描述所述一或多个音频对象的音频元数据;

用于分析在所述俘获所述音频数据的同时用所述装置俘获的视频数据的装置, 所述分析包括执行所述视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据, 所述音频元数据及所述视频元数据使用共同格式;

用于使所述一或多个音频对象中的至少一者与所述一或多个视频对象中的至少一者相关联的装置; 及

用于通过基于描述所述音频对象中的所述至少一者的所述音频元数据与描述所述视频对象中的所述相关联的至少一者的所述视频元数据之间的相关度, 将所述音频对象中的所述至少一者再现为扩散音频对象, 基于所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者的所述关联从所述音频数据产生所述多声道音频数据的装置。

17. 根据权利要求16所述的装置,

其中所述用于使得相关联的装置包括:

用于基于第一音频对象的数据分量与第一视频对象的数据分量的第一比较识别与所述第一视频对象对应物相关联的所述第一音频对象的装置;

用于基于第二音频对象的数据分量与第二视频对象的数据分量的第二比较识别不与所述第二视频对象对应物相关联的所述第二音频对象的装置, 及

其中所述装置进一步包括:

用于在第一区域中再现所述第一音频对象的装置;

用于在第二区域中再现所述第二音频对象的装置; 及

用于基于组合在所述第一区域中再现的所述第一音频对象及在所述第二区域中再现的所述第二音频对象产生所述音频输出信号的装置。

18. 根据权利要求17所述的装置, 其中所述第一音频对象的所述数据分量包括位置及大小中的一者。

19. 根据权利要求17所述的装置, 其中所述第一视频对象数据的所述数据分量包括位

置及大小中的一者。

20. 根据权利要求17所述的装置, 其中所述第二音频对象的所述数据分量包括位置及大小中的一者。

21. 根据权利要求17所述的装置, 其中所述第二视频对象的所述数据分量包括位置及大小中的一者。

22. 根据权利要求17所述的装置, 其中所述第一区域及第二区域为在音频前景内的不同区域或在音频背景内的不同区域。

23. 根据权利要求17所述的装置, 其中所述第一区域及第二区域为音频前景内的同一区域或音频背景内的同一区域。

24. 根据权利要求17所述的装置, 其中所述第一区域在音频前景内, 并且所述第二区域在音频背景内。

25. 根据权利要求17所述的装置, 其中所述第一区域在音频背景内, 并且所述第二区域在音频前景内。

26. 根据权利要求17所述的装置, 其进一步包括:

用于确定所述第一比较是否在可信度区间外的装置; 及

用于基于对所述第一比较是否在所述可信度区间外的所述确定加权所述第一音频对象的所述数据分量及第一视频对象的所述数据分量的装置。

27. 根据权利要求26所述的装置, 其中所述用于加权的装置包括用于对所述第一音频对象数据的所述数据分量及所述第一视频对象的所述数据分量求平均的装置。

28. 根据权利要求17所述的装置, 其进一步包括用于基于所述第一比较及所述第二比较中的一或多者分配不同数目的位的装置。

29. 一种非暂时性计算机可读存储媒体, 其上存储有在执行时致使装置的一或多个处理器执行以下操作的指令:

分析用装置俘获的音频数据, 所述分析包括执行所述音频数据的听觉场景分析来识别一或多个音频对象, 并识别描述所述一或多个音频对象的音频元数据;

分析在所述俘获所述音频数据的同时用所述装置俘获的视频数据, 所述分析包括执行所述视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据, 所述音频元数据及所述视频元数据使用共同格式;

使所述一或多个音频对象中的至少一者与所述一或多个视频对象中的至少一者相关联; 及

通过基于描述所述音频对象中的所述至少一者的所述音频元数据与描述所述视频对象中的所述相关联的至少一者的所述视频元数据之间的相关度, 将所述音频对象中的所述至少一者再现为扩散音频对象, 基于所述一或多个音频对象中的所述至少一者与所述一或多个视频对象中的所述至少一者的所述关联从所述音频数据产生多声道音频数据。

多声道音频数据的视频分析辅助产生

[0001] 本申请案请求2013年2月15日申请的美国临时申请案第61/765,556号的权利。

技术领域

[0002] 本发明涉及俘获音频数据,并且更确切地说涉及俘获多声道音频数据。

背景技术

[0003] 通常,视频俘获装置(例如视频可携式摄像机、平板或平板触摸计算机、移动电话(包含所谓的“智能电话”)、个人游戏装置、个人媒体装置和类似者)以摄像机为特征而以给定帧速率俘获一系列图像来产生视频数据。通常,这些视频俘获装置以麦克风为特征,以俘获视频数据中所描绘的场景的单声道音频数据。更复杂的视频俘获装置可能以两个或两个以上麦克风为特征,以增加能够俘获的音频声道(以单声道音频数据来自单个声道)的数目。这些更复杂的视频录制装置可包含至少两个麦克风来俘获立体声音频数据(其指代具有左右声道的音频数据)。

[0004] 鉴于更多地采用所谓的智能电话,智能电话逐渐变成俘获视频数据的主导方式。通常,由于智能电话的本质及其作为音频通信装置的用途,智能电话可包含两个、三个、四个或甚至五个麦克风。智能手机可在电话呼叫、视频会议或包含音频通信的其它形式通信期间出于噪音消除的目的而使用附加麦克风。尽管智能电话以大量麦克风为特征,但通常不使用这些麦克风来俘获除立体声音频数据外的多声道音频数据,因为这些麦克风通常放置于智能电话上限制其充分俘获除立体声音频数据外的任何数据的能力的位置中。

发明内容

[0005] 一般来说,本发明描述视频俘获装置可使用视频分析来协助多声道音频数据的俘获的技术。视频俘获装置可促进使用视频场景分析(或计算机视觉)技术产生环绕声音频数据(通常具有五个或五个以上声道)。在一些实例中,视频俘获装置可俘获音频数据和视频数据,从而处理视频数据来识别视频对象,同时也处理音频数据来识别音频对象。视频俘获装置可执行视频场景分析技术来识别所述视频对象并产生关于所述对象的各种元数据。视频俘获装置也可执行听觉场景分析,尝试识别音频对象及关于所述对象的各种元数据。通过比较所述对象,视频俘获装置可识别很可能为音频对象的来源的所述视频对象。

[0006] 鉴于视频分析技术可更确切地识别视频对象(相比仅音频对象)相对于视频俘获装置的位置,相比仅依赖于通常不准确的波束形成技术,视频俘获装置可更好地本地化音频对象。随后可使用将音频对象更好地本地化到一或多个前声道的分贝差来将所述音频对象再现到一或多个声道,借此使得能够更好地产生环绕声音频数据(相比由常规视频俘获装置产生的环绕声音频数据)。

[0007] 在一个方面中,一种方法包括分析用装置俘获的音频数据以识别一或多个音频对象及分析用装置在俘获音频数据的同时所俘获的视频数据来识别一或多个视频对象。所述方法进一步包括:使一或多个音频对象中的至少一者与一或多个视频对象中的至少一者相

关联,及基于一或多个音频对象中的至少一者与一或多个视频对象中的至少一者的关联从音频数据产生多声道音频数据。

[0008] 在另一方面中,一种装置包括一或多个处理器,所述处理器经配置以获得音频对象、获得视频对象、使音频对象与视频对象相关联、比较所述音频对象与相关联的视频对象及基于音频对象与相关联视频对象之间的比较再现音频对象。

[0009] 在另一方面中,产生音频输出信号的装置包括:用于基于第一音频对象的数据分量与第一视频对象的数据分量的第一比较识别与第一视频对象对应物相关联的第一音频对象的装置,及用于基于第二音频对象的数据分量与第二视频对象的数据分量的第二比较识别不与第二视频对象对应物相关联的第二音频对象的装置。所述装置进一步包括:用于在第一区域中再现第一音频对象的装置,用于在第二区域中再现第二音频对象的装置,及用于基于组合第一区域中的再现第一音频对象和第二区域中的再现第二音频对象产生音频输出信号的装置。

[0010] 在另一方面中,一种非暂时性计算机可读存储媒体上存储有在执行时致使装置的一或多个处理器执行以下操作的指令:分析用装置俘获的音频数据来识别一或多个音频对象,分析用装置在俘获所述音频数据的同时所俘获的视频数据来识别一或多个视频对象,使一或多个音频对象中的至少一者与一或多个视频对象中的至少一者相关联,及基于一或多个音频对象中的至少一者与一或多个视频对象中的至少一者的关联从所述音频数据产生多声道音频数据。

[0011] 在附图及以下描述中阐述所述技术的一或多个方面的细节。这些技术的其它特征、目标及优点将从所述描述及图式以及权利要求书而显而易见。

附图说明

[0012] 图1A为说明执行本发明中所描述技术的实例视频俘获装置10的各种视图的图。

[0013] 图1B为更详细说明执行本发明中所描述技术的视频俘获装置的方块图。

[0014] 图2A至2D为说明由图1的视频俘获装置在根据本发明中所描述技术使视频对象与音频对象相关联时执行的操作的图。

[0015] 图3为更详细说明图1B的辅助音频再现单元的方块图。

[0016] 图4为说明由图1B的实例中所展示的视频俘获装置的摄像机所俘获并且根据本发明中所描述技术经处理的场景的图。

[0017] 图5为说明由图1B的实例中所展示的视频俘获装置的摄像机所俘获并且根据本发明中所描述技术的扩增实境方面经处理的另一场景的图。

[0018] 图6为说明视频俘获装置在执行本发明中所描述技术时的示范性操作的流程图。

[0019] 图7为说明可如何根据本发明中所描述技术在多声道音频数据的前景和背景中再现各种音频对象的图。

具体实施方式

[0020] 图1A为说明执行本发明中所描述技术的实例视频俘获装置10的各种视图8A至8C(分别为前视图、后视图及侧视图)的图。视频俘获装置10可表示能够俘获视频及音频数据的任何类型的装置,例如视频可携式摄像机、平板或平板触摸计算机、移动电话(包含所谓

的“智能手机”)、个人游戏装置、个人媒体装置和类似者。出于说明的目的,假定视频俘获装置10表示智能手机。尽管关于特定类型的装置(即智能手机)进行了描述,但在本发明中,所述技术可由能够俘获视频数据及多声道音频数据的任何类型的装置实施。

[0021] 在图1A的实例中,从三个不同视图8A至8C展示视频俘获装置10。视图8A从前面展示视频俘获装置10。视图8B从后面展示视频俘获装置10。视图8C从侧面展示视频俘获装置10。

[0022] 如视图8A中所展示,视频俘获装置10包含耳机9、扬声器11A、11B及麦克风16A、16B和16E。耳机9表示当用靠近用户耳朵的装置10收听音频时用于再现声音或音频数据的小型扬声器。扬声器11A和11B各自表示当用距用户较远的装置10收听音频时(例如当用于再现音乐、观看视频或用作免提电话时)用于再现声音音频数据音频数据的扬声器。扬声器11A可被称为左扬声器11A(或“扬声器L”),原因在于扬声器11A可再现多声道音频数据的左声道。扬声器11B可被称为右扬声器11A(或“扬声器R”),原因在于扬声器11B可再现多声道音频数据的右声道。下文更详细描述麦克风16A、16B及16E。

[0023] 如视图8B中所展示,在一个实例中,视频俘获装置10还包含摄像机14及麦克风16C和16D。摄像机14可表示能够俘获图像的任何类型的装置。摄像机14可能以给定速率(其通常被称为“帧速率”)俘获一系列图像以形成视频数据。摄像机14可包含可促进光的俘获以产生或以其它方式产生图像的透镜及其它组件。摄像机14也可与闪光灯或其它光产生元件(其在图1A的实例中未展示)介接,其中,在一些情况下,摄像机14可与闪光灯集成。在假定的智能手机的上下文中,摄像机14通常包括数码摄像机,其包含光感应传感器(例如互补型金属氧化物半导体(CMOS)光图像传感器或电荷耦合装置(CCD)图像传感器)来感应进入透镜的光的亮度和色度,这与电影摄像机中常见的用于感应光的赛璐璐媒体形成对照。摄像机14可俘获光并产生一系列图像,其在下文的图1B的实例中展示为视频数据18。

[0024] 麦克风16A至16E(“麦克风16”)可各自表示能够俘获音频数据的任何类型的装置。麦克风16可一般涉及能够将声音转换成电信号的任何类型的声电换能器或传感器。存在大量不同类型的麦克风,其中的每一者在不同类型俘获声音的方式方面不同。提供几个实例,麦克风16可包含动态麦克风(其指代使用电磁感应俘获声音的麦克风)、调相器麦克风(其指代使用电容变化俘获声音的麦克风)及压电麦克风。虽然展示为并入视频俘获装置10内或在视频俘获装置10内部,麦克风16中的一或多者可在视频俘获装置10外部并通过有线连接或无线连接耦合到视频俘获装置10。麦克风16中的每一者可俘获单独音频数据20A至20E,如关于图1B的实例更详细展示。

