(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **3 013 574**

(51) Int. Cl.:

G06F 3/01 (2006.01)

G06F 3/16 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **29.03.2022 PCT/EP2022/058273**

(87) Fecha y número de publicación internacional: **13.10.2022 WO22214357**

(96) Fecha de presentación y número de la solicitud europea: **29.03.2022 E 22712436 (9)**

(97) Fecha y número de publicación de la concesión europea: **29.01.2025 EP 4320498**

(54) Título: **Aparato y procedimiento para generar una señal de audio**

(30) Prioridad:

08.04.2021 EP 21167514

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.04.2025

(73) Titular/es:

KONINKLIJKE PHILIPS N.V. (100.00%)
High Tech Campus 52
5656 AG Eindhoven, NL

(72) Inventor/es:

VAREKAMP, CHRISTIAAN y
KOPPENS, JEROEN GERARDUS HENRICUS

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Aparato y procedimiento para generar una señal de audio

Campo de la invención

La invención se refiere a un aparato y un procedimiento para generar una señal de audio para interacciones entre objetos de escena del mundo virtual y real, y en particular, pero no exclusivamente, para generar una señal de audio para aplicaciones de Realidad Aumentada.

Antecedentes de la invención

La variedad y el alcance de experiencias basadas en contenido audiovisual han aumentado sustancialmente en los últimos años con nuevos servicios y formas de utilizar y consumir dicho contenido que se desarrollan e introducen continuamente. En particular, se están desarrollando muchos servicios, aplicaciones y experiencias espaciales e interactivas para brindar a los usuarios una experiencia más completa e inmersiva.

Ejemplos de tales aplicaciones son aplicaciones de Realidad Virtual (VR), Realidad Aumentada (AR) y Realidad Mixta (MR), que se están convirtiendo rápidamente en algo común, con una serie de soluciones dirigidas al mercado de consumo. Un número de normas también se encuentran en desarrollo por un número de organismos de normalización. Tales actividades de estandarización están desarrollando activamente estándares para los diversos aspectos de los sistemas de VR/AR/MR, que incluyen, por ejemplo, transmisión en directo, radiodifusión, renderización, etc.

Las aplicaciones de VR tienden a proporcionar experiencias de usuario correspondientes a que el usuario esté en un mundo/entorno/escena diferente, mientras que las aplicaciones de AR (que incluyen la Realidad Mixta MR) tienden a proporcionar experiencias de usuario correspondientes a que el usuario esté en el entorno local real, pero con información u objetos virtuales adicionales que se añaden. Por lo tanto, las aplicaciones de VR tienden a proporcionar un mundo/escena completamente inmersiva generada sintéticamente, mientras que las aplicaciones de AR tienden a proporcionar un mundo/escena parcialmente sintético que se superpone a la escena real en la que el usuario está físicamente presente. Sin embargo, los términos a menudo se usan indistintamente y tienen un alto grado de superposición. En lo siguiente, el término Realidad Aumentada/AR se usará para denotar tanto la Realidad Aumentada como la Realidad Mixta (así como también algunas variaciones a veces denominadas Realidad Virtual). Un ejemplo de una aplicación de MR se divulga en US2021/0082191.

Como ejemplo, los servicios y aplicaciones para la realidad aumentada mediante el uso de dispositivos portátiles se han vuelto cada vez más populares y se han introducido API de software (interfaces de programación de aplicaciones) y kits de herramientas, tales como ARKit (desarrollado por Apple Inc.) y ARCore (desarrollado por Google Inc.) para admitir aplicaciones de realidad aumentada en teléfonos inteligentes y tabletas. En estas aplicaciones, las cámaras integradas y otros sensores de los dispositivos se usan para generar imágenes en tiempo real del entorno con gráficos virtuales que se superponen a las imágenes presentadas. Las aplicaciones pueden, por ejemplo, generar una transmisión de video en vivo con objetos gráficos superpuestos al video en vivo. Tales objetos gráficos pueden usarse, por ejemplo, para colocar objetos virtuales de manera que se perciban como presentes en la escena del mundo real.

Como otro ejemplo, se han desarrollado auriculares y gafas donde la escena del mundo real puede verse directamente a través de las gafas de AR, pero estas también son capaces de generar imágenes que el usuario ve al mirar a través de las gafas. Esto también puede usarse para presentar imágenes virtuales que se perciben como parte de la escena del mundo real vista. Los sensores de movimiento se usan para rastrear los movimientos de la cabeza y el objeto virtual presentado puede adaptarse correspondientemente para proporcionar una impresión de que el objeto virtual es un objeto del mundo real visto en el mundo real.

Los enfoques se conocen respectivamente como transparente y transparente y ambos pueden proporcionar una experiencia de usuario novedosa y emocionante.

Además de la representación visual, algunas aplicaciones de AR pueden proporcionar una experiencia de audio correspondiente. Además, a la experiencia visual, se ha propuesto proporcionar audio que puede corresponder a objetos virtuales. Por ejemplo, si un objeto virtual es uno que generaría un ruido, la presentación visual del objeto puede complementarse con un sonido correspondiente que se genera. En algunos casos, también puede generarse un sonido para reflejar una acción para el objeto virtual.

Típicamente, el sonido se genera como un sonido predeterminado al reproducir un clip de audio pregrabado en momentos apropiados. El sonido puede adaptarse en algunos casos para reflejar el entorno actual, tal como por ejemplo adaptando la reverberación percibida en dependencia del entorno actual, o puede, por ejemplo, procesarse para que se perciba que llega desde una posición correspondiente a la posición percibida del objeto virtual en el mundo real. A menudo, tal posicionamiento puede lograrse mediante el procesamiento binaural para generar una salida de audio de auriculares adecuada.

Sin embargo, aunque tales enfoques pueden proporcionar una aplicación y experiencia de usuario interesantes en muchas realizaciones, los enfoques convencionales tienden a ser subóptimos, y tienden a ser difíciles de implementar y/o tienden a proporcionar un rendimiento y experiencias de usuario subóptimos.

Por lo tanto, un enfoque mejorado sería ventajoso. En particular, un enfoque para generar una señal de audio que permita un funcionamiento mejorado, mayor flexibilidad, menor complejidad, implementación facilitada, una experiencia de audio mejorada, calidad de audio mejorada, menor carga computacional, mayor idoneidad y/o rendimiento para aplicaciones de realidad mixta/aumentada, mayor inmersión del usuario y/o mayor rendimiento y/u operación sería ventajoso.

Sumario de la invención

En consecuencia, la invención busca mitigar, aliviar o eliminar preferentemente una o más de las desventajas mencionadas anteriormente individualmente o en cualquier combinación.

De acuerdo con un aspecto de la invención, se proporciona un aparato para generar una señal de audio de salida, el aparato que comprende: un primer receptor dispuesto para recibir una secuencia de imágenes en tiempo real de una escena del mundo real de un sensor de imagen, la secuencia de imágenes en tiempo real que comprende una secuencia de tramas de imágenes, cada trama de imagen que comprende al menos uno de los datos de imagen visual y los datos de imagen de profundidad; un segundo receptor dispuesto para recibir un conjunto de objetos de audio y metadatos para objetos de audio del conjunto de objetos de audio, los metadatos son indicativos de enlaces entre objetos de audio del conjunto de objetos de audio y las características del material; un generador de imágenes dispuesto para generar una secuencia de imágenes de salida que comprende un objeto de imagen correspondiente a un objeto de escena virtual en la escena del mundo real; un detector dispuesto para detectar una interacción entre el objeto de escena virtual y un objeto de escena del mundo real de la escena del mundo real en respuesta a una detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real en un sistema de coordenadas tridimensional que representa la escena del mundo real; un estimador dispuesto para determinar una propiedad del material para el objeto de escena del mundo real en base a los datos de imagen comprendidos en las tramas de imágenes de la secuencia de tramas de imágenes; un selector dispuesto para seleccionar un primer objeto de audio del conjunto de objetos de audio en respuesta a la propiedad del material y las características del material vinculadas a objetos de audio del conjunto de objetos de audio; un circuito de salida dispuesto para generar la señal de audio de salida que comprende el primer objeto de audio; y un circuito dispuesto para presentar la secuencia de imágenes de salida a un usuario.

