(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108777832 B
(45)授权公告日 2021.02.09

(21)申请号 201810607331.X
(22)申请日 2018.06.13
(65)同一申请的已公布的文献号
申请公布号 CN 108777832 A
(43)申请公布日 2018.11.09
(73)专利权人 上海艺瓣文化传播有限公司
地址 200041 上海市静安区石门二路333弄3号28层E-02室
(72)发明人 王雨霓 秦明昌
(74)专利代理机构 上海科盛知识产权代理有限公司 31225
代理人 翁惠瑜
(51)Int.Cl.
H04R 5/02(2006.01)

审查员 李莎莎

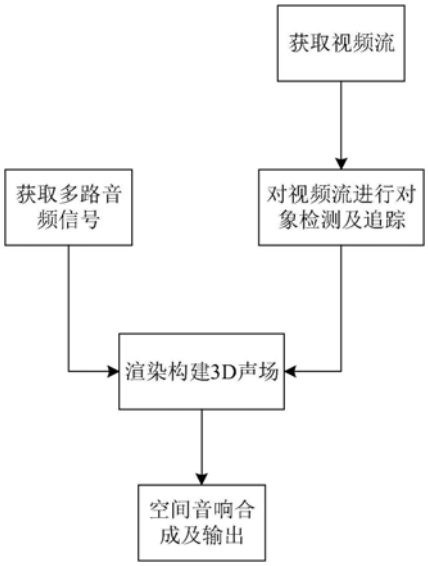
权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于视频对象追踪的实时3D声场构建和混音系统

(57)摘要

本发明涉及一种基于视频对象追踪的实时3D声场构建和混音方法及系统，所述方法包括以下步骤：获取视频流，对该视频流进行对象检测及追踪，形成多个对象的三维空间轨迹，各对象具有对象标签，所述对象包括声源对象和非声源对象；获取多路音频信号，所述音频信号包括实时收音信号和预制音频信号；根据所述对象标签将获得的各对象与音频信号进行匹配，基于所述三维空间轨迹渲染构建3D声场；空间音响合成，生成多种格式输出。与现有技术相比，本发明具有灵活、有效、精确等优点，且系统的各个模块间彼此间信息和数据的共享与交互，便于操作。



CN 108777832 B

1.一种基于视频对象追踪的实时3D声场构建和混音方法,其特征在于,该方法包括以下步骤:

获取视频流,对该视频流进行对象检测及追踪,形成多个对象的三维空间轨迹,各对象具有对象标签,所述对象包括声源对象和非声源对象;

获取多路音频信号,所述音频信号包括实时收音信号和预制音频信号;

根据所述对象标签将获得的各对象与音频信号进行匹配,基于所述三维空间轨迹渲染构建3D声场,基于Ambsonics与Binaural双耳录音算法,采用音源信号在空间传播衰减模型与观测点反向模型进行所述3D声场的构建,同时提供3D声场构建效果的多颗粒度选择;

所述各对象与音频信号进行匹配具体为:

对于声源对象,采用一对一或多对一的方式将各声源对象匹配到一路音频信号中,形成声场中的有效音源;对于非声源对象,根据各非声源对象的特性匹配获得吸音与反射声波系数;

空间音响合成,生成多种格式输出;

所述空间音响合成具体为:

将相关线路匹配到3D声场中带有声源信息的对象上或者一个包含多个声源对象和非声源对象的集合上,并根据声场环境参数,获得一个或多个位置上的声音频谱分布,实现收音线路和预制线路的混音。

2.根据权利要求1所述的基于视频对象追踪的实时3D声场构建和混音方法,其特征在于,通过机器学习方法提取声源相关对象的特征,建立一对象库,基于所述对象库进行对象检测,并利用图像边界追踪技术进行对象追踪。

3.一种基于视频对象追踪的实时3D声场构建和混音系统,其特征在于,该系统包括:

视频对象检测与追踪模块,用于获取视频流,对该视频流进行对象检测及追踪,形成多个对象的三维空间轨迹,各对象具有对象标签,所述对象包括声源对象和非声源对象;

音频信号采集模块,用于获取多路音频信号,所述音频信号包括实时收音信号和预制音频信号;

3D声场构建模块,用于根据所述对象标签将获得的各对象与音频信号进行匹配,基于所述三维空间轨迹渲染构建3D声场,基于Ambsonics与Binaural双耳录音算法,采用音源信号在空间传播衰减模型与观测点反向模型进行所述3D声场的构建,同时提供3D声场构建效果的多颗粒度选择;