[0025] 通常,视频俘获装置(例如视频可携式摄像机、平板或平板触摸计算机、移动电话(包含所谓的“智能电话”)、个人游戏装置、个人媒体装置和类似者)以摄像机为特征而以给定帧速率俘获一系列图像来产生视频数据。通常,这些视频俘获装置以麦克风为特征,以俘获视频数据中所描绘的场景的单声道音频数据。更复杂的视频俘获装置可能以两个或两个以上麦克风为特征,以增加能够俘获的声道(以单声道音频数据来自单个声道)的数目。这些更复杂的视频录制装置可包含至少两个麦克风来俘获立体声音频数据(其指代具有左右声道的音频数据)。

[0026] 三个或三个以上麦克风(例如图1A中展示为麦克风16的五个麦克风)可使得视频俘获装置能够执行被称作“波束形成”技术的操作,以促进具有前后及左右区别(或所谓的

音频数据的“声道”，例如前或中央声道、前左声道、前右声道、后左声道及后右声道)的环绕声音频的俘获。在俘获麦克风信号(其也可被称作“音频数据”)之后，智能手机可以算法方式形成至其它空间方向的空间波束(其可指代扩增某些方向的声音的过程)。通过用这些波束单独地过滤所俘获的声音，智能手机可产生不同输出环绕声声道。在一些实例中，智能手机可产生波束，使得波束区域与对应无波束区域之间的差异展现6dB的声级差异。作为一个实例，智能手机可基于这些波束产生5.1环绕声音频数据。

[0027] 尽管智能电话可使用波束形成技术俘获环绕音频并借此相比以仅一个或两个麦克风为特征的视频俘获装置俘获更真实的音频，但一些智能电话上的麦克风布置(例如图1A的实例的视图8A至8C中所展示的布置)时常不允许最好质量的环绕声音频。通常，针对转角的分贝差异并不十分重要。也就是说，当组合波束时的6dB差异并不产生多大差异，使得产生识别波束的声音并不在播放后显得十分本地化。当产生环绕声音频数据时，智能手机可将应为本本地化音频的音频放置在中央声道和前右声道中(例如当所述音频应对于前右声道更本地化时)。

[0028] 另外，鉴于一些前后麦克风(例如麦克风16B及16C)之间的接近度，智能电话可不能够充分区别前后音频。不能够充分区别前后音频可导致智能手机产生并不再现前后声道之间的音频的充分区别的环绕声或多声道音频数据。换句话说，前后声道可能听起来混乱，其中后部声音可通过前扬声器(通常，结合后部声音以使前后声音混合)再现，并且前部声音可通过后扬声器(通常，结合前部声音以使前后声音混合)再现。

[0029] 视频俘获装置10可实施本发明中所描述技术以促进更好地复制当俘获视频数据时听到的音频数据的环绕声或多声道音频数据的产生。为根据本发明中描述的技术产生这种多声道音频数据，视频俘获装置10可使用视频分析来协助多声道音频数据的俘获。视频俘获装置10可促进使用视频场景分析(或计算机视觉)技术产生多声道音频数据(通常具有五个或五个以上声道)。在一些实例中，视频俘获装置10可俘获音频数据和视频数据，从而处理视频数据来识别视频对象，同时也处理音频数据来识别音频对象。视频俘获装置10可执行视频场景分析技术来识别所述视频对象和关于所述对象的各种元数据。视频俘获装置10也可执行听觉场景分析，尝试识别音频对象及关于所述对象的各种元数据。通过比较所述对象，视频俘获装置可识别很可能为音频对象的来源的所述视频对象。

[0030] 鉴于视频分析技术可更确切地识别视频对象(相比仅音频对象)相对于视频俘获装置的位置，相比仅依赖于通常不准确的波束形成技术，视频俘获装置10可更好地本地化音频对象。随后可使用将音频对象更好地本地化到前声道中的一者的分贝差来将所述音频对象再现到一或多个声道，借此使得能够更好地产生环绕声或其它类型的多声道音频数据(相比由常规视频俘获装置产生的音频数据)。关于下图1B更详细地描述由视频俘获装置10执行的技术。

[0031] 图1B为更详细说明执行本发明中所描述技术的视频俘获装置10的方块图。在图1B的实例中，视频俘获装置10包含控制单元12、摄像机14及麦克风(“mic”)16A至16E(“麦克风16”或“mic 16”)。尽管为了易于说明目的在图1B的实例中未展示，但视频俘获装置10还可包含耳机9及扬声器11A和11B，以及执行通常与视频俘获装置10相关联的各种其它功能的额外模块、元件及/或单元。

[0032] 在任何情况下，控制单元12可表示一或多个中央处理单元(“CPU”，其在图1中未展

示)、图形处理单元(“GPU”,其在图1中同样未展示)或执行软件指令的其它处理单元,所述软件指令例如用于定义软件或计算机程序、存储到非暂时性计算机可读存储媒体(同样在图1中未展示)的软件指令,所述非暂时性计算机可读存储媒体例如存储装置(例如,磁盘驱动器或光盘驱动器)或存储器(例如快闪存储器、随机存取存储器或RAM)或存储致使一或多个处理单元执行本文中所描述的技术的指令的任何其它类型的易失性或非易失性存储器。

[0033] 或者或另外,控制单元12可表示专用硬件,例如一或多个集成电路、一或多个专用集成电路(ASIC)、一或多个特定应用专用处理器(ASSP)、一或多个现场可编程门阵列(FPGA)或用于执行本文中所描述技术的专用硬件的前述实例中的一或多者的任何组合。不论是否由执行软件、专用硬件或其某一组合的CPU及/或GPU组成,控制单元12在一些上下文中可被称作“处理器”。

[0034] 如上文所描述,摄像机14可表示能够俘获图像的任何类型的装置,而麦克风16可各自表示能够俘获音频数据的任何类型的装置。摄像机14可俘获光并产生一系列图像,其在下文的图1B的实例中展示为视频数据18。麦克风16中的每一者可俘获单独音频数据20A至20E。

[0035] 如图1的实例中进一步展示,控制单元12包含视觉分析单元22、听觉分析单元24、对象关联单元26、再现单元28A至28C(“再现单元28”)及音频混合单元30。视觉分析单元22可表示执行视频数据(例如视频数据18)的视觉场景分析的硬件或硬件与软件的组合。视觉场景分析可涉及计算机视觉的各方面,所述计算机视觉指代计算机或其它装置可处理及分析图像以检测及识别图像的各种对象、元件及/或方面的过程。计算机视觉在一些情况下可被称作机器视觉,因为计算机视觉及机器视觉具有许多重叠或相关概念。通常,机器视觉使用计算机视觉(但在不同上下文中)的方面或概念。尽管本发明当描述技术时涉及计算机视觉,但所述技术还可使用机器视觉结合计算机视觉或作为计算机视觉的一个替代方案来执行。出于这个原因,术语“机器视觉”及“计算机视觉”可互换使用。

[0036] 尽管在图1的实例中未展示,但视觉分析单元22在一些情况下可在执行视觉场景分析时与视频俘获装置10外部的图像服务器或其它数据库通信。视觉分析单元22可与所述图像服务器通信来分担通常资源(意指处理资源及/或存储器资源)密集的视觉场景分析过程的各种方面。举例来说,视觉分析单元22可执行某一初始分析来检测对象,从而将所述对象传递到图像服务器以供识别。图像服务器随后可分类或以其它方式识别对象,从而将所述分类对象传递回到视觉分析单元22。通常,视觉分析单元22通过无线会话与图像服务器通信。因而,视频俘获装置10可包含视频俘获装置10可借以无线地或通过有线连接与外围装置、服务器及任何其它类型的装置或附件通信的一或多个接口(但在图1的实例中未展示)。视觉分析单元22可输出视频对象32作为执行视觉场景分析的结果。

[0037] 听觉分析单元24可执行音频数据(例如音频数据20A至20N(“音频数据20”))的听觉场景分析并产生音频对象34。听觉分析单元24可分析音频数据来检测及识别音频对象。音频对象可指代独特的或可辨识的声音,其可经分类或以其它方式与给定对象相关联。举例来说,汽车引擎可发出可容易辨识的声音。听觉场景分析可尝试在音频数据中检测及识别或分类所述声音。

[0038] 类似于视觉分析单元22,听觉分析单元24在一些情况下可在执行听觉场景分析时与视频俘获装置10外部并且可能远离视频俘获装置10的音频网络服务器或其它数据库通

信(这在图1的实例中也未展示)。听觉分析单元24可与所述音频服务器通信来分担通常资源(意指处理资源及/或存储器资源)密集的听觉场景分析过程的各种方面。举例来说,听觉分析单元24可执行某一初始分析来检测对象,从而将所述对象传递到音频服务器以供识别。音频服务器随后可分类或以其它方式识别对象,从而将所述分类对象传递回到听觉分析单元24。听觉分析单元24可使用上文在描述视觉分析单元22时提及的接口与所述音频服务器通信。听觉分析单元24可输出音频对象34作为执行听觉场景分析的结果。

[0039] 对象关联单元26表示尝试使视频对象32与音频对象34相关联的硬件或硬件与软件的组合。视频对象32及音频对象34可各自根据可兼容的或常见的格式定义,意指视频对象32及音频对象34均以促进对象32与对象34之间的关联的方式定义。提供几个实例,对象32及34中的每一者可包含定义以下各者中的一或多者的元数据:对应对象的预测位置(例如,x、y、z坐标)、对应对象的大小(或预测大小)、对应对象的形状(或预测形状)、对应对象的速度(或预测速度)、位置可信程度及对象是否在焦点中或者对象属于近前景、远前景或近背景还是远背景。对象关联单元26可基于元数据使一或多个视频对象32与一或多个音频对象34相关联(通常使视频对象32中的单个一者与音频对象34中的单个一者相关联)。

[0040] 对象关联单元26可将对象32及34分类成三个类别中的一者。第一类别包含与具有元数据的视频对象32中的一者相关联的具有元数据的音频对象34中的所述音频对象。第二类别包含不与视频对象32中的任一者相关联的音频对象34中的所述音频对象。第三类别包含不与音频对象34中的任一者相关联的视频对象32中的所述视频对象。对象关联单元26可将分类为第一类别的音频对象34(展示为音频对象34')传递到辅助音频再现单元28A。对象关联单元26可将分类为第二类别的音频对象34(展示为音频对象34")传递到无辅助音频再现单元28B。对象关联单元26可将分类为第三类别的视频对象32(展示为视频对象32')传递到扩增实境音频再现单元28C。

[0041] 虽然关于三个类别进行了描述,但可仅关于前两个类别实施所述技术。换句话说,可基于可用资源适应性地执行第三类别。在一些情况下,不利用第三类别,尤其是在功率受限或资源受限装置中不利用。在一些情况下,由于不利用第三类别,这些功率受限或资源受限装置可不包含扩增实境音频再现单元28C。此外,对象关联单元26可不传递视频对象或以其它方式将视频对象分类成第三类别。因此,所述技术应不限于本发明中所描述但可关于第一和第二类别而不关于第三类别执行的实例。

[0042] 在任何情况下,再现单元28各自表示经配置以分别从音频对象34'、34"及视频对象32'中的一或多者再现音频数据38A至38C的硬件或硬件与软件的组合。辅助音频再现单元28A可被称为“辅助”音频再现单元28A,原因在于辅助音频再现单元28A接收通过视频对象32中的匹配或关联者潜在地扩增的具有元数据的音频对象34'。在这种意义上讲,再现单元28A可从视频对象32中的对应或关联者接收更准确地再现音频对象34'的辅助。鉴于单元28A接收与视频对象相关联的音频对象,这指示所述音频对象与由摄像机所俘获的视频对象相关联并且因此存在于前景中,辅助音频再现单元28A可被称为前景再现单元28A。

[0043] 无辅助音频再现单元28B可被称为“无辅助的”,原因在于再现单元28B再现分类为第二类别的音频对象34" (意指所述音频对象34"不与视频对象32中的任一者相关联)。因此,再现单元28B并不从视频对象32中的任一者接收再现音频对象34"的任何辅助。无辅助音频再现单元28B也可被称作背景再现单元28B,原因在于音频对象单元28B过程不与任何