La invención puede proporcionar una experiencia de usuario mejorada en muchas realizaciones y puede, en particular, en muchas realizaciones, proporcionar una experiencia de Realidad Aumentada mejorada y más inmersiva. El enfoque puede lograrse en muchas realizaciones mientras se mantiene una complejidad y/o implementación baja. El enfoque puede ser muy adecuado para un sistema AR donde el audio y los metadatos de soporte pueden proporcionarse mediante un servidor remoto. El enfoque puede facilitar y/o admitir un sistema donde un servidor centralizado puede proporcionar soporte para un número de clientes remotos y puede facilitar sustancialmente la implementación del cliente remoto. El enfoque puede admitir una generación y gestión centralizada de audio para mejorar una aplicación y experiencia de AR.

Los objetos de audio pueden ser fragmentos/ clips de audio/ etc. y pueden representarse de cualquier manera adecuada. En muchas realizaciones, cada objeto de audio puede representar un sonido en un intervalo de tiempo. En muchas realizaciones, el intervalo de tiempo no puede exceder 5 segundos, 10 segundos o 20 segundos para cualquiera de los objetos de audio.

Las características del material pueden ser características del material de objetos del mundo real.

La secuencia de imágenes de salida puede comprender imágenes visuales.

La interacción puede detectarse en un sistema de coordenadas tridimensional que representa la escena del mundo real. La interacción puede ser una proximidad/colisión/contacto entre el objeto de escena virtual y el objeto de escena del mundo real.

El aparato puede ser un aparato de realidad aumentada. El aparato puede ser un aparato para proporcionar una secuencia de imágenes de salida y una señal de audio de salida para una aplicación de realidad aumentada. La aplicación de realidad aumentada puede presentar un objeto de escena virtual en una escena del mundo real.

De acuerdo con una característica opcional de la invención, el estimador está dispuesto para: determinar una región de imagen de interacción en al menos una trama de imagen de la secuencia de tramas de imagen, la región de imagen de interacción es una región de imagen de la al menos una trama de imagen en la que ocurre la interacción; y determinar la propiedad del material para el objeto de escena en respuesta a los datos de imagen de la región de imagen de interacción.

Esto puede proporcionar una estimación de propiedades de material particularmente eficiente y/o ventajosa en muchas realizaciones, y puede permitir específicamente una estimación de propiedades de material más precisa en muchas realizaciones. El enfoque puede proporcionar una experiencia de usuario mejorada como resultado.

De acuerdo con una característica opcional de la invención, el segundo receptor se dispone para recibir los metadatos de un servidor remoto.

El enfoque puede proporcionar una aplicación particularmente eficiente donde el audio puede generarse y gestionarse de forma remota y posiblemente centralmente mientras se adapta eficazmente a las condiciones actuales apropiadas.

De acuerdo con una característica opcional de la invención, los metadatos para al menos algunos objetos de audio comprenden indicaciones de enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena del mundo real y enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena virtual; y en el que el selector se dispone para seleccionar el primer objeto de audio en respuesta a la propiedad material y características materiales de objetos del mundo real vinculados al conjunto de objetos de audio y en respuesta a una propiedad material del objeto de escena virtual y características materiales de objetos de escena virtual vinculados al conjunto de objetos de audio.

Esto puede permitir un rendimiento mejorado en muchas realizaciones y puede permitir específicamente una adaptación mejorada a la interacción específica. A menudo puede lograrse una experiencia de usuario más inmersiva.

De acuerdo con una característica opcional de la invención, el selector está dispuesto para seleccionar el primer objeto de audio en respuesta a una propiedad dinámica del objeto de escena virtual.

Esto puede permitir un rendimiento mejorado en muchas realizaciones y puede permitir específicamente una adaptación mejorada a la interacción específica. A menudo puede lograrse una experiencia de usuario más inmersiva.

De acuerdo con una característica opcional de la invención, el detector está dispuesto para determinar una propiedad de la interacción y el selector está dispuesto para seleccionar el primer objeto de audio en respuesta a la propiedad de la interacción.

Esto puede permitir un rendimiento mejorado en muchas realizaciones y puede permitir específicamente una adaptación mejorada a la interacción específica. A menudo puede lograrse una experiencia de usuario más inmersiva.

De acuerdo con una característica opcional de la invención, la propiedad de la interacción es al menos una propiedad seleccionada del grupo de: una velocidad de la interacción; una fuerza de una colisión entre el objeto de escena virtual y el objeto de escena del mundo real; una elasticidad de una colisión entre el objeto de escena virtual y el objeto de escena del mundo real; una duración de la interacción; y una dirección de movimiento del objeto de escena virtual con relación al objeto de escena del mundo real.

De acuerdo con una característica opcional de la invención, el selector está dispuesto para seleccionar el primer objeto de audio en respuesta a una orientación del objeto virtual con relación al objeto de la escena del mundo real.

Esto puede permitir un rendimiento mejorado en muchas realizaciones y puede permitir específicamente una adaptación mejorada a la interacción específica. A menudo puede lograrse una experiencia de usuario más inmersiva.

De acuerdo con una característica opcional de la invención, el estimador está dispuesto para determinar una indicación de coincidencia para el objeto de escena del mundo real a al menos una primera categoría de una pluralidad de categorías de objetos; y para determinar la propiedad del material en respuesta a la indicación de coincidencia y las propiedades del material vinculadas a las categorías de objetos.

Esto puede proporcionar una determinación particularmente ventajosa y a menudo de menor complejidad de una propiedad del material que, sin embargo, puede seguir siendo de alta precisión. En muchas realizaciones, la categorización/clasificación puede lograrse ventajosamente mediante el uso de una red neuronal.

De acuerdo con una característica opcional de la invención, el aparato comprende además un receptor de audio para recibir una señal de audio de audio en tiempo real capturada en la escena del mundo real, y en el que el estimador se dispone para determinar la indicación de coincidencia en respuesta a la señal de audio.

Tal enfoque puede en muchas realizaciones mejorar sustancialmente la precisión de la estimación de las propiedades del material, lo que conduce a un rendimiento general mejorado.

De acuerdo con una característica opcional de la invención, el selector está dispuesto para seleccionar el primer objeto de audio como un objeto de audio predeterminado si no se detecta ningún objeto de audio para el cual se cumpla un criterio de selección.

De acuerdo con una característica opcional de la invención, al menos una trama de imagen comprende datos de imagen en profundidad y en el que el estimador se dispone para determinar la propiedad del material para el objeto de escena del mundo real en respuesta a una detección de que al menos parte de una región de imagen de la al menos una trama de imagen que representa el objeto de escena del mundo real tiene un nivel de confianza para los datos de imagen en profundidad que no excede un umbral.

En algunas realizaciones, la trama de imagen comprende datos de imagen visual y datos de imagen de profundidad y el estimador está dispuesto para determinar que el objeto de escena del mundo real tiene un componente metálico en respuesta a una detección de que para al menos parte de la región de la imagen un brillo de los datos de imagen visual excede un umbral y un nivel de confianza para los datos de imagen de profundidad no excede un umbral.

De acuerdo con otro aspecto de la invención, se proporciona un procedimiento para generar una señal de audio de salida, el procedimiento comprende: recibir una secuencia de imágenes en tiempo real de una escena del mundo real de un sensor de imagen, la secuencia de imágenes en tiempo real comprende una secuencia de tramas de imágenes, cada trama de imagen comprende al menos uno de los datos de imagen visual y los datos de imagen de profundidad; recibir un conjunto de objetos de audio y metadatos para objetos de audio del conjunto de objetos de audio, los metadatos son indicativos de enlaces entre objetos de audio del conjunto de objetos de audio y las características del material; generar una secuencia de imágenes de salida que comprende un objeto de imagen correspondiente a un objeto de escena virtual en la escena del mundo real; detectar una interacción entre el objeto de escena virtual y un objeto de escena del mundo real de la escena del mundo real en respuesta a una detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real en un sistema de coordenadas tridimensional que representa la escena del mundo real; determinar una propiedad del material para el objeto de escena del mundo real en base a los datos de imagen comprendidos en las tramas de imágenes de la secuencia de tramas de imágenes; seleccionar un primer objeto de audio del conjunto de objetos de audio en respuesta a la propiedad del material y las características del material vinculadas a objetos de audio del conjunto de objetos de audio; generar la señal de audio de salida que comprende el primer objeto de audio; y presentar la secuencia de imágenes de salida al usuario.

El procedimiento puede incluir mostrar la secuencia de imágenes de salida y/o renderizar la señal de audio de salida.