终混模块,用于整合以上模块所得到的参数,空间音响合成生成多种格式输出,具体地,将相关线路匹配到3D声场中带有声源信息的对象上或者一个包含多个声源对象和非声源对象的集合上,并根据声场环境参数,获得一个或多个位置上的声音频谱分布,实现收音线路和预制线路的混音;

所述3D声场构建模块中,各对象与音频信号进行匹配具体为:

对于声源对象,采用一对一或多对一的方式将各声源对象匹配到一路音频信号中,形成声场中的有效音源;对于非声源对象,根据各非声源对象的特性匹配获得吸音与反射声波系数。

4.根据权利要求3所述的基于视频对象追踪的实时3D声场构建和混音系统,其特征在于,所述视频对象检测与追踪模块中,通过机器学习方法提取声源相关对象的特征,建立一

对象库，基于所述对象库进行对象检测，并利用图像边界追踪技术进行对象追踪。

一种基于视频对象追踪的实时3D声场构建和混音系统

技术领域

[0001] 本发明涉及一种信号处理技术，尤其是涉及一种基于视频对象追踪的实时3D声场构建和混音系统。

背景技术

[0002] 在当前虚拟现实、电影、游戏娱乐、多媒体展厅等迅速发展与推广的背景下，音视频交互领域成为了关注的重点。然而，在现有的技术条件下，无论是音频、视频还是音视频的交互方面，都没有将每一个对象进行逐一的对象化的处理，因此这就直接造成了无法自动寻找轨迹而需手动跟踪，实时性、用户的交互性体验感较差的尴尬局面，与此同时音频混音也仅局限于一个大的声场的概念下，没有进行精细的划分。在系统层面，也并无一体化的系统可供直接使用。

[0003] 究其原因，主要有二大技术瓶颈：

[0004] (1)视频对象的对象化处理难点：无法对视频对象进行充分的对象化处理，传统的技术在最终呈现手段方面也极为复杂，无法自动追踪声音、位置等信息，而需要手动操作。

[0005] (2)3D声场的实时构建难点：对声场的还原与构建若仅对空间感进行处理，会造成声音的层次感和方位感不清晰，整体混响过大等缺陷。

[0006] 目前已知的技术和系统均无法彻底解决以上技术难点。

发明内容

[0007] 本发明的目的就是为了克服上述现有技术存在的缺陷而提供一种更为灵活、有效和精确的、适用于音视频交互领域的基于视频对象追踪的实时3D声场构建和混音系统。

[0008] 本发明的目的可以通过以下技术方案来实现：

[0009] 一种基于视频对象追踪的实时3D声场构建和混音方法，该方法包括以下步骤：

[0010] 获取视频流，对该视频流进行对象检测及追踪，形成多个对象的三维空间轨迹，各对象具有对象标签，所述对象包括声源对象和非声源对象；

[0011] 获取多路音频信号，所述音频信号包括实时收音信号和预制音频信号；

[0012] 根据所述对象标签将获得的各对象与音频信号进行匹配，基于所述三维空间轨迹渲染构建3D声场；

[0013] 空间音响合成，生成多种格式输出。

[0014] 进一步地，通过机器学习方法提取声源相关对象的特征，建立一对象库，基于所述对象库进行对象检测，并利用图像边界追踪技术进行对象追踪。

[0015] 进一步地，所述各对象与音频信号进行匹配具体为：

[0016] 对于声源对象，采用一对一或多对一的方式将各声源对象匹配到一路音频信号中，形成声场中的有效音源；对于非声源对象，根据各非声源对象的特性匹配获得吸音与反射声波系数。

[0017] 进一步地，采用音源信号在空间传播衰减模型与观测点反向模型进行所述3D声场

的构建,同时提供多颗粒度选择。

[0018] 进一步地,所述空间音响合成具体为:

[0019] 将相关线路匹配到3D声场中带有声源信息的对象上或者一个包含多个声源对象和非声源对象的集合上,并根据声场环境参数,获得一个或多个位置上的声音频谱分布,实现收音线路和预制线路的混音。

[0020] 一种基于视频对象追踪的实时3D声场构建和混音系统,该方法包括以下步骤:

[0021] 视频对象检测与追踪模块,用于获取视频流,对该视频流进行对象检测及追踪,形成多个对象的三维空间轨迹,各对象具有对象标签,所述对象包括声源对象和非声源对象;

[0022] 音频信号采集模块,用于获取多路音频信号,所述音频信号包括实时收音信号和预制音频信号;

[0023] 3D声场构建模块,用于根据所述对象标签将获得的各对象与音频信号进行匹配,基于所述三维空间轨迹渲染构建3D声场;