视频对象相关联,意指所述对象很可能驻留在背景中或俘获场景作为视频数据18的用户后方。

[0044] 扩增实境音频再现单元28C可在以下意义上“扩增实境”:再现单元28C可存取音频库(位于装置10内部或外部)或其它音频存储库来检索对应于不匹配或不相关联的视频对象32'的音频对象并再现音频数据38C来扩增反映由麦克风16所俘获的音频数据20的音频数据38A和38B。鉴于单元28C处理在由摄像机14所俘获的作为视频数据18的场景中检测的视频对象32',扩增实境音频再现单元28C可在前景中再现音频数据。

[0045] 再现单元28中的每一者可以空间化方式再现音频数据38A至38C。换句话说,再现单元28可产生空间化音频数据38A至38C,其中在假定用于播放的某一扬声器配置的情况下分配及再现音频对象34'、34''及34'''中的每一者(其中音频对象34'''指代由扩增实境音频再现单元28C检索的扩增实境音频对象34'')。再现单元28可使用头部相关转移函数(HRTF)及当再现空间化音频数据时常用的其它再现算法来再现音频对象34'、34''及34'''。

[0046] 音频混合单元30表示将音频数据38A至38C(“音频数据38”)混合成特定多声道音频数据格式的硬件或硬件与软件的组合。本发明中对多声道音频数据的参考可指代立体声或更高阶的多声道音频数据。更高阶多声道音频数据可包含5.1环绕声音频数据或7.1环绕声音频数据,其中句点之前的第一数字指代声道的数目,并且句点之后的数字指代低音或低频声道的数目。举例来说,5.1环绕声音频数据包含左声道、中央声道、右声道、左后或环绕左声道和右后或环绕右声道以及单个低频声道。混合单元30可将音频数据38混合成所述多声道音频数据格式中的一或多者来产生多声道音频数据40。

[0047] 在操作中,视频俘获装置10可以经配置以调用摄像机14来俘获视频数据18,而同时也经配置以调用麦克风16中的一或多者(并且通常为全部麦克风)来俘获音频数据20A至20E(“音频数据20”)。响应于接收视频数据18及音频数据20,视频俘获装置10的控制单元12可以经配置以执行本发明中描述的技术以产生多声道音频数据40。

[0048] 在接收音频数据20后,控制单元12可即刻调用听觉分析单元24,其可分析音频数据20来识别音频对象34中的一或多者。如上文简要描述,听觉分析单元24可执行听觉场景分析来识别及产生音频对象34。同样,在接收视频数据18后,控制单元12可即刻经配置以调用视觉分析单元22,其可在分析及/或俘获音频数据20的同时分析视频数据18来识别一或多个视频对象32。而且,如上文简要描述,视觉分析单元22可执行视觉场景分析(使用计算机视觉算法)来识别及产生视频对象32中的一或多者。

[0049] 视觉分析单元22及音频分析单元24可以经配置以使用共同或共用格式分别产生视频对象32及音频对象34。通常,这种共用格式包含文本分量,其可被称作元数据。所述元数据可描述视频对象32及音频对象34中的对应者的各种性质或方面。描述视频对象32中的对应者的视频元数据可指定(作为几个非限制性实例)音频元数据中的一或多者包括对应视频对象的位置、形状、速度及位置可信程度中的一或多者。描述音频对象32中的对应者的音频元数据可同样指定(提供几个非限制性实例)音频对象位置、音频对象形状、音频对象速度及对应音频对象的位置可信程度中的一或多者。

[0050] 由于音频元数据及视频元数据均抽象化为同一语义层级(即,在此实例中的同一文本语义层级),所以视频俘获装置10可直接比较由所述元数据所指定的相应标记(其可指代上文所描述的不同类型的元数据中的每一者)并将所述标记映射(或,换句话说,使对象

相关联)在文本域中。在映射对象的情况下,视频俘获装置10可直接使装置如何“看到”对象与装置如何“听到”所述场景中的对象相关联。

[0051] 控制单元12可接收视频对象32及音频对象34并调用对象关联单元26。对象关联单元26可使音频对象34中的至少一者与视频对象32中的至少一者相关联。对象关联单元26可在执行所述关联时通常基于元数据(其在一些情况下可定义音频对象的类型)将音频对象34中的每一者分类为一种类型的音频对象。同样,对象关联单元26可在执行所述关联时通常基于对应元数据(其在一些情况下也可定义视频对象的类型)将视频对象32中的每一者分类为一种类型的视频对象。视频对象的实例类型可包括汽车、沙滩、波浪、流水、音乐、人、狗、猫等。对象关联单元26随后可确定音频对象34中的一者的类型为与视频对象32中的一者相同的类型。响应于确定音频对象34中的一者的类型与视频对象32中的一者的类型相同,对象关联单元26可使音频对象34中的一者与视频对象32中的一者相关联。

[0052] 对象关联单元26可基于将音频对象34分类成上文所描述的三个不同类别中的一者来产生多种音频对象。同样,第一类别包含与具有元数据的视频对象32中的一者相关联的具有元数据的音频对象34中的所述音频对象。第二类别包含不与视频对象32中的任一者相关联的音频对象34中的所述音频对象。第三类别包含不与音频对象34中的任一者相关联的视频对象32中的所述视频对象。

[0053] 对象关联单元26可将分类为第一类别的音频对象34(展示为音频对象34')传递到辅助音频再现单元28A。对象关联单元26可将分类为第二类别的音频对象34(展示为音频对象34'')传递到无辅助音频再现单元28B。对象关联单元26可将分类为第三类别的视频对象32(展示为视频对象32')传递到扩增实境音频再现单元28C。

[0054] 关于确定为属于第一类别的所述音频对象34,对象关联单元26可确定音频对象34中的一者的音频元数据与相关联的一个视频对象32的视频元数据之间的相关度,从而基于确定的相关度产生用于与所述一个视频对象32相关联的音频对象34中的一者的组合元数据。在某一实例中,对象关联单元26可用对应视频元数据或其部分替换音频元数据或其部分(如由音频元数据指定的位置)。对象关联单元26随后可将所述音频对象34作为音频对象34'中的一者传递到辅助音频再现单元28A。辅助音频再现单元28A随后可基于针对音频对象34'中的一者产生的组合元数据将音频对象34'中的一者再现在多声道音频数据40的一或多个前景声道中。辅助音频再现单元28A将多声道音频数据40的所述部分作为音频数据38A传递到音频混合单元30。

[0055] 关于确定为属于第二类别的所述音频对象34,对象再现单元26可确定音频对象34中的一者不与视频对象32中的任一者相关联。对象再现单元26可将所述音频对象34作为音频对象34''中的一者传递到无辅助音频再现单元28B。无辅助音频再现单元28B可产生多声道音频数据40,使得音频对象34''中的一者起源于多声道音频数据40的一或多个背景声道中。也就是说,因为所述音频对象34不与视频对象32中的任一者相关联,所以无辅助音频再现单元28B经配置以假定所述音频对象34''为出现在由摄像机14所俘获的场景外部的对象。因而,无辅助音频再现单元28B可以经配置以通常将音频对象34''作为扩散声再现在背景中。无辅助音频再现单元28B将多声道音频数据40的所述部分作为音频数据38B传递到音频混合单元30。

[0056] 关于确定为属于第三类别的所述视频对象32(即,其中视频对象32不与图1B的实

例中的音频对象34中的任一者相关联),对象关联单元26可将所述视频对象32作为视频对象32'传递到扩增实境音频再现单元28C。扩增实境音频再现单元28C可响应于接收视频对象32'而从音频库检索原本与视频对象32'中的各者(如果可能)相关联的参考音频对象。扩增实境音频再现单元28C随后可再现参考音频对象(其可被称作音频对象34'')中的每一者来产生多声道音频数据40的至少一部分。扩增实境音频再现单元28C将多声道音频数据40的所述部分作为音频数据38C传递到音频混合单元30。

[0057] 音频混合单元30接收音频数据38并将所述音频数据38混合以形成多声道音频数据40。音频混合单元30可以上文所描述的方式混合所述音频数据38来产生任何形式的多声道音频数据40。所述格式可包含5.1环绕声格式、7.1环绕声格式、10.1环绕声格式、22.2环绕声格式或任何其它专有或非专有格式。

[0058] 以这种方式,视频俘获装置10的控制单元12可以经配置以分析音频数据来识别一或多个音频对象并分析在俘获音频数据的同时用装置所俘获的视频数据来识别一或多个视频对象。控制单元12可进一步经配置以使音频对象34中的一者与视频对象32中的一者相关联并基于音频对象34中的一者与视频对象32中的一者的关联从音频数据20产生多声道音频数据40。

[0059] 鉴于视频场景分析可更确切地识别视频对象(相比仅音频对象)相对于视频俘获装置的位置,相比仅依赖于通常不准确的波束形成技术,视频俘获装置10可更好地本地化音频对象。随后可使用将音频对象更好地本地化到前声道中的一者的分贝差来将所述音频对象再现到一或多个声道,借此使得能够更好地产生环绕声或多声道音频数据(相比由常规视频俘获装置产生的音频数据)。

[0060] 此外,在一些实例中,视频俘获装置可将音频对象32作为离散音频源再现在前景中(收听者前方180度)。对于视频俘获装置10“听到”但未“看到”的音频对象32,视频俘获装置10可将所述音频对象32再现在背景中,因为所述音频对象32更可能在收听者后方。

[0061] 虽然上文描述为由视频俘获装置10执行,但所述技术可由与俘获视频数据18及音频数据20的装置不同的装置实施。换句话说,智能手机或其它视频俘获装置可俘获视频数据及音频数据,从而将所述视频数据及音频数据上载到一不同的装置,例如专用处理服务器、台式计算机、膝上型计算机、平板计算机或平板触摸计算机或能够处理数据的任何其它类型的装置。所述另一装置随后可执行本发明中描述的技术以促进产生可认为更准确的环绕声或多声道音频数据。因此,虽然描述为由俘获视频及音频数据的装置来执行,但所述技术可由与俘获视频及音频数据的装置不同的装置执行并且在这方面不应限于本发明中所描述的实例。

[0062] 图2A至2D为说明由图1的视频俘获装置10在根据本发明中所描述技术使视频对象32与音频对象34相关联时执行的操作的图。在上文图2A中,音频对象34中的一者(在图2A的实例中标示为“音频对象34A”)及视频对象32中的一者(在图2A的实例中标示为“视频对象32A”)包含相应音频元数据54A及视频元数据52A。视频俘获装置10的对象关联单元26可使音频对象34A与视频对象32A相关联,从而使用视频元数据52A来扩增音频元数据54A以产生具有扩增元数据56A的扩增音频对象34A'(其为图1B的实例中所展示的音频对象34中的一者)。所述扩增元数据56A可包含音频元数据54A及视频元数据52A,其中在一些情况下,视频元数据52A可替换音频元数据54A中的一些或全部。在一些情况下,对象关联单元26可确定

音频元数据54A及视频元数据52A具有高相关性。

[0063] 在其他情况下,对象关联单元26可确定音频元数据54A及视频元数据52A具有低相关性。在这种情况下,当产生扩增元数据56A时,对象关联单元26可加权视频元数据52A以相比音频元数据52A更偏向视频元数据52A。当再现及混合所述音频对象34A'来产生多声道音频数据40时,由于音频元数据54A与视频元数据52A之间缺乏相关性,辅助音频再现单元28A可在前景中将所述音频对象34A'再现为跨更多声道的更扩散的散布音频对象34A'。视频俘获装置10可对所述对象执行各种扩散算法(例如声音去相关)来扩散所述对象。

[0064] 在图2B的实例中,听觉分析单元24识别音频对象34中的另一者(在图2B的实例中标示为音频对象34B)但不能够识别音频对象34B的任何元数据。所述实例反映多个麦克风不可用于视频俘获装置10上并且因此视频俘获装置10无法确定音频元数据的情况。因此,当再现所述音频对象来产生音频对象34B'(其指代音频对象34'中的一者)时,对象关联单元26可利用相关联视频对象32B的视频元数据52B代替音频元数据。如图2B的实例中所展示,音频对象34B'包含视频元数据52B。

[0065] 在图2C的实例中,听觉分析单元24识别音频对象34中的一者(标示为“音频对象34C”)并确定所述音频对象的音频元数据54C,但不能够识别视频对象32中对应于所述音频对象34C的任一者。因为未识别用于所述音频对象34C的视频对象,所以对象关联单元26可确定音频对象34C位于视频俘获装置10后方。对象关联单元26可基于此确定将音频对象34C作为音频对象34'中的一者(即,在图2C的实例中的音频对象34C')传递到无辅助再现单元28B,其随后可在多声道音频数据40的背景声道中再现所述音频对象。当再现所述音频对象34C'时,无辅助音频再现单元28B可基于音频元数据54C中的预测位置或跨背景声道十分扩散地再现音频对象34C'。也就是说,视频俘获装置10可基于音频元数据估计实际位置,或十分扩散地再现对象(使用上文经识别声音扩散过程),因此其具有无特定感知角度空间云状形状。