Estos y otros aspectos, características y ventajas de la invención serán evidentes y se describirán con referencia a la(s) realización(es) descrita(s) en la presente descripción.

Breve descripción de las figuras

Las realizaciones de la invención se describirán, a manera de ejemplo solamente, con referencia a las figuras, en las que

La Figura 1 ilustra un ejemplo de elementos de un sistema de realidad aumentada;
La Figura 2 ilustra un ejemplo de un aparato de audio para generar una señal de audio de salida de acuerdo con algunas realizaciones de la invención;
La Figura 3 ilustra un ejemplo de una imagen de una escena del mundo real con un objeto de escena virtual; y
La Figura 4 ilustra un ejemplo de un enfoque para generar una señal de audio de salida de acuerdo con algunas realizaciones de la invención.

Descripción detallada de algunas realizaciones de la invención

La siguiente descripción se centrará en la generación de una señal de audio para complementar la generación de imágenes de un objeto virtual en una escena del mundo real como parte de una aplicación de realidad aumentada. Sin embargo, se apreciará que los principios y conceptos descritos pueden usarse en muchas otras aplicaciones y realizaciones.

Las experiencias de realidad aumentada que permiten presentar información y objetos virtuales para complementar un entorno del mundo real se están volviendo cada vez más populares y se están desarrollando servicios para satisfacer tal demanda.

En muchos enfoques, la aplicación de AR puede proporcionarse localmente a un espectador mediante, por ejemplo, un dispositivo independiente que no usa, o incluso tiene acceso a, ningún servidor remoto de AR. Sin embargo, en otras aplicaciones, una aplicación de AR puede basarse en datos recibidos de un servidor remoto o central. Por ejemplo, los datos de audio o gráficos pueden proporcionarse al dispositivo de AR desde un servidor central remoto y pueden procesarse localmente para generar una experiencia de AR deseada.

La Figura 1 ilustra un ejemplo de un sistema AR en el que un dispositivo cliente AR remoto 101 se comunica con un servidor AR 103, por ejemplo, a través de una red 105, tal como la Internet. El servidor 103 puede disponerse para admitir simultáneamente un número potencialmente grande de dispositivos cliente 101.

El servidor AR 103 puede, por ejemplo, admitir una experiencia aumentada mediante la transmisión de datos que definen elementos de un entorno virtual y objetos al dispositivo cliente 101. Los datos pueden describir específicamente características visuales y propiedades geométricas de varios objetos virtuales que pueden usarse por el dispositivo cliente 101 para generar gráficos de superposición que pueden presentarse a un usuario. En algunas realizaciones, los datos también pueden incluir diversa información que puede presentarse al usuario. Además, el servidor 103 puede proporcionar datos de audio al dispositivo cliente 103 que pueden usarse para generar localmente sonidos/audios virtuales que pueden mejorar aún más la experiencia del usuario y específicamente la inmersión.

La Figura 2 ilustra un dispositivo de acuerdo con algunas realizaciones de la invención. El dispositivo puede ser específicamente un dispositivo cliente 101 de la Figura 1 y se describirá con referencia a tal realización.

El aparato comprende un primer receptor 201 que se dispone para recibir datos de imagen de uno o más sensores de imagen. Los datos de imagen de un sensor de imagen comprenden específicamente una secuencia de imágenes/ fotogramas en tiempo real de una escena del mundo real.

En muchas realizaciones, los cuadros pueden comprender datos de imagen visual de una cámara de imagen visual. Los cuadros/datos de imagen pueden comprender una secuencia en tiempo real de imágenes visuales donde cada píxel representa una intensidad de luz recibida desde una dirección de visión del píxel. Cada píxel puede incluir, por ejemplo, un conjunto de valores de brillo/intensidad de luz para un intervalo (posiblemente ponderado) del espectro visible. Por ejemplo, los valores de píxeles de las imágenes visuales pueden representar uno o más niveles de brillo, tales como, por ejemplo, un conjunto de valores de brillo/intensidad de canales de color. Por ejemplo, las imágenes pueden ser imágenes RGB. En muchas realizaciones, una imagen visual puede ser una imagen en color y/o una imagen RGB (y las referencias a la imagen visual pueden en algunas realizaciones ser reemplazadas por referencias a la imagen en color o imagen RGB).

Alternativamente o adicionalmente, el sensor de imagen puede ser un sensor de profundidad que proporciona una secuencia en tiempo real de imágenes de profundidad/tramas. Para tales imágenes de profundidad, cada valor de píxel puede representar una profundidad/ distancia a un objeto en una dirección de vista del píxel. Por ejemplo, cada valor de píxel puede ser un valor de disparidad o profundidad. Una imagen de profundidad también puede denominarse mapa de profundidad.

Por lo tanto, en diferentes realizaciones, las tramas/ imágenes recibidas del sensor de imagen pueden ser, por ejemplo, imágenes visuales/color, imágenes infrarrojas, imágenes de profundidad, imágenes de radar, imágenes multiespectrales, imágenes de sonar, imágenes de fase, imágenes de diferencia de fase, imágenes de intensidad, imágenes de magnitud de coherencia y/o imágenes de confianza. La secuencia de imágenes/tramas puede ser por lo tanto una estructura bidimensional de píxeles que comprenden valores que representan una propiedad del mundo real en una dirección de vista para el píxel.

En muchas realizaciones, el aparato se dispondrá para procesar imágenes tanto visuales como de profundidad y el primer receptor 201 puede disponerse para recibir tanto una secuencia de imágenes visuales en tiempo real de la escena del mundo real como una secuencia de imágenes de profundidad en tiempo real de la escena del mundo real. En muchas realizaciones, las imágenes pueden ser imágenes de profundidad visual combinadas y el primer receptor 201 puede disponerse para recibir una secuencia de imágenes de profundidad y visuales en tiempo real de la escena del mundo real.

Se apreciará que se conocen muchos sensores y enfoques diferentes para generar imágenes visuales y/o de profundidad y que puede usarse cualquier enfoque y sensores de imagen adecuados. Por ejemplo, puede usarse una cámara de vídeo convencional como un sensor de imagen para generar imágenes visuales. Las imágenes de profundidad pueden generarse, por ejemplo, mediante el uso de una cámara de profundidad dedicada tal como una cámara de alcance infrarroja, o pueden generarse, por ejemplo, mediante la estimación de disparidad en base a, por ejemplo, dos cámaras visuales con un desplazamiento físico conocido.

El(los) sensor(es) de imagen está(n) capturando una escena del mundo real y, por lo tanto, las imágenes recibidas comprenden una captura de esta escena del mundo real. Por ejemplo, para muchas aplicaciones de AR, las imágenes recibidas pueden ser de una escena del mundo real correspondiente al entorno del usuario. Como ejemplo específico, un usuario puede usar un auricular o gafas de AR que también comprenden uno o más sensores de imagen para capturar la escena del mundo real en la dirección en la que el espectador está mirando.

El aparato comprende además un generador de imágenes 203 que se dispone para generar y emitir una secuencia de imágenes visuales que comprende un objeto de imagen que representa un objeto de escena virtual. La secuencia de imágenes visuales de salida puede presentarse a un usuario mediante el uso de una pantalla adecuada, tal como una pantalla (o pantallas) de un casco AR o gafas AR. Por lo tanto, el usuario cuando se le presenta la secuencia de imágenes visuales de salida percibirá que el objeto de imagen corresponde al objeto virtual que está presente en la escena del mundo real.

El generador de imágenes 203 puede adaptar el objeto de imagen (por ejemplo, la orientación en la imagen, la posición en la imagen, la dirección de visión, etc.) en respuesta a cambios en la postura del usuario (posición y/u orientación) como, por ejemplo, proporcionado por datos recibidos de sensores de movimiento adecuados. Por lo tanto, la presentación del objeto de imagen puede ser de manera que proporcione la impresión de que un objeto virtual está presente en la escena del mundo real que se está viendo.

En algunas realizaciones, las secuencias de imágenes visuales de salida generadas pueden incluir solo el objeto virtual (o los objetos virtuales en caso de que se incluya más de uno) y pueden, por ejemplo, presentarse mediante el uso de gafas transparentes que permiten ver la escena del mundo real a través de las gafas. Tal enfoque se conoce como aplicación transparente de AR.