[0024] 终混模块,用于整合以上模块所得到的参数,空间音响合成,生成多种格式输出。

[0025] 进一步地,所述视频对象检测与追踪模块中,通过机器学习方法提取声源相关对象的特征,建立一对象库,基于所述对象库进行对象检测,并利用图像边界追踪技术进行对象追踪。

[0026] 进一步地,所述3D声场构建模块中,各对象与音频信号进行匹配具体为:

[0027] 对于声源对象,采用一对一或多对一的方式将各声源对象匹配到一路音频信号中,形成声场中的有效音源;对于非声源对象,根据各非声源对象的特性匹配获得吸音与反射声波系数。

[0028] 进一步地,所述3D声场构建模块中,采用音源信号在空间传播衰减模型与观测点反向模型进行所述3D声场的构建,同时提供多颗粒度选择。

[0029] 进一步地,所述终混模块中,将相关线路到3D声场中带有声源信息的对象上或者一个包含多个声源对象和非声源对象的集合上,并根据声场环境参数,获得一个或多个位置上的声音频谱分布,实现收音线路和预制线路的混音。

[0030] 与现有技术相比,本发明具有以下有益效果:

[0031] (1)本发明基于对象与音频信号的融合形成3D声场,具有更好的准确性和精确度且具有较高的灵活性,所生成的3D声场具有高度的三维空间感和沉浸感,是对声场的高度的还原与再现。

[0032] (2)本发明的视频对象追踪能够有效的解决传统手动跟踪方式操作复杂的难题,其基于对声源对象与非声源对象的定义、分类与学习,追踪产生对象相对视频空间的左右,上下和景深三个参数相对于视频时间变量的函数,并且根据视频播放格式进行转换。

[0033] (3)本发明系统各模块采用多项技术相结合,构成一个完整的系统集成,实现各个模块间彼此间信息和数据的共享与交互便于操作,且充分发挥了各个模块间的交互作用。

附图说明

[0034] 图1为本发明的流程示意图。

具体实施方式

[0035] 下面结合附图和具体实施例对本发明进行详细说明。本实施例以本发明技术方案为前提进行实施，给出了详细的实施方式和具体的操作过程，但本发明的保护范围不限于下述的实施例。

[0036] 如图1所示，本发明提供一种基于视频对象追踪的实时3D声场构建和混音方法，该方法包括以下步骤：获取视频流，对该视频流进行对象检测及追踪，形成多个对象的三维空间轨迹，各对象具有对象标签，所述对象包括声源对象和非声源对象；获取多路音频信号，所述音频信号包括实时收音信号和预制音频信号；根据所述对象标签将获得的各对象与音频信号进行匹配，基于所述三维空间轨迹渲染构建3D声场；空间音响合成，生成多种格式输出。

[0037] (1)视频对象检测与追踪

[0038] 在进行对象检测与追踪前期，通过机器学习方法(如监督学习、深度学习、迁移学习等)在图片和视频参考库中提取声源相关对象的特征，建立一定规模的对象库。获取视频流后，基于所述对象库在视频帧上对对象进行识别，形成场景动态分割并且链接对应语义，设置每一对象的对象标签。

[0039] 在视频流中采用图像边界追踪技术等方法对对象进行跟踪并且形成三维空间上的轨迹。该方法也支持特定对象的手动标注功能。采用的图像边界追踪技术具体步骤为：

[0040] a)选取图像中的坐标原点，并标记为P0(0,0)，最靠近P0即为最小行数与最小列数的像素。

[0041] b)定义变量dir为边界移动的方向，存储从前一个边界元素到当前元素沿着边界的移动方向。同时判断dir的奇偶性，同时对dir进行计算更改。

[0042] c)边界是一个闭合空间，搜索到首尾相接时即结束。同时，删除重复计算的Pn-1与Pn，得出最后边界为P0到Pn-2。

[0043] 通过上述对象追踪可以获得对象相对视频空间的左右，上下和景深三个参数相对于视频时间变量的函数，并且根据视频播放格式进行转换。

[0044] (2)音频信号采集

[0045] 接收音频信号，包括实时收音信号也可以是预制音频信号。

[0046] (3)3D声场构建

[0047] 根据所述对象标签将获得的各对象与音频信号进行匹配，有些对象为音源对象，有些为非音源对，基于所述三维空间轨迹渲染构建3D声场。对于声源对象，采用一对一或多对一的方式将各声源对象匹配到一路音频信号中，形成声场中的有效音源；对于非声源对象，作为声场中的障碍对象，根据各非声源对象的特性匹配获得吸音与反射声波系数。声场构建采用音源信号在空间传播衰减模型与观测点反向模型，同时提供多颗粒度选择。