[0066] 在图2D的实例中,对象关联单元26接收视频对象32中包含视频元数据52D的一者(在图2D的实例中标示为“视频对象32D”),但不能够使视频对象32D与音频对象34中的任一者相关联。因此,对象关联单元26将视频对象32D作为视频对象32'中的一者(即,在图2D的实例中的视频对象32D')传递到扩增实境音频再现单元28C。视频对象32D包含视频元数据52D。扩增实境音频再现单元28C可利用视频元数据52D来存取参考音频对象34''的库,从而检索原本与视频对象32D'相关联的参考音频对象34''中的一者(例如,匹配识别对象32D'的类型的视频元数据52D中所指定的类型的参考音频对象34''')。扩增实境音频再现单元28C随后可使用视频元数据52D再现所述参考音频对象34'''来优化或以其它方式空间化音频对象34'''的再现。

[0067] 以此方式,视频俘获装置10可基于由音频对象所指定的元数据与由相关联视频对象所指定的元数据的相关性再现音频对象,从而尝试本地化起源于某一视频对象或其部分的音频对象。考虑到视频场景分析通常比听觉场景分析准确得多,在一些实例(例如图2A)中,视频俘获装置10相比音频对象元数据可更偏向(使用权数)视频对象元数据。在一些情况下,视频俘获装置可产生不具有元数据或具有十分不确定的元数据的音频对象(例如图2B的实例中所展示),其中视频俘获装置可导入“匹配”视频对象元数据用作在再现音频对象时使用的元数据。

[0068] 为了说明, 扩增元数据56A可包含音频元数据54A及视频元数据52A, 其中在一些情况下, 视频元数据52A可替换音频元数据54A。在一些情况下, 视频俘获装置10可确定音频元数据54A及视频元数据52A具有高相关性。换句话说, 视频俘获装置10可确定音频元数据54A中所指定的发出声音的对象的位置与由视频元数据52A定义的对应对象的位置高度(例如, 其可由某一可信度阈值定义, 通常表达为百分比)相关。视频俘获装置10随后可再现及混合音频对象来产生具有高可信度的多声道音频数据40。

[0069] 在其它情况下, 视频俘获装置10可确定音频元数据54A及视频元数据52A具有低相关性。在这种情况下, 当产生扩增元数据56A时, 视频俘获装置10可加权视频元数据52A以相比音频元数据54A更偏向视频元数据52A。当再现及混合音频对象34A'来产生多声道音频数据40时, 由于音频元数据54A与元数据52A之间缺乏相关性, 视频俘获装置10可在前景中将音频对象34A'再现为跨更多声道的更扩散的散布音频对象34A'。

[0070] 图3为更详细说明图1B的辅助音频再现单元28A的方块图。在图3的实例中, 辅助音频再现单元28A包含大量空间音频再现单元60A至60N(“空间音频再现单元60”)。虽然图3的实例中展示大量空间音频再现单元60, 在一些情况下, 但辅助音频再现单元28可仅包含能够并行处理多个对象的单个空间音频再现单元60。或者, 辅助音频再现单元28可包含能够处理仅单个音频对象的单个空间音频再现单元60。因此, 所述技术在这方面不应限于图3的实例。

[0071] 在图3的实例中, 空间音频再现单元60中的每一者可表示执行关于音频对象34A'至34N'(“音频对象34”, 其展示于图1B的实例中)的空间音频再现来产生音频对象38A的单独音频再现过程空间音频再现可指代用于再现音频数据的各种算法或过程并可包含(作为几个实例)立体混响、波场合成(WFS)及基于向量的振幅平移(VBAP)。空间音频再现单元60可基于扩增元数据56A至56N(“扩增元数据56”)处理音频对象34'中的相应者。也就是说, 空间音频再现单元60可使用扩增元数据56再现音频对象34'来进一步优化或以其它方式更准确定位音频对象34'中的对应者, 使得可在播放多声道音频数据40时更准确再现音频对象34'中的所述一者。空间音频再现单元60可将再现的音频数据38A输出到音频混合单元30, 其随后可混合再现的音频数据38A以产生多声道音频数据40。在一些情况下, 可跨多声道音频数据40的两个或两个以上声道混合对应于给定音频对象34'的音频数据38A。

[0072] 虽然在图3的实例中关于辅助音频再现单元28A进行描述, 但再现单元28中的每一者可包含类似于空间音频再现单元60的空间音频再现单元, 其同样可处理音频对象34'及34''(其同样指代从参考音频库检索的参考音频对象34''并且其原本与视频对象32'相关联)来产生音频数据38B及38C。此外, 虽然描述为包含再现单元28C, 但视频俘获装置10可不包含再现单元28C, 其中视频俘获装置10可不执行本发明中所描述技术的扩增实境音频再现方面。

[0073] 图4为说明由图1B的实例中所展示的视频俘获装置10的摄像机14所俘获并且根据本发明中所描述技术经处理的场景70的图。场景70可表示图1B的实例中所展示的视频数据18的一部分。视频俘获装置10可响应于接收场景70而调用视觉分析单元22, 其处理场景70来识别视频对象32。

[0074] 如图4中所展示, 场景70(例如)按帧的时间顺序包含第一帧或图像72A、第二帧或图像72B及第三帧或图像72C。虽然为了易于说明目的展示为包含仅三个帧或图像72A至72C

(“图像72”),但场景70可包含大量图像72或单个图像72,并且所述技术在这方面不应限于图4中所展示的实例。

[0075] 在任何情况下,视觉分析单元22可使用计算机视觉算法处理图像72A来识别视频对象32A至32G。视觉分析单元22可产生视频对象32A至32G以包含视频元数据52A至52G或以其它方式与其相关联。视频元数据52A至52G可定义视频对象32A至32G相对于俘获场景70的摄像机14的对应位置。视频元数据52A至52G一般也可(例如)基于基于机器视觉的对象辨识识别视频对象32中的对应者的类型,所述基于机器视觉的对象辨识可完全支持于视觉分析单元22内或通过视觉分析单元22结合一或多个外部及可能远程的网络服务器支持。举例来说,与视频对象32A相关联的视频元数据52A可将视频对象32A识别为汽车。作为另一实例,视频元数据52B至32F可将视频对象32B至32F的对应者的类型识别为人。作为又一实例,视频元数据52G可将对应视频对象32G的类型识别为立体声系统。

[0076] 视觉分析单元22可同时分析图像72中的一或多者来产生呈视觉元数据52A至52G形式的位置信息,以表达描述视频对象32A至32G在场景70期间如何移动的动作、速度或其它位置相关度量标准。为了说明,考虑从图像72A到图像72C的视频对象32A,其中视频对象32A沿几乎水平的线从第一位置移动到第二位置并且随后到第三位置。视觉分析单元22可识别对象32A,从而产生视频元数据52A来指示,从图像72A到图像72B并且随后到图像72C,视频对象32A从第一位置移动到第二位置并且随后到第三位置。所述视频元数据52A在与音频对象34中的对应一者(例如,音频对象34A)相关联时可使得对象关联单元26能够扩增音频元数据54A来更准确限定识别为音频对象34A的发出音频数据的对象的位置(在视觉场景分析通常比听觉场景分析更准确的条件下)。对象关联单元26随后可产生具有扩增元数据56A(如例如在图2A中所展示)的音频对象34’。

[0077] 作为另一实例,考虑视频对象32G在场景70内移动。最初,图像72A展示在第一位置的视频对象32G。图像72B展示在第二位置的视频对象32G。图像72C并不包含视频对象32G,表明视频对象32G已离开所述场景并且在背景中或到由摄像机14所俘获的场景70的左侧或右侧。对象关联单元26随后可产生包含指定视频对象32G在贯穿场景70移动时的位置的视觉元数据52G的视频对象32G。对象关联单元26可使视频对象32G与音频对象34中的具有指定同一类型(即在所述实例中的立体声系统)的元数据的一者相关联。但是,鉴于视频对象32G离开场景,对象关联单元26可不替换或以其它方式利用由视频元数据52G所指定的位置信息,而是实际上维持由与音频对象34中的所述者相关联的音频元数据54所指定的位置信息。

[0078] 当再现音频对象34中的关联者(例如,音频对象34G)以供关于图像72A、72B的播放时,对象关联单元26可利用由视频元数据52G所指定的位置。但是,视频元数据52G可在所述时间指定位置的高可信程度,但在对应于图像72C的时间指定位置信息的低至零可信度。因此,当再现相关联音频对象34G以供在再现图像72C期间的时间播放时,对象关联单元26可不替换或以其它方式利用由视频元数据52G所指定的位置信息。实际上,当在再现图像72C的时间期间再现音频对象34G时,对象关联单元26可利用由音频对象34G所指定的位置信息。

[0079] 如上所述,对象关联单元26可能不能够识别对应于音频对象34G的视频对象32G,例如在图像72C的实例中。也就是说,视频对象32G可离开场景70,如图像72C中所展示,但从

立体声系统播放的音乐可仍被俘获并识别为音频对象34G。在所述实例中,对象关联单元26可执行上文关于图2C所描述的操作。也就是说,对象关联单元26可将音频对象34G从与视频对象相关联的音频对象的当前分类重新分类成不与视频对象32中的任一者相关联的音频对象34G,从而以以上文关于图2C所描述的方式处理音频对象34G。对象关联单元26可产生音频对象34G”,从而将所述音频对象34G”传递到无辅助音频再现单元28B。在这方面,音频对象34G可从以上文关于图2A所描述的方式处理转变成以上文关于图2C所描述的方式处理。

[0080] 以这种方式,视频俘获装置10可动态地执行本发明中描述的技术来潜在地产生更准确的多声道音频数据40。为此,视频俘获装置10可适应性地分类音频对象34,从而使所述音频对象34及视频对象32在上文所描述的三个类别中的各者之间转变。在一些情况下,视频俘获装置10可适应性地分类音频对象34及视频对象32,从而使处理音频对象34及视频对象32从上文关于图2A至2D所描述的方式中的一者转变成上文关于图2A至2D所描述的方式中的不同者。

[0081] 图5为说明由图1B的实例中所展示的视频俘获装置10的摄像机14所俘获并且根据本发明中所描述技术的扩增实境方面经处理的另一场景80的图。在图5的实例中,场景80可表示图1B的实例中所展示的视频数据18的一部分。视频俘获装置10可响应于接收场景80而调用视觉分析单元22,其处理场景80来识别视频对象32I及32H。场景80包含图像82。虽然展示为包含仅单个图像(图像82),但为了易于说明目的,场景80可包含额外图像,并且所述技术在这方面不应限于图5中所展示的实例。

[0082] 在任何情况下,视觉分析单元22可识别并产生分别包含视频元数据52I及52H的视频对象32I及32H。视觉分析单元22可将视觉对象32I及32H传递到对象关联单元26,其可尝试使视觉对象32I及32H与音频对象34中的一者相关联。出于实例目的假定对象关联单元26使视觉对象32I与音频对象34中的一者(例如,音频对象34I)相关联。对象关联单元26随后可以类似于上文关于图2A的实例所描述的方式考虑到相关联视频对象32I来处理音频对象34I。对象关联单元26随后可产生具有扩增元数据56I的音频对象34I’。

[0083] 除识别为视频对象32I的人之外,场景80还包含视觉分析单元22识别为视频对象32H的沙滩,其中出于说明的目的假定波浪的声音不由麦克风16俘获。也就是说,假定视频俘获装置10距沙滩足够遥远,使得撞到沙石上的波浪的声音由于所述距离、人的谈话、风声或某一其它干扰根本不可听。因此,对象关联单元26可将视频对象32H分类为属于第三类别,即,在本发明的实例中,视频对象32中不与音频对象34中的任一者相关联的所述视频对象。因此,对象关联单元26可以上文关于图2D的实例所描述的方式处理视频对象32H,从而产生视频对象32H’。对象关联单元26随后可将视频对象32H’转发到扩增实境音频再现单元28C。

[0084] 音频再现单元28C可接收视频对象32H’并检索参考音频对象34”’中为同一类型的对应者,所述类型在所述实例中可为波浪、沙滩等的类型。音频再现单元28C随后可基于视频元数据52H再现参考音频对象34”’中的所述者(例如,音频再现对象34H”’)。扩增实境音频再现单元28C可将所述再现的音频数据作为音频数据38C传递到混合单元30,其混合音频数据38A至38C来以上文所描述的方式形成多声道音频数据40。