En otras realizaciones, la secuencia de imágenes visuales de salida puede generarse para incluir también una presentación de la escena del mundo real capturada por los sensores de imagen. Por lo tanto, en tal enfoque, la escena del mundo real también puede (y solo) verse a través de las imágenes generadas. Tal enfoque se conoce como aplicación de paso a través de AR.

Se apreciará que se conocen muchos enfoques diferentes para generar tales imágenes AR y para adaptar y modificar objetos de imagen generados para proporcionar una impresión de que un objeto virtual está presente en una escena del mundo real, y por brevedad tales características no se describirán con más detalle. Se apreciará que puede usarse cualquier enfoque adecuado sin restar valor a la invención. Por ejemplo, pueden usarse algoritmos y enfoques de kits de herramientas AR desarrollados, tales como, por ejemplo, procesos de ARKit o ARCore.

El generador de imágenes 203 puede por lo tanto generar un objeto de imagen de manera que se perciba que un objeto de escena virtual se añade a la escena del mundo real. En muchas realizaciones, el objeto de escena virtual puede ser móvil con respecto a al menos un objeto de escena del mundo real, y típicamente el objeto de escena virtual puede ser móvil con respecto a toda la escena del mundo real. Por ejemplo, el objeto de escena virtual puede moverse en función de una acción del usuario, una propiedad de la escena del mundo real que cambia, un movimiento predeterminado para crear un efecto predeterminado, etc.

El aparato comprende además un detector 205 que se dispone para detectar una interacción entre el objeto de escena virtual y un objeto de escena del mundo real de la escena del mundo real. El detector 205 puede detectar específicamente que ocurre una colisión o contacto entre el objeto de escena virtual y un objeto de escena del mundo real. Otras posibles interacciones pueden ser, por ejemplo, una detección de la proximidad del objeto de escena virtual y el objeto de escena del mundo real, o una duración de contacto que puede ser un intervalo de tiempo(s) en el que estamos seguros de que el objeto virtual y el objeto del mundo real se han tocado entre sí. Otro ejemplo de una interacción es cuando llueve en el mundo real y las gotas de lluvia caen sobre el objeto virtual. Otro ejemplo es cuando se genera un flujo de aire en el mundo real y el aire fluye hacia el objeto virtual.

Por lo tanto, el detector 205 puede detectar la interacción entre el objeto de escena virtual y el objeto de escena del mundo real de la escena del mundo real en respuesta a una detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real. La detección de la proximidad puede ser de acuerdo con cualquier criterio de proximidad adecuado.

La detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real puede ser una detección de una distancia entre una posición del objeto de escena del mundo real y una posición del objeto de escena virtual que cumple un criterio de proximidad. La detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real puede ser una detección de una distancia entre una posición del objeto de escena del mundo real y una posición del objeto de escena virtual que es menor que un umbral. Las posiciones y distancias pueden determinarse en un sistema de coordenadas de escena para la escena del mundo real. El objeto de escena virtual puede ser un objeto que se presenta/visualiza como si fuera un objeto de escena del mundo real presente en la escena del mundo real.

En algunas realizaciones, el detector 205 puede determinar una (primera) posición del objeto de escena del mundo real en un sistema de coordenadas (escena) (para la escena) y una (segunda) posición del objeto de escena virtual en el sistema de coordenadas. El detector 205 puede detectar una interacción en respuesta a, y específicamente si, las posiciones (primera y segunda) cumplen un criterio de proximidad. Específicamente, el detector 205 puede detectar una interacción en respuesta a, y específicamente si, una distancia (de acuerdo con cualquier medida de distancia adecuada) entre las posiciones (primera y segunda) cumple con un criterio de proximidad de distancia, y específicamente si la distancia es menor que un umbral.

El generador de imágenes puede disponerse para generar el objeto de imagen para representar un objeto de escena virtual como un objeto que tiene una postura/presencia espacial dada en la escena del mundo real/sistema de coordenadas de la escena. La proximidad puede determinarse en respuesta a una posición indicativa de esta postura/presencia espacial en la escena del mundo real/sistema de coordenadas de la escena y una posición del objeto de la escena del mundo real en la escena del mundo real/sistema de coordenadas de la escena. ¡La proximidad puede determinarse en respuesta a una posición de este objeto de escena virtual! proceso espacial en la escena del

mundo real/ el sistema de coordenadas de la escena y una posición del objeto de escena del mundo real en la escena del mundo real/ el sistema de coordenadas de la escena.

Pueden usarse diferentes enfoques para detectar interacciones en diferentes realizaciones en dependencia de la operación y el rendimiento deseados para la realización específica y se describirán más adelante varios enfoques.

El detector 205 está acoplado a un estimador 207 que se dispone para determinar una propiedad del material para el objeto de escena con el que el objeto de escena virtual ha sido detectado para interactuar. La propiedad del material puede ser una indicación del material/ materia/ cosa de la que está hecho el objeto en tiempo real (que incluye la composición del material en caso de que el objeto comprenda diferentes materiales). La propiedad del material puede generarse, por ejemplo, para indicar si el objeto del mundo real está hecho de una de una pluralidad de categorías, tales como, por ejemplo, si está hecho de madera, tela, metal, plástico, etc.

El estimador 207 está dispuesto para determinar la propiedad del material en respuesta a los datos de imagen recibidos de los sensores de imagen, es decir, se determina en respuesta a los datos de la secuencia de cuadros de imagen recibidos de los sensores. En muchas realizaciones, la secuencia de tramas de imagen puede comprender tanto datos de profundidad como de imagen visual y el estimador 207 puede disponerse para estimar la propiedad del material en respuesta tanto a los datos de profundidad como a los de imagen visual. En otras realizaciones, el estimador 207 puede considerar solo datos visuales o datos de profundidad.

Se apreciará que pueden usarse diferentes algoritmos y enfoques para estimar la propiedad del material en diferentes realizaciones. También se apreciará que, en muchas realizaciones, incluso una estimación muy inexacta e inestable de las propiedades del material puede ser útil y puede proporcionar una experiencia de usuario mejorada.

Como ejemplo de baja complejidad, el estimador 207 puede disponerse, por ejemplo, para determinar la propiedad del material mediante la aplicación de una segmentación basada en el color a las imágenes visuales recibidas y para determinar un objeto de imagen que se considera que coincide con el objeto de imagen en tiempo real. Después, puede comparar el color promedio con un conjunto de categorías predefinidas para encontrar una categoría con una coincidencia más cercana. La propiedad del material del objeto de imagen en tiempo real puede establecerse entonces en una propiedad del material asociada con la categoría más cercana. Como ejemplo simplista, si se determina que el objeto de imagen del mundo real es predominantemente marrón, se puede estimar que está hecho de madera, si es predominantemente plateado se puede estimar que es metal, si es predominantemente de un color primario brillante, se puede estimar que es plástico, etc.

Se apreciará que, en la mayoría de las realizaciones, puede usarse un enfoque más complejo y más preciso para estimar la propiedad del material, y se describirán más ejemplos más adelante.

El aparato comprende además un segundo receptor 209, que se dispone para recibir una serie de objetos de audio, así como también metadatos para los objetos de audio. Los objetos de audio pueden ser un sonido/ clips de audio/ fragmentos de audio y pueden ser específicamente una señal de audio de duración limitada. Se apreciará que cualquier representación y formato puede usarse por los objetos de audio para representar el sonido. Por lo tanto, los objetos de audio pueden corresponder a sonidos que pueden seleccionarse mediante el aparato y renderizarse para proporcionar una salida de sonido cuando ocurren varias acciones.

Específicamente, el aparato comprende un selector 211 que se dispone para seleccionar un objeto de audio de entre los objetos de audio recibidos cuando se detecta una interacción entre el objeto de escena virtual y un objeto de audio del mundo real. El objeto de audio seleccionado se alimenta a un circuito de salida 213 que se dispone para generar una señal de audio de salida que comprende el primer objeto de audio. Por lo tanto, el circuito de salida 213 puede renderizar el objeto de audio seleccionado generando una señal de audio de salida que incluye una renderización del objeto de audio.

Se apreciará que el circuito de salida 213 puede, en dependencia del formato del objeto de audio, en algunos casos simplemente generar la señal de salida como el audio del objeto de audio, puede en algunas realizaciones incluir una decodificación del objeto de audio, puede generar la señal de audio de salida para incluir una mezcla entre varios componentes de audio (por ejemplo, sonido ambiental, audio del narrador, etc.) para combinarse con el objeto de audio, etc. También se apreciará que el circuito de salida 213 puede incluir, por ejemplo, conversión de digital a analógico y amplificación de señal analógica en realizaciones en las que el circuito de salida 213 puede generar una señal de audio de salida analógica, por ejemplo, para accionar directamente altavoces.