[0048] (4)混音

[0049] 把相关线路，即各路收音信号和预制线路信号匹配到声场中带有声源信息的对象或者一个包含多个声源对象和非声源对象的集合，并且依据声场环境参数，即对各非声源对象的特性匹配获得吸音与反射声波系数分析所得出的空间参数，从而给出在一个特定位置或者多个位置上声音频谱分布，实现对实时线路与预制线路的混音。

[0050] 实现上述基于视频对象追踪的实时3D声场构建和混音方法的系统包括视频对象

检测与追踪模块、音频信号采集模块、3D声场构建模块和终混模块，多个模块以及多项技术相结合，使其构成一个完整的系统集成，实现各个模块间彼此间信息和数据的共享与交互。其中，所述视频对象检测与追踪基于对声源对象与非声源对象的定义、分类与学习，并形成三维空间(包括左右、上下、景深三个参数)上的轨迹，同时可根据视频播放格式进行转换；音频信号采集模块用于实时接收音频信号以及预制音频信号；3D声场实时构建模块用于将视频对象检测与追踪模块所获取的各对象匹配到声场中关键元素，同时对声场中的有效声源对象和和无声对象进行相应的数据处理以获得多维的环境参；终混模块整合以上模块所得到的各参数，把相关线路匹配到声场中带有声源信息的对象或者一个包含多个声源对象和非声源对象的集合，并且依据声场环境参数，给出一个特定位置或者多个位置上声音频谱分布，实现收音线路和预制线路的混音。

[0051] 在某些实施例中，上述3D声场实时构建模块和终混模块可以采用Max来实现。Max程序主要分为以下几个功能：

[0052] 1)收音及预制声音

[0053] 实时拾取现场声源和采集预制声音，并将这些声源实时匹配至相关线路，并发送至声场重建效果模块。

[0054] 2)3D声场构建效果

[0055] 基于Ambsonics与Binaural双耳录音算法，采用音源信号在空间传播衰减模型与观测点反向模型，同时提供多颗粒度选择。有声对象通过一对一、多对一的方式匹配到一路输入线路，形成声场中有效音源对象；无声对象最为声场中的障碍对象，并根据对象特性匹配吸音与反射声波系数。最终将拾取到的声源实时转制成可供用户在佩戴耳机的回放条件下试听的3D立体声音频文件。

[0056] 3)混音

[0057] 对声音拾取和声场重建信息中不足的声音频段进行弥补与对特殊声音效果的加强。

[0058] 以上详细描述了本发明的较佳具体实施例。应当理解，本领域的普通技术人员无需创造性劳动就可以根据本发明的构思作出诸多修改和变化。因此，凡本技术领域中技术人员依本发明的构思在现有技术的基础上通过逻辑分析、推理或者有限的实验可以得到的技术方案，皆应在由权利要求书所确定的保护范围内。

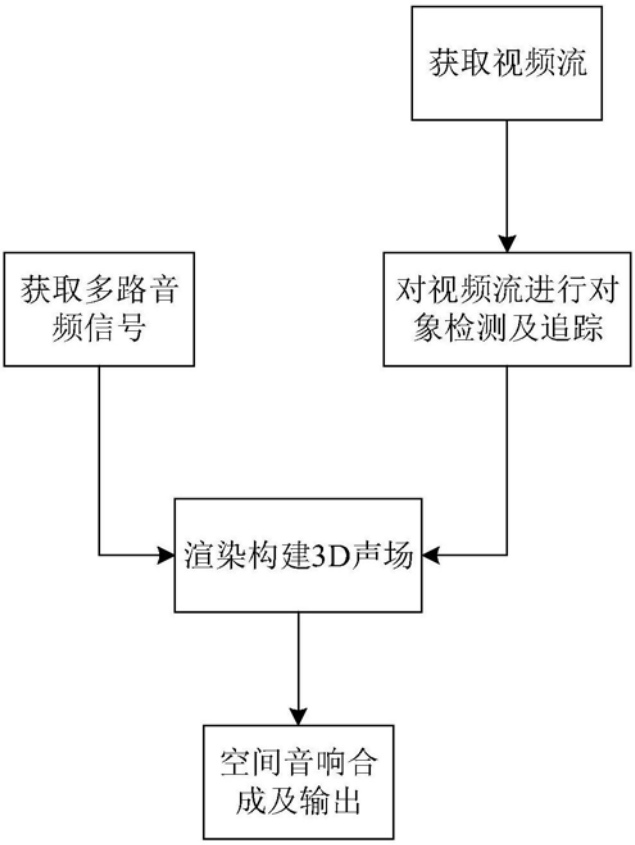


图1