[0085] 图6为说明视频俘获装置(例如图1B的实例中所展示的视频俘获装置10)在执行本

发明中描述的技术时的示范性操作的流程图。最初,视频俘获装置10可以经配置以调用摄像机14来俘获视频数据18,而同时也经配置以调用麦克风16中的一或多个(并且通常为全部麦克风)来俘获音频数据20(90、92)。响应于接收视频数据18及音频数据20,视频俘获装置10的控制单元12可以经配置以执行本发明中描述的技术以产生多声道音频数据40。

[0086] 在接收视频数据18后,控制单元12可以经配置以调用视觉分析单元22,其可执行关于视频数据18的视觉场景分析来识别一或多个视频对象32(94)。在接收音频数据20后,控制单元12可调用听觉分析单元24,其可执行关于音频数据20的听觉场景分析来识别音频对象34中的一或多者(96)。

[0087] 控制单元12可接收视频对象32及音频对象34并调用对象关联单元26。对象关联单元26可比较音频对象34与视频对象32以尝试使音频对象34中的至少一者与视频对象32中的至少一者相关联(98)。如上文所描述,对象关联单元26可在执行所述关联时通常基于元数据(其在一些情况下可定义音频对象的类型)将音频对象34中的每一者分类为一种类型的音频对象。同样,对象关联单元26可在执行所述关联时通常基于对应元数据(其在一些情况下也可定义视频对象的类型)将视频对象32中的每一者分类为一种类型的视频对象。实例类型可包括汽车、沙滩、波浪、流水、音乐、人、狗、猫、风等。对象关联单元26随后可确定音频对象34中的一者的类型为与视频对象32中的一者相同的类型并且借此确定匹配(100)。响应于确定音频对象34中的一者的类型与视频对象32中的一者的类型相同或(换句话说)已识别匹配(“是”,100),对象关联单元26可使音频对象34中的一者与视频对象32中的匹配者相关联(102)。

[0088] 关于确定为属于第一类别的所述音频对象34,对象关联单元26可确定音频对象34中的一者的音频元数据与视频对象32中的相关联的一者的视频元数据之间的相关度,从而基于确定的相关度产生用于与所述一个视频对象32相关联的音频对象34中的一者的组合元数据。在某一实例中,对象关联单元26可用对应视频元数据或其部分替换音频元数据或其部分(如由音频元数据指定的位置)。以这种方式,对象关联单元26可基于视频对象32中的关联者更新音频对象34中的一或多者来产生更新或扩增的音频对象34’(104)。

[0089] 对象关联单元26随后可将所述音频对象34’传递到辅助音频再现单元28A。辅助音频再现单元28A随后可基于针对音频对象34’中的一者产生的组合元数据将音频对象34’中的一者再现在多声道音频数据40的一或多个前景声道中(106)。辅助音频再现单元28A将多声道音频数据40的所述部分作为音频数据38A传递到音频混合单元30。

[0090] 关于确定为属于第二类别的所述音频对象34,即在本发明的实例中未确定为对应于视频对象32中的任一者的所述音频对象34(或,换句话说,不存在匹配的音频对象,“否”100,“是”108),对象关联单元26可将所述音频对象34作为音频对象34”中的一者传递到无辅助音频再现单元28B。无辅助音频再现单元28B可产生多声道音频数据40,使得音频对象34”中的一者起源于多声道音频数据40的一或多个背景声道中。无辅助音频再现单元28B可以经配置以通常将不匹配的音频对象34”作为扩散声再现在背景中(110)。无辅助音频再现单元28B将多声道音频数据40的所述部分作为音频数据38B传递到音频混合单元30。

[0091] 关于确定为属于第三类别的所述视频对象32,即在图1B的实例中不与音频对象34中的任一者相关联的视频对象32(或,换句话说,不匹配音频对象34中的任一者并且为视频对象的所述视频对象32,“否”100,“否”108),对象关联单元26可将所述视频对象32作为视

频对象32'传递到扩增实境音频再现单元28C。扩增实境音频再现单元28C可响应于接收视频对象32'而从音频库检索原本与视频对象32'中的每一者(如果可能)相关联的参考音频对象,并且随后再现参考音频对象(其可被称为音频对象34'')中的每一者来产生多声道音频数据40的至少一部分(112)。扩增实境音频再现单元28C将多声道音频数据40的所述部分作为音频数据38C传递到音频混合单元30。

[0092] 音频混合单元30接收音频数据38并将所述音频数据38混合以形成多声道音频数据40(114)。音频混合单元30可以上文所描述的方式混合所述音频数据38来产生任何形式的多声道音频数据40。所述格式可包含5.1环绕声格式、7.1环绕声格式、10.1环绕声格式、22.2环绕声格式或任何其它专有或非专有格式。音频混合单元30随后可输出所述多声道音频数据40(116)。

[0093] 以这种方式,视频俘获装置10的控制单元12可以经配置以分析音频数据来识别一或多个音频对象并分析在俘获音频数据的同时用装置所俘获的视频数据来识别一或多个视频对象。控制单元12可进一步经配置以使音频对象34中的一者与视频对象32中的一者相关联并基于音频对象34中的一者与视频对象32中的一者的关联从音频数据20产生多声道音频数据40。

[0094] 虽然在产生多声道音频数据40的上下文中进行描述,但视频俘获装置10可进一步编码视频数据。当编码视频数据时,扩散音频对象可使得视频俘获装置10能够使用更少位来编码所述音频对象。也就是说,可不需要以高质量再现后部背景或远距离中的音频对象,因为其不被眼睛所见或与聚焦的近距离空间中的音频对象相比较不重要,并且其十分可能在与其它音频对象一起再现时被遮挡。因此,视频俘获装置10可在编码及传输所述音频对象以用于播放系统时将较少位分配给所述音频对象。

[0095] 而且,虽然描述为在俘获音频数据及视频数据(或由于通常参考这种形式的处理而“离线”)之后执行或不在实时或近实时系统中执行,但所述技术可在俘获音频数据及视频数据的至少一部分期间在实时或近实时系统中加以实施。存在近实时或实时系统对视频场景分析的实施例,而音频场景分析通常不如视频场景分析复杂,意指音频场景分析可在近实时或实时装置中执行。

[0096] 此外,虽然关于音频及视觉域进行描述,但可关于其它域执行所述技术。举例来说,触摸、运动、指南针、海拔高度、温度及其它传感器域也可一起被视为通过对3D空间本质的可能聚焦增强媒体再现质量。因此,所述技术在这方面不应限于本发明中阐述的实例。

[0097] 图7为说明可如何根据本发明中所描述技术在多声道音频数据的前景和背景中再现各种音频对象126A至126K的图。图7的图详细说明展示根据自顶向下视角或鸟眼视野通常被称作“甜点(sweet spot)”处的视图120。所述甜点指代当扬声器恰当地配置成用于5.1或更高阶环绕声播放时房间中环绕声体验最为理想的位置。

[0098] 在图7的实例中,将视图120分段成两个部分,将其标示为前景部分122A及背景部分122B。在圆圈内,收听者124位于甜点的中间在分隔前景部分122A与背景部分122B的水平通路上。在播放多声道音频数据40期间,收听者124可在视图120中所再现的声场中听到音频对象126A至126K。也就是说,音频对象126A至126D从收听者124的视角显得起源于更遥远的前景。可能已以以上文关于图2B所描述的方式由对象关联单元26处理音频对象126A至126D,结果由于缺乏任何音频元数据,辅助音频再现单元28A将所述音频对象作为更扩散的

音频对象再现于远前景中。

[0099] 音频对象126E至126G从收听者124的视角可显得如同更聚焦的对象一样起源于更近前景中。可能已以上文关于图2A所描述的方式由对象关联单元26处理音频对象126E至126G,结果由于对象关联单元26提供具有高音频与视频元数据相关性的扩增元数据的能力,辅助音频再现单元28A将所述音频对象再现于更聚焦的前景中。

[0100] 音频对象126A至126G中的一或多个可为以上文关于扩增实境音频再现单元28C所描述的方式从参考库检索的参考音频对象。在这种意义上讲,对象关联单元26可识别不匹配音频对象34中的任一者的所述视频对象32,从而将视频对象32中的所述各者作为视频对象32'传递到扩增实境音频再现单元28C。扩增实境音频再现单元28C随后可检索对应或匹配视频对象32'中的一者的参考音频对象34''中的一者并基于包含在视频对象32'中的关联者内的视频元数据再现参考音频对象34''中的所述者。

[0101] 音频对象126H至126K从收听者124的视角可显得起源于背景中。可能已以上文关于图2C所描述的方式由对象关联单元26处理音频对象126H至126K,结果由于对象关联单元26不能使所述音频对象34''与视频对象32中的任一者相关联,无辅助音频再现单元28B将所述音频对象再现于背景中。也就是说,因为听觉场景分析在定位声音的起源方面通常不如视觉场景分析精确,所以无辅助音频再现单元28B可能不能够精确定位音频对象34''的来源。无辅助音频再现单元28B可最多基于对应音频元数据54而仅再现音频对象34'',这可导致音频再现单元28B将所述音频对象34''作为更扩散对象再现在背景中。

[0102] 以这种方式,所述技术可使得装置能够分析用装置俘获的音频数据以识别一或多个音频对象及分析用装置在俘获音频数据的同时所俘获的视频数据来识别一或多个视频对象。所述装置可进一步使一或多个音频对象中的至少一者与一或多个视频对象中的至少一者相关联并基于一或多个音频对象中的至少一者与一或多个视频对象中的至少一者的关联从音频数据产生多声道音频数据。

[0103] 在一些情况下,所述装置可在分析音频数据时执行音频数据的听觉场景分析来识别一或多个音频对象及描述所述一或多个音频对象的音频元数据,其中音频元数据包括对应音频对象的位置、形状、速度及位置可信程度中的一或多个。所述装置可在分析视频数据时执行视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据,其中视频元数据包括对应音频对象的位置、形状、速度及位置可信程度中的一或多个。

[0104] 在一些情况下,所述装置可在使一或多个音频对象中的至少一者与一或多个视频对象中的至少一者相关联时将一或多个音频对象中的每一者分类为一类型的音频对象、将一或多个视频对象中的每一者分类为一类型的视频对象、确定音频对象中的至少一者的类型为与视频对象中的至少一者相同的类型并响应于确定一或多个音频对象中的至少一者的类型与一或多个视频对象中的至少一者的类型相同而使一或多个音频对象中的至少一者与一或多个视频对象中的至少一者相关联。

[0105] 在一些情况下,所述装置可在产生多声道音频数据时确定一或多个音频对象中的至少一者的音频元数据与关联于一或多个音频对象中的至少一者的一或多个视频对象中的至少一者的视频元数据之间的相关度、基于所确定的相关度产生用于与一或多个视频对象中的至少一者相关联的一或多个音频对象中的至少一者的组合元数据并基于针对一或

多个音频对象中的至少一者产生的组合元数据将一或多个音频对象中的至少一者再现在多声道音频数据的一或多个前景声道中。

[0106] 在一些情况下,一或多个音频对象中的至少一者包括一或多个音频对象中的第一者。在这些情况下,所述装置可进一步确定一或多个音频对象中的第二者不与一或多个视频对象中的任一者相关联,并且在产生多声道音频数据时,产生多声道音频数据以使得音频对象中的第二者起源于多声道音频数据的一或多个背景声道中。

[0107] 所述装置可在产生多声道音频数据时产生多声道音频数据以使得音频对象中的第二者作为扩散音频对象起源于多声道音频数据的一或多个背景声道中。

[0108] 在一些情况下,一或多个视频对象中的至少一者包括一或多个视频对象中的第一者。在这些情况下,所述装置可确定视频对象中的一或多者中的第二者与一或多个音频对象中的任一者相关联。响应于确定视频对象中的一或多者中的第二者与一或多个音频对象中的任一者相关联,所述装置可从音频库检索原本与一或多个视频对象中的第二者相关联的参考音频对象。此外,所述装置可基于一或多个视频对象中的第二者再现参考音频对象来产生多声道音频数据的至少一部分。

[0109] 在一些情况下,所述装置可在分析音频数据时执行音频数据的听觉场景分析来识别一或多个音频对象及描述所述一或多个音频对象的音频元数据。所述装置也可在分析视频数据时执行视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据。在这些情况下,以与用于定义视频元数据的文本格式共同的文本格式定义音频元数据。