El circuito de salida 213 puede incluir el procesamiento de renderización del objeto de audio mediante, por ejemplo, el procesamiento binaural con Respuestas de Impulso Relacionadas con la Cabeza (HRIR), Funciones de Transferencia Relacionadas con la Cabeza (HRTF) o Respuestas de Impulso de Habitación Binaural (BRIR), o la renderización a una configuración de altavoces, por ejemplo, Panorámica de Amplitud Basada en Vectores (VBAP), y/o una simulación acústica adicional de, por ejemplo: oclusión, difracciones, reflexiones, reverberación, extensión de la fuente, etc. Específicamente, el procesamiento de renderizado puede configurarse de manera que el sonido del objeto de audio se perciba como que se origina desde la ubicación de la interacción detectada.

En algunas realizaciones, el circuito de salida 213 puede incluir una adaptación de temporización para controlar una temporización de la renderización del objeto de audio seleccionado, tal como específicamente un retardo posiblemente variable. Por ejemplo, puede incluirse un retardo entre un tiempo del contacto/interacción "visual" y un tiempo de renderización del sonido. El retardo puede ajustarse, por ejemplo, en base a la velocidad del sonido y la distancia entre el objeto virtual y el observador, por ejemplo, para garantizar que el sonido generado se perciba simultáneamente a la interacción visual. En algunas realizaciones, tal ajuste de tiempo puede realizarse en otras partes del aparato que en el circuito de salida 213.

Los metadatos recibidos por el segundo receptor 209 incluyen metadatos que son indicativos de enlaces entre al menos algunos de los objetos de audio y las características del material. Los metadatos pueden indicar, por ejemplo, para cada objeto de audio, un material de un objeto del mundo real con el que está vinculado. Por ejemplo, los metadatos pueden indicar que un objeto de audio está vinculado a un objeto del mundo real que está hecho de madera, que otro objeto de audio está vinculado a un objeto del mundo real que está hecho de metal, que otro objeto de audio está vinculado a un objeto del mundo real que está hecho de plástico, etc.

Como se describirá con más detalle más adelante, en muchas realizaciones, los metadatos pueden incluir otros enlaces adicionales, tales como, por ejemplo, un enlace de objetos de audio a objetos de escena virtual, a características de material de objetos de escena virtual, tipos de interacción, etc.

El selector 211 responde al detector 205 que detecta una interacción entre el objeto de escena virtual y un objeto de escena del mundo real dispuesto para seleccionar un objeto de audio en base a la propiedad del material estimada para el objeto de escena del mundo real y los metadatos que indican un enlace entre objetos de audio y características del material

En una realización de baja complejidad, el selector 211 puede simplemente seleccionar el objeto de audio para el cual la propiedad material estimada y la característica material vinculada coinciden, por ejemplo, si la propiedad estimada es "madera", el objeto de audio vinculado a "madera" se selecciona y se genera una señal de audio de salida correspondiente.

El aparato puede generar una experiencia de usuario mejorada en muchas realizaciones y puede proporcionar en particular una experiencia más inmersiva y de sonido natural donde el audio percibido puede reflejar más de cerca las interacciones del objeto de escena virtual con objetos del mundo real. Por ejemplo, en lugar de acompañar meramente las interacciones, tales como colisiones, entre objetos de escena virtual y objetos del mundo real con audio estandarizado, el aparato puede proporcionar una salida de audio que se ajuste más a la percepción visual del usuario y puede proporcionar una correspondencia más cercana entre las entradas visuales y de audio percibidas por el usuario. Esto puede lograrse mediante el aparato que adapta y genera audio que coincide más estrechamente con las interacciones.

El sistema puede, por ejemplo, usar objetos/fragmentos de audio múltiples grabados o sintetizados proporcionados para un objeto de escena virtual. Cada fragmento de audio puede modelar la interacción con una clase específica de objetos del mundo real que podrían ocurrir en la escena. En el momento de la ejecución, el aparato puede usar una cámara integrada para clasificar, por ejemplo, la superficie sobre la que se coloca el objeto virtual, y en base al resultado de la clasificación, el aparato puede reproducir el fragmento de audio correcto para la interacción específica.

En algunas realizaciones, el segundo receptor 209 puede disponerse para recibir los objetos de audio y/o los metadatos de una fuente interna. Sin embargo, en muchas realizaciones, el segundo receptor 209 puede disponerse para recibir los metadatos, y a menudo los objetos de audio, de un servidor remoto, tal como específicamente el servidor 103 de la Figura 1.

Por ejemplo, un servidor remoto puede mantener una gran biblioteca de objetos de audio correspondientes a una variedad de posibles interacciones entre objetos, y puede mantener además metadatos que incluyen, por ejemplo, datos que definen la propiedad del material de los objetos involucrados en la interacción, propiedades de la interacción, etc.

En la inicialización de la aplicación, o a una tasa de repetición, el segundo receptor 209 puede recuperar los metadatos para todos los objetos de audio. Cuando el detector 205 detecta una interacción, el estimador 207 puede estimar la propiedad del material del objeto de la escena del mundo real y el selector 211 puede evaluar los metadatos para encontrar un objeto de audio coincidente. Después, puede proceder a controlar el segundo receptor 209 para recuperar el objeto de audio seleccionado del servidor 103 y, cuando se reciba, el aparato de audio puede generar la señal de audio para incluir el objeto de audio recibido.

En algunas realizaciones, tal enfoque puede ser demasiado lento para el funcionamiento en tiempo real y el segundo receptor 209 puede en algunas realizaciones recuperar un subconjunto o posiblemente todos los objetos de audio del servidor remoto 103 y almacenarlos localmente para un rápido acceso. Por ejemplo, en algunas realizaciones, se puede realizar un análisis de la escena del mundo real al iniciar, y/o a intervalos regulares, con el aparato que detecta tantos objetos como sea posible. El aparato puede proceder a recuperar todos los objetos de audio para objetos o

materiales correspondientes a los detectados en la escena del mundo real, y/o correspondientes a los objetos de la escena virtual activos o ubicados cerca de la posición del usuario. En otras realizaciones, un usuario puede, por ejemplo, en la inicialización proporcionar una entrada de la escena del mundo real (por ejemplo, sala de estar, arena deportiva, etc.) y el aparato puede proceder a recuperar un conjunto de objetos de audio que se almacenan en el aparato como posibles objetos en tal entorno.

El enfoque puede proporcionar por lo tanto un enfoque eficiente para permitir que un servidor central interactúe con aparatos de dispositivos remotos para proporcionar soporte de audio para aplicaciones AR. Se aprecia que el servidor central como se mencionó anteriormente también podría representarse por un servidor relativamente local, tal como un servidor de borde 5G.

En el enfoque, el objeto de escena virtual se presenta al usuario de manera que puede percibirse que está presente en la escena del mundo real, y por lo tanto se asocia con una geometría/forma espacial (típicamente tridimensional) en la escena del mundo real. El detector 205 puede disponerse para detectar la interacción como una detección de que la distancia entre el objeto de escena virtual y el objeto de escena del mundo real en el espacio tridimensional del mundo real es menor que un umbral dado, o específicamente que entran en contacto entre sí. En algunos casos, pueden tenerse en cuenta consideraciones adicionales, tales como la velocidad a la que los objetos se mueven entre sí, etc.

Por ejemplo, algunos conjuntos de herramientas desarrollados para admitir aplicaciones de AR proporcionan detección de superficies planas en 3D donde, a través del seguimiento de puntos de características a lo largo del tiempo, las regiones planas se detectan a través del movimiento específico que sigue un grupo de puntos de características a lo largo del tiempo.

En algunas realizaciones, el detector 205 puede disponerse para detectar una interacción al rastrear la distancia más corta entre cualquier vértice en una malla delimitadora del objeto de escena gráfica virtual y uno cualquiera de los planos detectados:

$$d(t) \equiv \min_{i \in V} \left(\min_{j \in P} \left(\text{dist}\left(\mathbf{x}_j(t), \mathbf{x}_i(t) \right) \right) \right)$$

donde i E V denota el vértice i del conjunto de todos los vértices V presentes en la malla delimitadora del objeto de escena virtual, j E P es el plano para el cual se calcula la distancia más corta y la función dist ($x_j(t)$, $x_i(t)$) evalúa esta distancia. Se debe señalar que tanto la malla de contorno del objeto de escena virtual como los sensores de imagen están generalmente en movimiento, lo que significa que las posiciones de los vértices en la malla y las posiciones en los planos detectados, y por lo tanto también la distancia más corta final $d(t)$, todos varían con el tiempo.