[0110] 在一些情况下,所述装置可在分析音频数据时执行音频数据的听觉场景分析来识别一或多个音频对象及描述所述一或多个音频对象的音频元数据。当分析视频数据时,所述装置可执行视频数据的视觉场景分析来识别一或多个视频对象及描述所述一或多个视频对象的视频元数据。在这些情况下,所述装置可在产生多声道音频数据时确定针对音频对象中的至少一者识别的音频元数据与针对视频对象中的关联者识别的视频元数据之间的相关度,并基于在产生多声道音频数据时所确定的相关度将音频对象中的至少一者再现为扩散音频对象。通常所述相关度基于某一形式的可信度区间,其中相关度可作为音频及对应物视频对象元数据与可信区间之间的百分比差异的函数导出。

[0111] 所述技术的各种方面还可使得装置能够包括一或多个处理器,以获得音频对象、获得视频对象、使音频对象与视频对象相关联、比较所述音频对象与相关联的视频对象并基于音频对象与相关联视频对象之间的比较再现音频对象。

[0112] 在一些情况下,音频对象包含音频元数据。在一些情况下,音频元数据包括大小及位置。在一些情况下,视频对象包含视频元数据。在一些情况下,视频元数据包括大小及位置。

[0113] 在一些情况下,一或多个处理器在比较音频对象与相关联视频对象时经进一步配置以至少部分产生包括大小及位置中的一或多者的组合元数据。

[0114] 在一些情况下,音频对象包含位置元数据,并且视频对象包含位置元数据。当产生组合元数据时,一或多个处理器经进一步配置以比较音频对象的位置元数据与视频对象的位置元数据来确定相关值,并且基于对相关值是否超过可信度阈值的确定产生组合元数据的位置元数据。

[0115] 此外,所述技术的各种方面可提供一种方法,所述方法包含获得音频对象、获得视频对象、使音频对象与视频对象相关联、比较所述音频对象与相关联的视频对象并基于音频对象与相关联视频对象之间的比较再现音频对象。

[0116] 另外,当比较音频对象与相关联视频对象时,方法可进一步包括至少部分产生包括大小及位置中的一或多者的组合元数据。

[0117] 而且,当音频对象包含位置元数据并且视频对象包含位置元数据时,产生组合元数据可包括比较音频对象的位置元数据与视频对象的位置元数据来确定相关值及基于对相关值是否超过可信度阈值的确定产生组合元数据的位置元数据。

[0118] 此外,所述技术的各种方面可提供一种装置,所述装置包括:用于获得音频对象的装置,用于获得视频对象的装置,使音频对象与视频对象相关联的装置,用于比较音频对象与相关联视频对象的装置,及用于基于音频对象与相关联视频对象之间的比较再现音频对象的装置。

[0119] 另外,用于比较音频对象与相关联视频对象的装置可包括用于至少部分产生包括大小及位置中的一或多者的组合元数据的装置。

[0120] 而且,当音频对象包含位置元数据并且视频对象包含位置元数据时,用于产生组合元数据的装置可包括用于比较音频对象的位置元数据与视频对象的位置元数据来确定相关值的装置及用于基于对相关值是否超过可信度阈值的确定产生组合元数据的位置元数据的装置。

[0121] 在一些情况下,一种非暂时性计算机可读存储媒体上存储有在执行时致使一或多个处理器执行以下操作的指令:获得音频对象、获得视频对象、使音频对象与视频对象相关联、比较音频对象与相关联视频对象及基于音频对象与相关联视频对象之间的比较再现音频对象。

[0122] 本发明中描述的技术的各种方面还可由产生音频输出信号的装置执行。所述装置可包括:用于基于第一音频对象的数据分量与第一视频对象的数据分量的第一比较识别与第一视频对象对应物相关联的第一音频对象的装置,及用于基于第二音频对象的数据分量与第二视频对象的数据分量的第二比较识别不与第二视频对象对应物相关联的第二音频对象的装置。所述装置另外可包括:用于在第一区域中再现第一音频对象的装置,用于在第二区域中再现第二音频对象的装置,及用于基于组合第一区域中再现的第一音频对象和第二区域中再现的第二音频对象产生音频输出信号的装置。本文中所描述的各种装置可包括经配置以执行关于装置中的每一者描述的功能的一或多个处理器。

[0123] 在一些情况下,第一音频对象的数据分量包括位置及大小中的一者。在一些情况下,第一视频对象数据的数据分量包括位置及大小中的一者。在一些情况下,第二音频对象的数据分量包括位置及大小中的一者。在一些情况下,第二视频对象的数据分量包括位置及大小中的一者。

[0124] 在一些情况下,第一区域及第二区域为音频前景内的不同区域或音频背景内的不同区域。在一些情况下,第一区域及第二区域为音频前景内的同一区域或音频背景内的同一区域。在一些情况下,第一区域在音频前景内,并且第二区域在音频背景内。在一些情况下,第一区域在音频背景内,并且第二区域在音频前景内。

[0125] 在一些情况下,第一音频对象的数据分量、第二音频对象的数据分量、第一视频对

象的数据分量及第二视频对象的数据分量各自包括元数据。

[0126] 在一些情况下,所述装置进一步包括:用于确定第一比较是否在可信度区间外的装置,及用于基于第一比较是否在可信度区间外的确定加权第一音频对象的数据分量及第一视频对象的数据分量的装置。在一些情况下,用于加权的装置包括用于对第一音频对象数据的数据分量及第一视频对象的数据分量求平均的装置。

[0127] 在一些情况下,所述装置也可为用于基于第一比较及第二比较中的一或多者分配不同数目的位的装置。

[0128] 在一些情况下,所述技术可提供一种非暂时性计算机可读存储媒体,其上存储有在执行时致使一或多个处理器执行以下操作的指令:基于第一音频对象的数据分量与第一视频对象的数据分量的第一比较识别与第一视频对象对应物相关联的第一音频对象,基于第二音频对象的数据分量与第二视频对象的数据分量的第二比较识别不与第二视频对象对应物相关联的第二音频对象,在第一区域中再现第一音频对象,在第二区域中再现第二音频对象,及基于组合第一区域中再现的第一音频对象与第二区域中再现的第二音频对象产生音频输出信号。

[0129] 应理解,取决于实例,本文中所描述方法中的任一者的某些动作或事件可依不同顺序执行、可以添加、合并或全部省略(例如,实践所述方法并不需要所有的所描述动作或事件)。此外,在某些实例中,可同时(例如,通过多线程处理、中断处理或多个处理器)而非顺序地执行动作或事件。另外,虽然为了清晰起见,将本发明的某些方面描述为是通过单个模块或单元执行,但应理解,本发明的技术可以通过与视频译码器相关联的单元或模块的组合执行。

[0130] 在一或多个实例中,所描述功能可在硬件、软件、固件或其任何组合中实施。如果在软件中实施,那么所述功能可以作为一或多个指令或代码存储在计算机可读媒体上或经由计算机可读媒体传输,并且由基于硬件的处理单元执行。计算机可读媒体可包含计算机可读存储媒体,其对应于有形媒体(例如数据存储媒体),或包含促进将计算机程序从一处传送到另一处(例如,根据通信协议)的任何媒体的通信媒体。

[0131] 以这种方式,计算机可读媒体一般可对应于(1)有形计算机可读存储媒体,其为非暂时性的,或(2)通信媒体,例如信号或载波。数据存储媒体可为可由一或多个计算机或一或多个处理器存取以检索用于实施本发明中描述的技术的指令、代码和/或数据结构的任何可用媒体。计算机程序产品可包含计算机可读媒体。

[0132] 借助于实例而非限制,此类计算机可读存储媒体可包括RAM、ROM、EEPROM、CD-ROM或其它光盘存储装置、磁盘存储装置或其它磁性存储装置、快闪存储器或可用来存储呈指令或数据结构的形式的所要程序代码并且可由计算机存取的任何其它媒体。而且,任何连接被适当地称为计算机可读媒体。举例来说,如果使用同轴电缆、光纤缆线、双绞线、数字订户线(DSL)或例如红外线、无线电和微波等无线技术从网站、服务器或其它远程源传输指令,那么同轴电缆、光纤缆线、双绞线、DSL或例如红外线、无线电和微波等无线技术包含在媒体的定义中。

[0133] 但是,应理解,所述计算机可读存储媒体和数据存储媒体并不包含连接、载波、信号或其它暂时媒体,而是实际上针对于非暂时性有形存储媒体。如本文中所使用,磁盘和光盘包含压缩光盘(CD)、激光光盘、光学光盘、数字多功能光盘(DVD)、软性磁盘和蓝光光盘,

其中磁盘通常以磁性方式再现数据,而光盘用激光以光学方式再现数据。以上各者的组合也应包含在计算机可读媒体的范围内。

[0134] 可由例如一或多个数字信号处理器(DSP)、通用微处理器、专用集成电路(ASIC)、现场可编程逻辑阵列(FPGA)或其它等效集成或离散逻辑电路的一或多个处理器来执行指令。因此,如本文中所使用的术语“处理器”可指代上述结构或适合于实施本文中所描述的技术的任何其它结构中的任一者。另外,在一些方面中,本文所述的功能性可在经配置用于编码和解码的专用硬件及/或软件模块内提供,或者并入在组合编解码器中。而且,可将所述技术完全实施于一或多个电路或逻辑元件中。

[0135] 本发明的技术可在广泛多种装置或设备中实施,包含无线手持机、集成电路(IC)或一组IC(例如,芯片组)。本发明中描述各种组件、模块或单元来强调经配置以执行所公开的技术的装置的功能方面,但未必需要通过不同硬件单元实现。确切地说,如上文所描述,各种单元可结合合适的软件和/或固件组合在编解码器硬件单元中,或者由互操作硬件单元的集合来提供,所述硬件单元包含如上文所描述的一或多个处理器。

[0136] 已描述所述技术的各种实施例。所述和其它实施例在所附权利要求书的范围内。

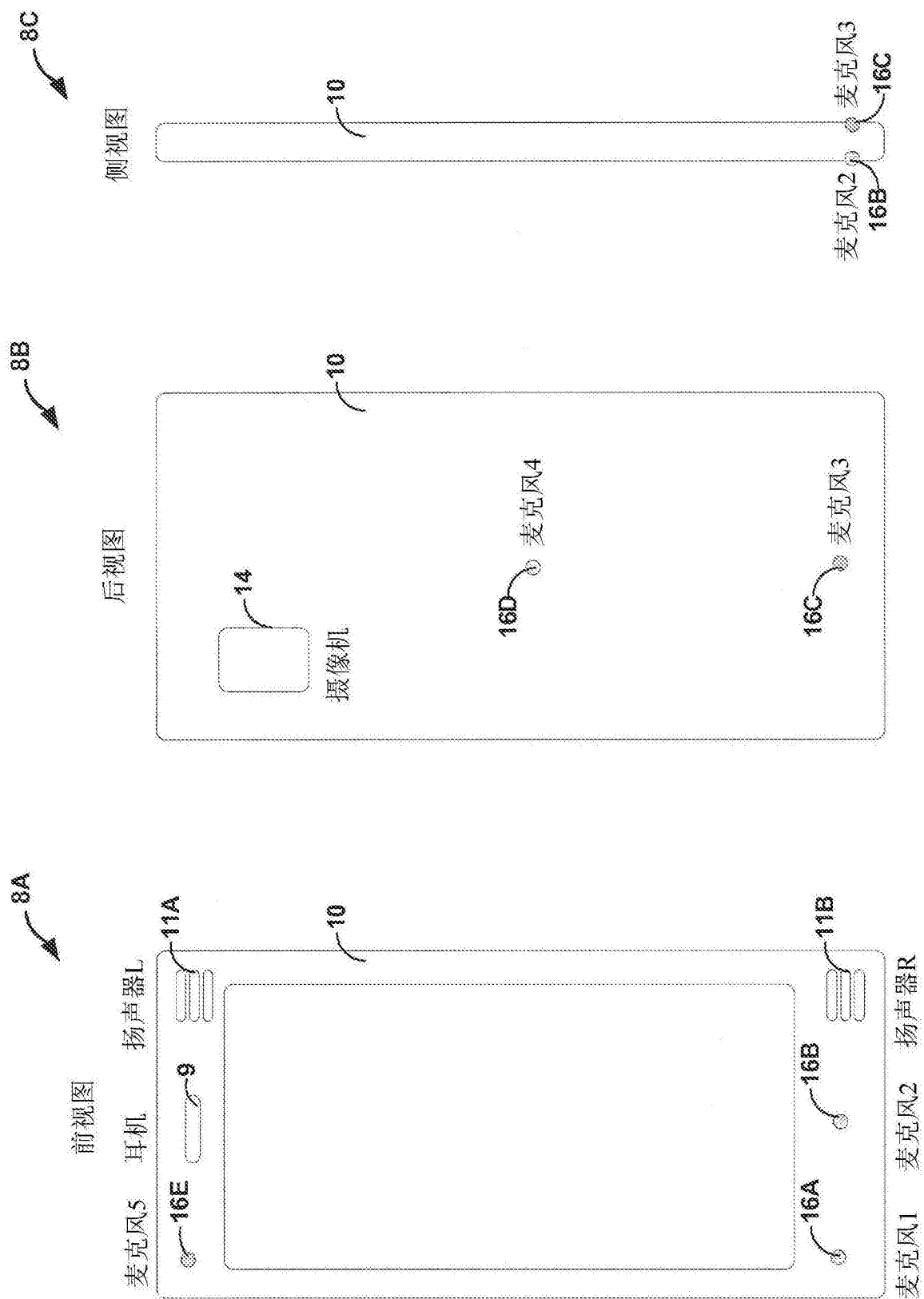


图1A

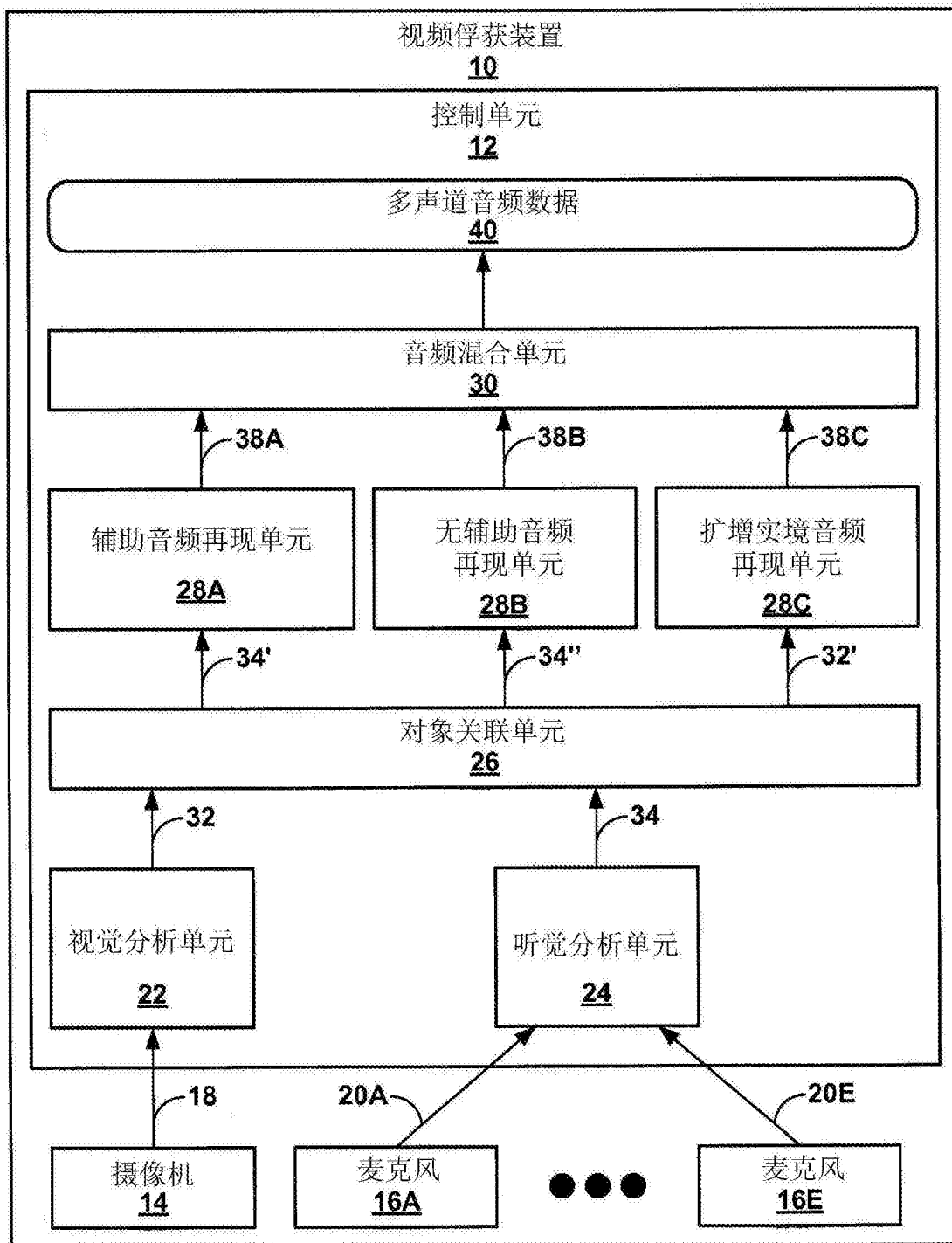


图1B

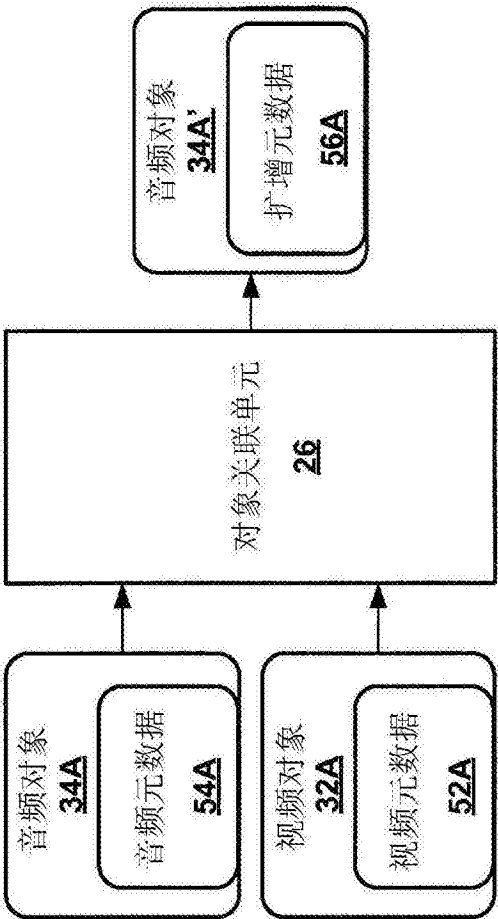


图2A

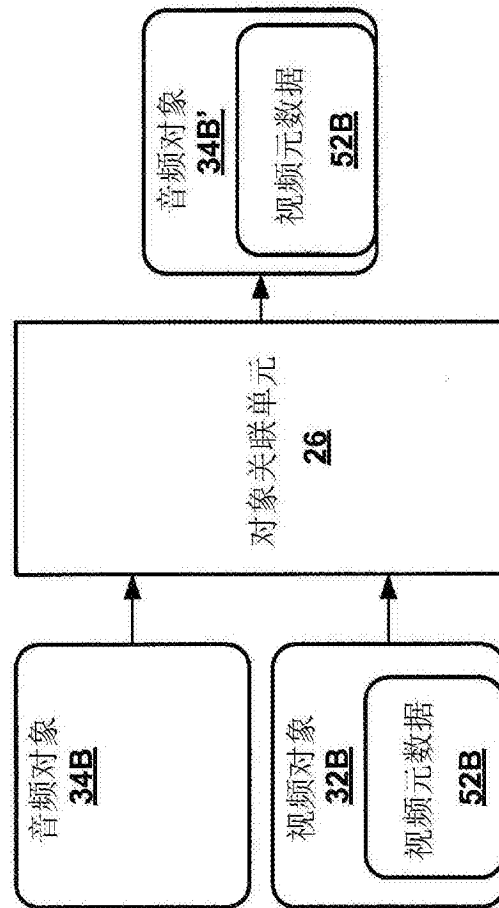


图2B