Puede detectarse que se produjo una interacción si la distancia es menor que un umbral. Específicamente, se puede detectar que se produjo una colisión si la distancia es menor que un umbral pequeño, o por ejemplo, que alcanza cero.

En algunas realizaciones, y una interacción, y específicamente una colisión, puede detectarse que ha ocurrido cuando la distancia mínima de un vértice a un plano ha disminuido a una distancia umbral mínima $\Delta_{\text{colisión}}$ y la velocidad instantánea del objeto $v(t)$ excede un umbral mínimo dado v_{min}:

$$(d(t) \leq \Delta_{\text{colisión}}) \wedge (v(t) > v_{\text{min}})$$

El razonamiento detrás de la condición de velocidad es que tiende a ser difícil en la práctica determinar si ocurrió una colisión real cuando la velocidad del objeto es baja, ya que el objeto virtual puede detenerse justo delante de o al lado de una superficie detectada. Los valores de parámetros útiles son, por ejemplo, $\Delta_{\text{colisión}}$= 5 mm y v_{min} = 0,05 m/s.

El estimador 207 puede proceder a determinar la propiedad del material para el objeto de escena del mundo real con el que se ha detectado que se produce la interacción. En muchas realizaciones, la detección puede basarse en que el estimador 207 determine una región de imagen de interacción en al menos uno de los cuadros de imagen de entrada, donde la región de imagen de interacción es una región de imagen del cuadro de imagen en la que ocurre la interacción. Por lo tanto, el estimador 207 puede determinar una región bidimensional en la trama de imagen que incluye el punto de contacto entre los objetos detectados por el detector 205. Por ejemplo, la proyección del punto de contacto en el marco de imagen bidimensional puede determinarse y puede identificarse una región alrededor de este.

En algunas realizaciones, la región de imagen puede ser, por ejemplo, una forma predeterminada que se centra alrededor del punto de contacto proyectado. En otras realizaciones, la región de imagen puede adaptarse a las propiedades de la imagen. Por ejemplo, la región de la imagen puede identificarse como una región alrededor del punto de contacto proyectado que cumple un criterio similar, tal como que las propiedades visuales y/o la profundidad no se desvían de las propiedades del punto de contacto proyectado en más de una cantidad predeterminada.

El estimador 207 puede entonces determinar la propiedad del material para el objeto de escena en respuesta a los datos de imagen de la región de imagen de interacción. Por ejemplo, las variaciones de color y textura pueden compararse con las propiedades correspondientes de un conjunto predeterminado de clases y la propiedad del material puede establecerse en la propiedad de la clase coincidente más cercana.

En muchas realizaciones, la propiedad del material puede determinarse en respuesta a una comparación de una propiedad de imagen determinada para los datos de imagen con una pluralidad de referencias de propiedades de imagen, cada referencia de propiedad de imagen está vinculada con un valor de propiedad del material. La propiedad de imagen puede determinarse en respuesta a/comunicación con un valor de propiedad de material vinculado con una referencia de propiedad de imagen coincidente de la pluralidad de referencias de propiedad de imagen. Una referencia de propiedad de imagen coincidente puede determinarse como una para la cual la propiedad de imagen y la referencia de propiedad de imagen coincidente cumplen un criterio de coincidencia. La propiedad de imagen puede ser una propiedad de brillo y/o color y/o textura.

Específicamente, una vez que se detecta una colisión, se puede determinar la región de interacción de la imagen. Esto puede hacerse proyectando primero el punto de malla más cercano x_i en la imagen mediante el uso de la matriz de vista de la cámara y la matriz de proyección de la cámara. El punto de imagen 2D resultante (u_i, v_i) se usa después para determinar una subimagen de tamaño fijo, centrada en (u_i, v_i), de la imagen de la cámara. La Figura 3 ilustra un ejemplo de una trama de imagen el que un objeto de escena virtual 301 en forma de una bola puede interactuar con un objeto de escena del mundo real 303 en forma de una mesa del mundo real que da como resultado la detección de una región de imagen 305.

El estimador 207 puede disponerse en muchas realizaciones para determinar una indicación de coincidencia para el objeto de escena del mundo real a al menos una primera categoría/clase de una pluralidad de categorías/clases de objetos. El estimador 207 puede tener, por ejemplo, un número de propiedades almacenadas para regiones de imagen correspondientes a un material específico, tal como color, variaciones de textura, variaciones de profundidad, etc. El estimador 207 puede determinar las propiedades correspondientes para la región de la imagen y comparar estas con las propiedades almacenadas para todas las categorías. Puede determinarse una indicación de coincidencia para reflejar qué tan bien coinciden las propiedades, y la categoría puede seleccionarse como la categoría para la cual la indicación de coincidencia indica una coincidencia más alta.

La propiedad del material para el objeto de escena del mundo real puede seleccionarse entonces como la propiedad almacenada para la categoría o clase seleccionada. Por lo tanto, cada una de las categorías/clases se asocia con una o más propiedades del material, tales como simplemente una indicación del material del que está hecho un objeto, y la propiedad del material del objeto de la escena del mundo real se establece en la almacenada para la categoría seleccionada.

Se apreciará que, en muchas realizaciones, el estimador 207 puede usar enfoques sustancialmente más complejos para determinar la propiedad del material para el objeto de la escena del mundo real, y específicamente para identificar una clase o categoría considerada que coincide más estrechamente con las propiedades de la región de la imagen correspondiente a la región de la imagen.

De hecho, en muchas realizaciones, el estimador 207 puede comprender una red neuronal dispuesta para estimar la propiedad del material del objeto de la escena del mundo real en base a los datos de imagen de la región de imagen de interacción. La red neuronal puede usarse específicamente para identificar una categoría o clase para el objeto de escena del mundo real.

En muchas realizaciones, una región de interacción extraída (imagen) puede tener un tamaño mucho menor que la imagen completa. Cuando la imagen completa tiene a menudo un tamaño de 2K o 4K, el subimagen de interacción puede tener un tamaño más pequeño constante de, por ejemplo, 256x256 píxeles.

Una imagen en color de, por ejemplo, 256x256 píxeles puede representarse como un tensor de dimensiones 3x256x256 y alimentarse a una red neuronal clasificador entrenada. La salida de la red neuronal puede ser un vector con, para cada clase de material predefinida, la probabilidad de que la imagen pertenezca a la clase dada.

Las siguientes clases de objetos (superficies) pueden ser particularmente ventajosas en muchas realizaciones:

- Alfombra de superficie del suelo (variación de altura corta del material)
- Alfombra de superficie del suelo (la variación de altura larga del material suprimirá el sonido de colisión)
- Piedra de superficie de suelo
- Madera de superficie de suelo

- Vidrio de superficie de la mesa
- Madera de la superficie de la mesa
- Pared
- Sofá
- Cojín
- Superficie de plástico de objetos domésticos (la estufa, el teclado, la computadora portátil)

Cada una de estas clases puede tener estadísticas de color y textura específicas o puede diferenciarse debido a su geometría delimitadora (por ejemplo, mesa o cojín).

Como conjunto de datos de entrenamiento, las imágenes de las diferentes superficies de objetos pueden capturarse y anotarse manualmente con una etiqueta de clase de verdad de la tierra.

Para la arquitectura de red neuronal, una opción puede ser usar como bloque básico una capa de convolución 2D que duplica el número de canales, seguido de una capa de agrupación media 2D (paso=2) y una activación lineal rectificada. Cada bloque reduce la resolución espacial del tensor con un factor 2 pero duplica el número de canales. El enfoque puede continuar conectando estos bloques hasta que el tensor tenga una resolución de N-canalesx1x1. Después se pueden añadir dos capas completamente conectadas seguidas de una función de activación sigmoide.