图2C



图2D

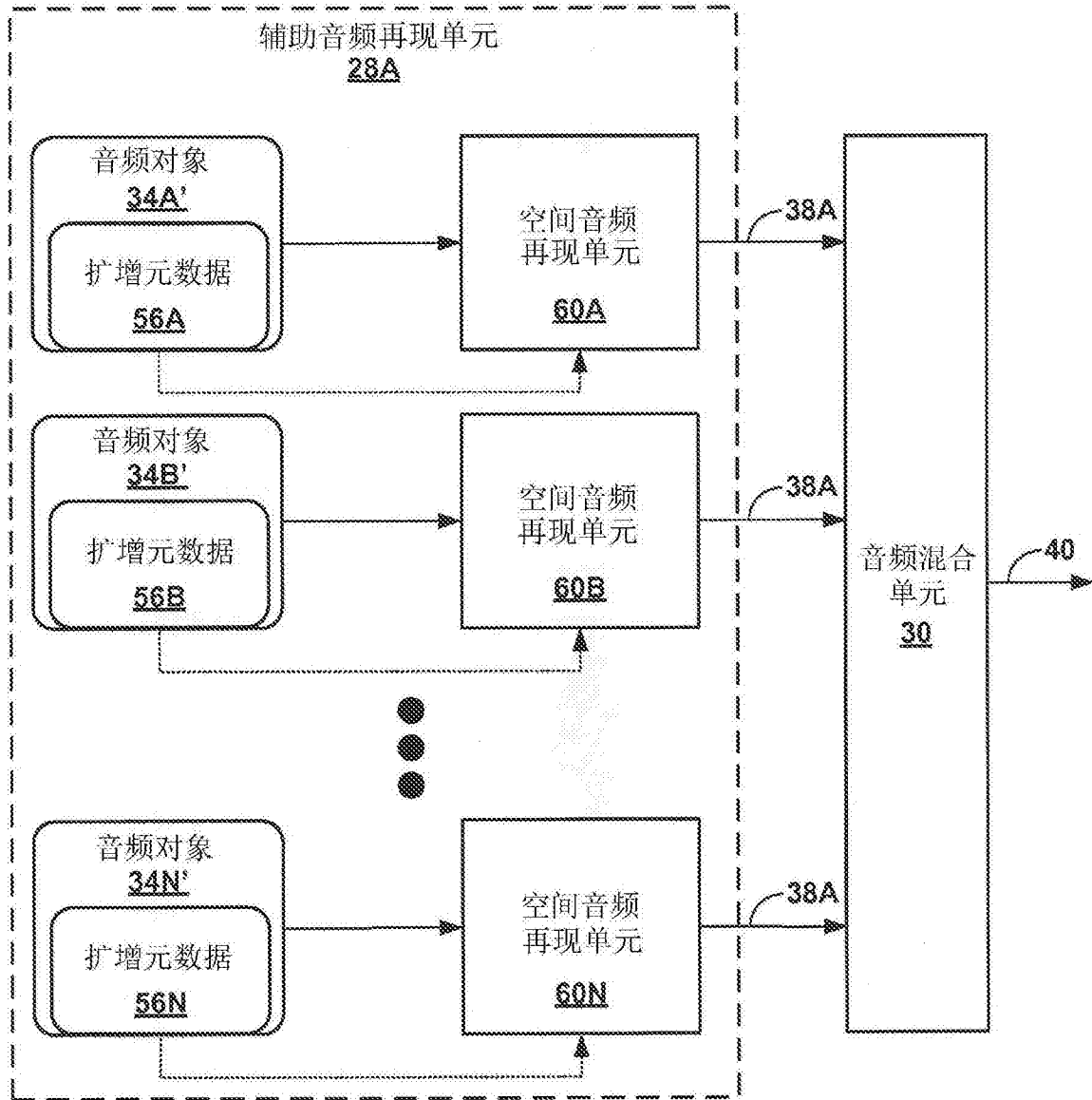


图3

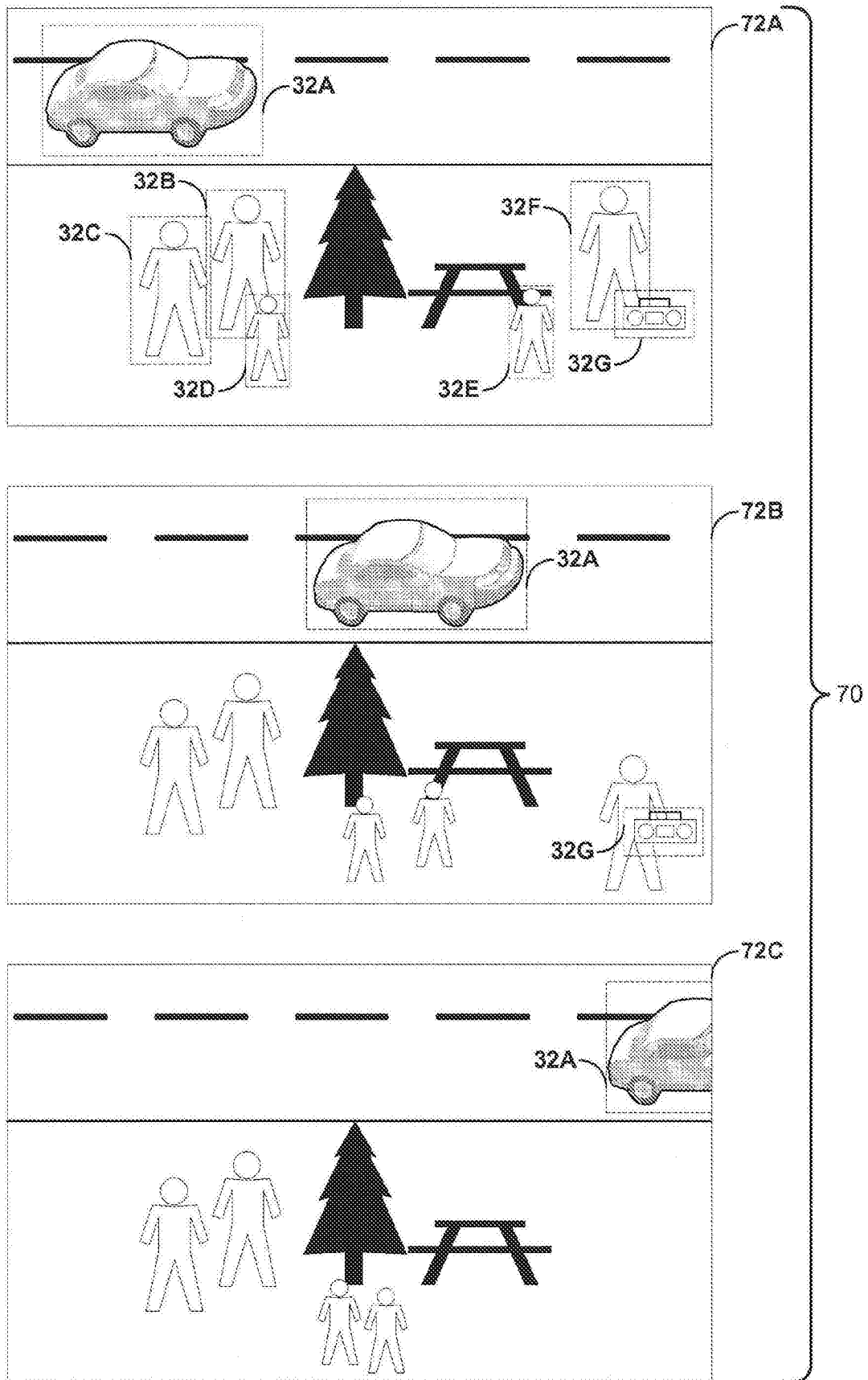


图4

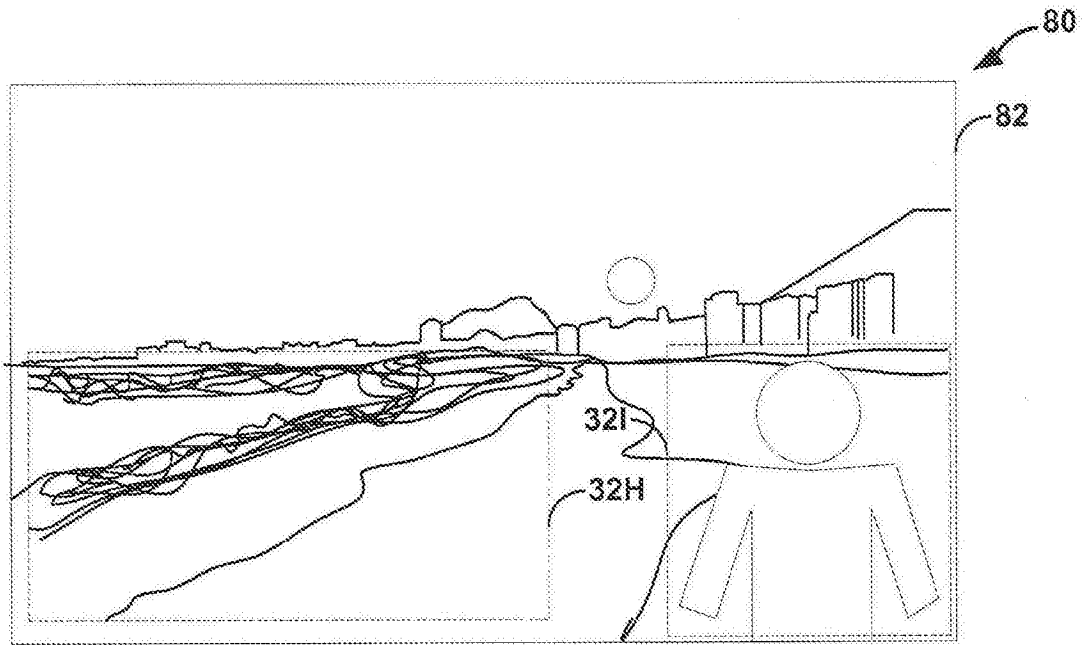


图5

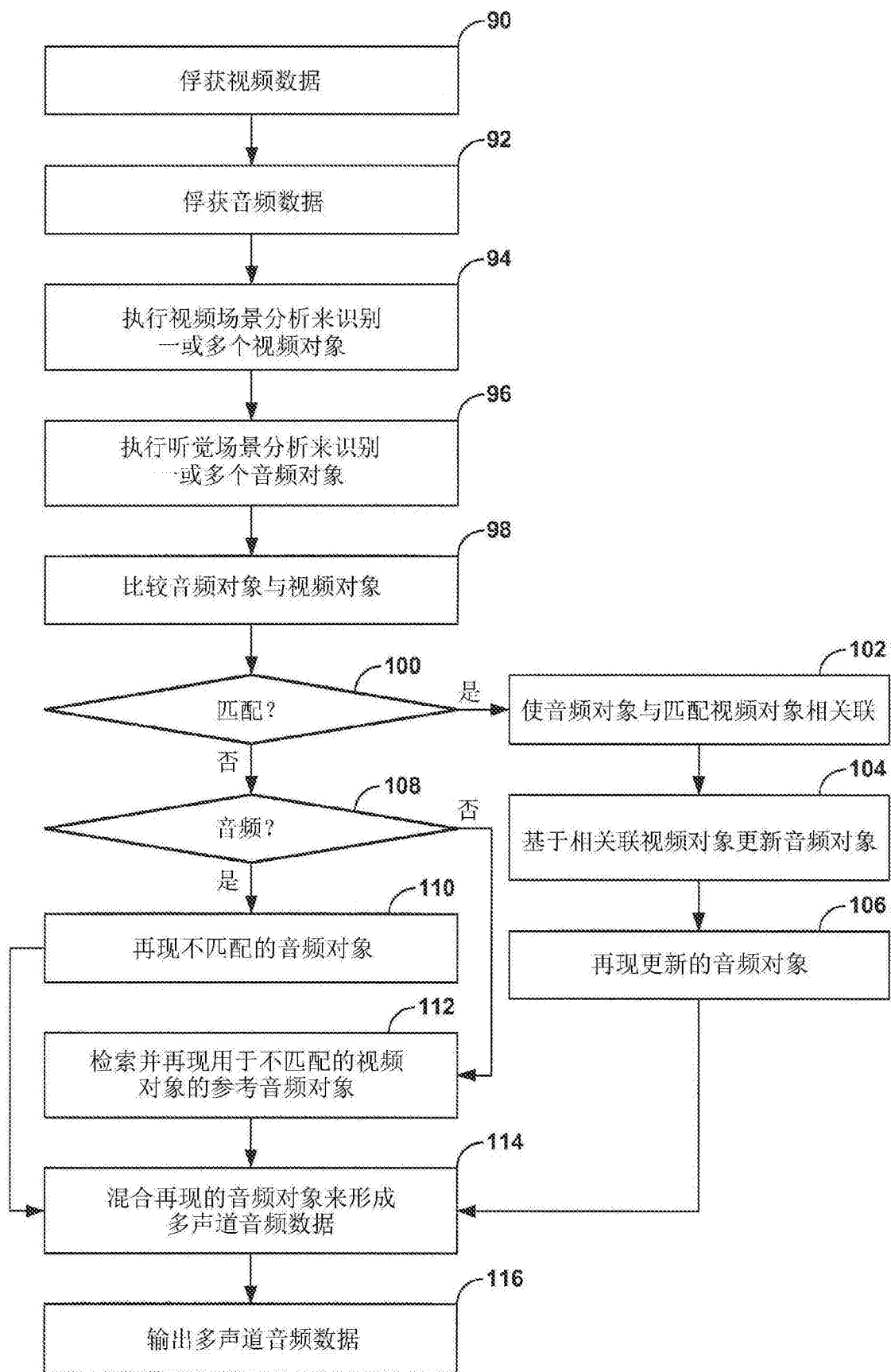


图6

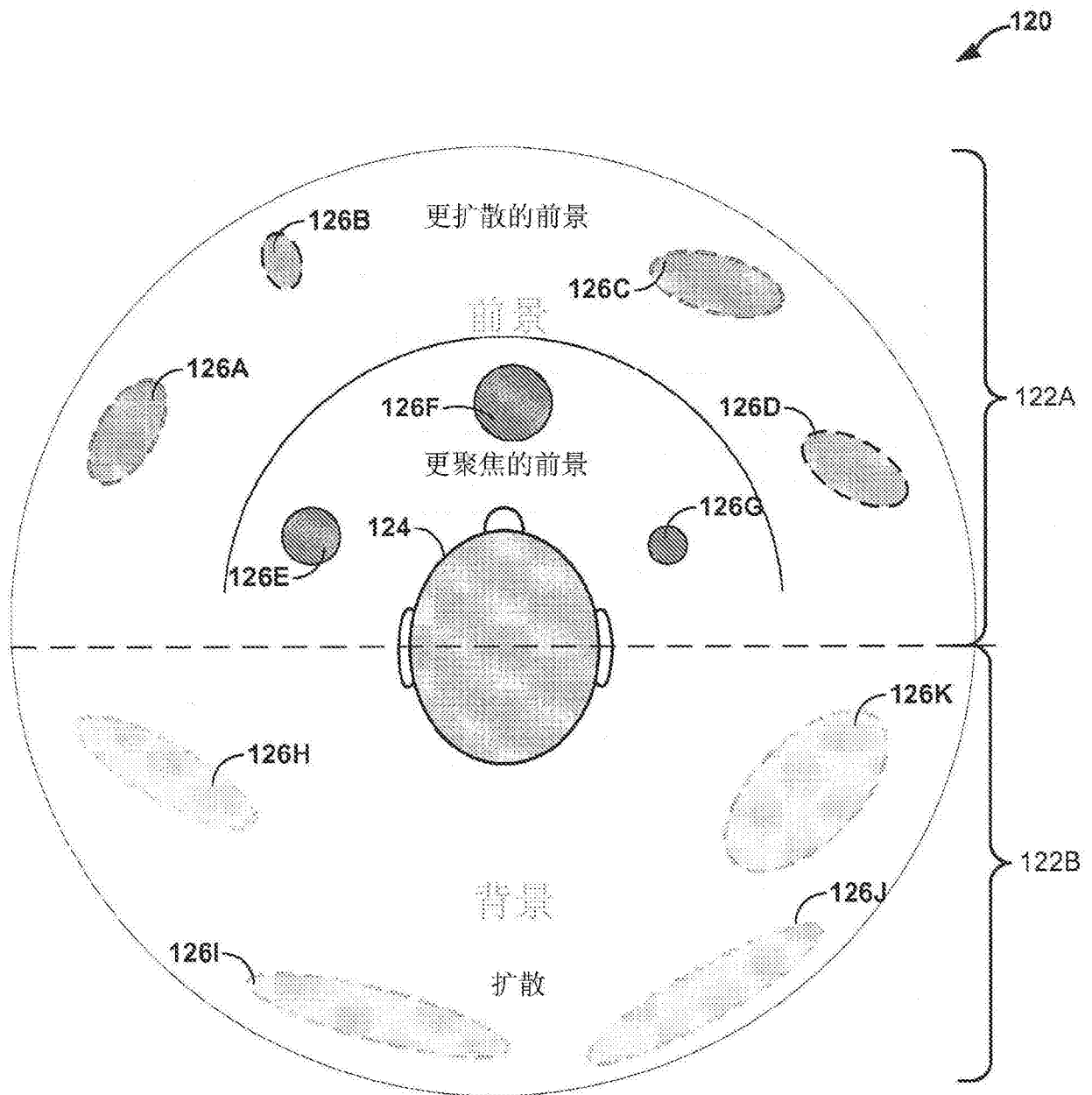


图7