Si el máximo de todas las probabilidades de clase está por debajo de un umbral dado (0,5 parece un valor apropiado en muchas realizaciones), entonces se puede reproducir un sonido de interacción predeterminado (neutral), es decir, se puede seleccionar un objeto de audio predeterminado si el estimador 207 no es capaz de determinar con suficiente precisión el material del objeto de escena del mundo real. En todos los demás casos, el selector 211 puede seleccionar el objeto de audio/archivo de sonido, tal como específicamente el objeto de audio correspondiente al material que tiene la mayor probabilidad según lo determinado por la categorización.

En algunas realizaciones, el aparato puede comprender además un receptor de audio 215 para recibir una señal de audio de audio en tiempo real capturada en la escena del mundo real y el estimador 207 puede disponerse para determinar la indicación de coincidencia en respuesta a la señal de audio.

Por lo tanto, la clasificación de materiales puede ser ayudada por la clasificación basada en audio, que registra sonidos en el espacio físico con un micrófono. Si se producen interacciones entre objetos físicos, el sonido resultante puede proporcionar información valiosa para la clasificación del material o propiedades más específicas. Puede usarse para proporcionar una mejor clasificación del material, o proporcionar atributos adicionales del material (sólido/ahuecado, delgado/grueso, tenso/holgado, grande/pequeño).

Por ejemplo, el sonido ambiental actual y reciente S_{env} puede capturarse durante un corto período de tiempo (digamos 10 segundos) y puede alimentarse al clasificador de materiales. Por ejemplo, un usuario que sostiene un teléfono inteligente o usa un auricular de AR, u otras personas pueden estar caminando en la habitación y el sonido de sus zapatos en el suelo puede proporcionar información sobre el material de la superficie del suelo. Un clasificador de materiales puede producirse mediante la recopilación de imágenes y sonidos de diferentes materiales. Una red neuronal más avanzada como se analizó anteriormente concatenaría entonces piezas de los fragmentos de audio con las imágenes capturadas y usaría esas como entrada para la red entrenada.

Como ejemplo específico, en algunas realizaciones, la determinación de la propiedad del material puede ser en respuesta a los datos de imagen en profundidad, y específicamente el estimador puede estar dispuesto para determinar la propiedad del material para el objeto de escena del mundo real en respuesta a una detección de que al menos parte de una región de imagen de al menos una trama de imagen que representa el objeto de escena del mundo real tiene un nivel de confianza para los datos de imagen en profundidad que no excede un umbral.

Las estimaciones de profundidad a menudo se generan con datos de confianza que son indicativos de la confiabilidad de los valores/estimaciones de profundidad generados. Por ejemplo, para la estimación de la disparidad, el nivel de confianza puede determinarse para reflejar cuán de cerca coinciden las regiones de la imagen que forman la base para la estimación de la disparidad en las imágenes. Como otro ejemplo, para una cámara de profundidad, por ejemplo, basada en reflejos de luz infrarroja, el nivel de confianza puede generarse para reflejar la cantidad de luz infrarroja que se recibe para un píxel dado. Si solo se recibe una pequeña cantidad de luz, la estimación de distancia/proceso de alcance puede no ser tan precisa como si se recibiera una gran cantidad de luz. En algunos casos, cada píxel de una imagen/mapa de profundidad puede comprender tanto un valor de profundidad estimado como un valor/nivel de confianza que indica cuán confiable se considera el valor de profundidad para el píxel.

En algunas realizaciones, estos datos de confianza de profundidad pueden tenerse en cuenta cuando se determina la propiedad del material. Por ejemplo, se puede conocer que algunos materiales proporcionan un nivel de confianza reducido en comparación con otros materiales. Por ejemplo, algunos materiales pueden variar mucho en apariencia visual local, lo que puede hacer que la estimación de la disparidad sea menos confiable.

Como otro ejemplo, para la estimación de la profundidad en base a la luz reflejada, y específicamente la luz infrarroja reflejada, emitida desde el sensor, algunos materiales pueden resultar en una confianza mucho menor debido a la cantidad de luz reflejada de vuelta al sensor que se reduce sustancialmente. Este puede ser el caso, por ejemplo, de materiales que no reflejan, pero tienen una alta absorción de luz infrarroja. También puede ser particularmente el caso de materiales que exhiben reflexión especular (por ejemplo, un objeto de metal) en cuyo caso muy poca luz infrarroja se refleja de vuelta al transmisor y al sensor. Tenga en cuenta que para un sensor de profundidad activo basado en tiempo de vuelo o luz estructurada, el sensor de luz infrarroja (por ejemplo, CMOS) a menudo se ubica junto con el transmisor/emisor de luz. Al mismo tiempo, la luz visual recibida de la superficie metálica puede, debido a la reflexión de la luz ambiental por la superficie, ser bastante sustancial. En muchas realizaciones, el estimador 207 puede detectar, por ejemplo, que un área de la región de la imagen de interacción tiene una confianza de profundidad baja, pero es bastante brillante. Tal área puede ser indicativa de una superficie altamente reflectante pero dispersiva, tal como una superficie metálica.

En algunas realizaciones, la trama de imagen comprende datos de imagen visual y datos de imagen de profundidad, y el estimador 207 puede disponerse para determinar que el objeto de escena del mundo real tiene un componente metálico en respuesta a una detección de que para al menos parte de la región de la imagen un brillo de los datos de imagen visual excede un umbral y un nivel de confianza para los datos de imagen de profundidad no excede un umbral.

Como se describió anteriormente, los metadatos pueden recibirse de una fuente remota y pueden comprender enlaces entre objetos de audio y una propiedad de material para un objeto, que específicamente puede ser el objeto de escena del mundo real. Sin embargo, en algunas realizaciones, los metadatos pueden comprender enlaces a propiedades de material de dos objetos, y de hecho pueden diferenciar entre el objeto de escena virtual y el objeto de escena del mundo real. En tales casos, el selector 211 puede seleccionar el objeto de audio en base a la propiedad material estimada del objeto de escena del mundo real y también en base a una propiedad material del objeto de escena virtual. Como el objeto de escena virtual puede ser un objeto virtual generado por el aparato, el material del que se pretende que esté hecho este objeto se conocerá típicamente. En este caso, el objeto de audio puede seleccionarse de modo que coincida con ambas propiedades lo más cerca como sea posible.

Por lo tanto, en algunas realizaciones, los metadatos para al menos algunos objetos de audio comprenden indicaciones de enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena del mundo real y enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena virtuales. En tales sistemas, el selector 211 puede disponerse para seleccionar el objeto de audio en respuesta a la propiedad del material y las características del material de objetos del mundo real vinculados al conjunto de objetos de audio y en respuesta a una propiedad del material del objeto de la escena virtual y las características del material de objetos de la escena virtual vinculados al conjunto de objetos de audio.

Como ejemplo, el servidor remoto 103 puede proporcionar una estructura bidimensional que vincula objetos de audio a propiedades del material. Un ejemplo de tal estructura es la siguiente tabla de consulta bidimensional:

	Material del mundo real detectado			
Material de objeto virtual	**Madera**	**metal**	**hormigón**	**desconocido**
madera	MaderaEnMadera	MaderaEnMetal	MaderaEnHormigón	MaderaEnDesconocido
metal	MetalEnMadera	MetalEnMetal	MetalEnHormigón	MetalEnDesconocido
caucho	CauchoEnMadera	CauchoEnMetal	CauchoEnHormigón	CauchoEnDesconocido

En otras realizaciones, los metadatos pueden proporcionar enlaces a otras características y la selección del objeto de audio puede tener en cuenta tales características adicionales.

Por ejemplo, en muchas realizaciones, puede considerarse además una propiedad dinámica del objeto de escena virtual. En tales realizaciones, los metadatos pueden indicar que diferentes objetos de audio están vinculados a diferentes estados/propiedades del objeto de escena virtual. Los estados/propiedades del objeto de escena virtual pueden cambiar dinámicamente, y la selección puede tener en cuenta además el estado/propiedad actual al encontrar un objeto de audio coincidente.

Por ejemplo, una bola virtual desinflada genera un sonido diferente a una bola virtual inflada. En algunas realizaciones, el objeto de escena virtual de una pelota puede incluirse dos veces en la tabla de consulta, una vez desinflado y una vez inflado, cada uno vinculado a un objeto de audio/efecto de sonido diferente.

Otros atributos del objeto virtual pueden influir en la selección o generación de un efecto de sonido apropiado. Por ejemplo, la orientación del objeto virtual puede provocar diferentes efectos de sonido.

El objeto de escena virtual puede tener, por ejemplo, diferentes tipos de superficies, tales como, por ejemplo, un cubo que tiene diferentes propiedades de superficie para los diferentes lados. En tal caso, la orientación del objeto de escena virtual puede determinar qué superficie colisiona con el objeto de escena del mundo real y la selección del objeto de audio puede tener en cuenta la orientación.

Como ejemplo específico, la tabla de consulta puede comprender diferentes entradas para diferentes orientaciones y, por lo tanto, la selección del objeto de audio puede incluir la selección de la entrada correspondiente a la orientación actual. Los objetos de escena virtual pueden representarse mediante una malla, donde cada cara, o grupo de caras, se asocia con un intervalo de orientación del objeto, propiedad de material virtual, o subconjunto de objetos de audio.

Finalmente, podrían detectarse y usarse atributos adicionales del objeto de escena del mundo real en la selección o generación de efectos de sonido. Una puerta abierta puede sonar diferente a una puerta cerrada. O un vaso suena diferente cuando contiene líquido.

En algunas realizaciones, la selección del objeto de audio puede ser además en respuesta a una propiedad de la interacción, tal como que depende específicamente del tipo de interacción. Por lo tanto, en algunas realizaciones, el detector 205 puede disponerse para determinar una propiedad de la interacción y el selector 211 puede disponerse para seleccionar el primer objeto de audio en respuesta a la propiedad de la interacción.

Por lo tanto, en algunas realizaciones, los atributos de la interacción misma pueden influir en la selección o generación de un efecto de sonido. La velocidad o la fuerza de una interacción puede provocar diferentes sonidos, ya que las colisiones físicas pueden ser no lineales con la intensidad de la colisión (por ejemplo, las pelotas rebotadas, el vidrio puede romperse a cierta intensidad de colisión, etc.). Otros ejemplos son la dirección de la colisión, o interacciones de no colisión tales como el deslizamiento/frotamiento de un objeto virtual a través de una superficie física.

Como ejemplos específicos, la propiedad de interacción que se considera además para la selección del objeto de audio puede ser una o más de las siguientes:

Una velocidad de la interacción: El detector 205 puede, por ejemplo, determinar la velocidad relativa entre el objeto de escena virtual y el objeto de escena del mundo real y puede seleccionar el objeto de audio en base a esto. Por ejemplo, una pelota lenta que golpea el suelo puede tener un sonido diferente al de una que golpea el suelo a una velocidad mucho mayor. La tabla de consulta puede tener, por ejemplo, diferentes entradas para diferentes intervalos de velocidad, con cada entrada vinculada a objetos de audio que representan grabaciones de pelotas que golpean el material de piso adecuado con diferentes velocidades. Como ejemplo, la detección de la interacción puede incluir un vector de velocidad que se compara con la orientación de la superficie física y, opcionalmente, su comportamiento poco después de que la interacción permita la detección del tipo de interacción. Por ejemplo, si el vector de velocidad del objeto es perpendicular y hacia la superficie física, la interacción puede considerarse un tipo de interacción de 'impacto' o 'unión'. Un vector de velocidad junto a la superficie del objeto daría como resultado un tipo de interacción de 'deslizamiento'. Tales tipos de interacción diferentes pueden vincularse a diferentes objetos de audio.

Una fuerza de una colisión entre el objeto de escena virtual y el objeto de escena del mundo real. La fuerza puede determinarse, por ejemplo, como una función de una velocidad actual y un peso estimado o supuesto de al menos uno del objeto de escena virtual y el objeto de escena del mundo real. Diferentes impactos de fuerza pueden conducir a diferentes sonidos y se pueden proporcionar y seleccionar diferentes objetos de audio en base a la fuerza.

Una elasticidad de una colisión entre el objeto de la escena virtual y el objeto de la escena del mundo real. Por ejemplo, cuando se golpea un piso, una pelota inflada a alta presión puede sonar diferente a una pelota que está parcialmente desinflada. La elasticidad de las bolas puede ser diferente y dar como resultado sonidos muy diferentes. En algunas realizaciones, en particular, pueden proporcionarse diferentes objetos de audio para diferentes elasticidades del objeto de escena virtual, y los metadatos pueden incluir datos que permiten diferenciar los objetos de audio en base a tal elasticidad.

Una duración de la interacción: Por ejemplo, el aparato puede incluir, o comunicarse con, un motor de física que predice una trayectoria actualizada del objeto de escena virtual después de la interacción detectada con el objeto del mundo real. La trayectoria pronosticada puede producir una duración de la interacción debido al momento, la inercia, la dirección o la elasticidad de la interacción o los materiales virtuales y del mundo real involucrados en la interacción. Una larga duración de la interacción puede resultar en un objeto de audio con una duración larga o nominal, mientras que una duración más corta puede resultar en un objeto de audio con una duración más corta, o una representación más corta del objeto de audio nominal.

En otros ejemplos, la interacción puede ser impulsada por el usuario que interactúa con el contenido virtual, tal como arrastrar el objeto de escena virtual a través del entorno del mundo real. La interacción puede durar tanto como el usuario tenga el objeto de escena virtual en proximidad, o en contacto (virtual) con el objeto del mundo real.

Una dirección de movimiento del objeto de escena virtual con relación al objeto de escena del mundo real. Por ejemplo, deslizar un objeto virtual mientras se toca a lo largo de una superficie de madera del mundo real generalmente producirá un sonido más agudo que cuando se golpea la superficie de madera del mundo real localmente en la dirección paralela a la normal de la superficie.

En muchas realizaciones, los metadatos pueden proporcionarse en forma de una tabla de consulta que puede ser multidimensional en dependencia de cuántas propiedades pueden/deben tenerse en cuenta al seleccionar el objeto de audio. Por ejemplo, puede proporcionarse una tabla de consulta con las siguientes dimensiones:

- Dimensión 1: Material virtual
- Dimensión 2: Material del mundo real detectado
- Dimensión 3: Tipo de interacción (por ejemplo, impacto/deslizamiento/rebote/unir/separar)
- Dimensión 4: Velocidad de interacción

En algunas realizaciones, se puede proporcionar una tabla de metadatos que básicamente vincula dos tipos de materiales a cada efecto de sonido. Sin embargo, puede haber muchos efectos de sonido para los mismos materiales e incluso tipo de interacción (material 1, material 2 y tipo de interacción). En base a esta información, todos ellos pueden ser adecuados para representar una determinada interacción. Un enfoque es elegir siempre el primero, pero la experiencia puede ser más convincente si se elige un efecto de sonido al azar. Para interacciones repetitivas, pequeñas diferencias en el audio hacen que suene menos artificial. La elección entre los clips de sonido incluso puede variar en base a los atributos de interacción que no se cubren en los metadatos de la biblioteca (por ejemplo, velocidad, o dirección del impacto, o el rebote n-ésimo de una pelota).

Un enfoque específico del aparato de la Figura 2 se representa mediante el diagrama de flujo de la Figura 4. En el ejemplo, una imagen de video actual *I* se analiza para predecir una región de la imagen el que es probable que ocurra la interacción. Este análisis puede basarse en la información de posición 3D del objeto de escena virtual/objeto gráfico *G.* La parte de imagen de la región de imagen espacial resultante *R* en el que el objeto de escena virtual es probable que colisione con un objeto del mundo real del entorno real se alimenta entonces a un clasificador de materiales (el estimador 207) que determina una clase de material *M.* Finalmente, en dependencia de la clase de material *M* y el objeto de escena virtual/objeto gráfico *G* se produce un sonido.

La clasificación del material de superficie en base a la información visual solamente es probable que sea suficiente para la aplicación en muchas implementaciones y usos prácticos. Además, si no se identifica una categoría o sonido coincidente, se reproduce un sonido de interacción predeterminado para el objeto de escena virtual *G* puede usarse en su lugar uno específico para el material del objeto del mundo real. Opcionalmente, el sonido ambiental actual y del pasado reciente S_{env} puede alimentarse al clasificador de materiales.

El enfoque puede, por ejemplo, permitir una aplicación donde un sonido preregistrado comienza a reproducirse tan pronto como se produzca una colisión entre un objeto de escena virtual y un objeto de escena del mundo real. En lugar de especificar para cada objeto virtual, los sonidos de interacción con todos los materiales del mundo real posibles, cada objeto virtual puede tener una propiedad de material y puede usarse una sola tabla de consulta como base para seleccionar el sonido apropiado. Una clase 'desconocida' puede usarse para producir aún un sonido aproximadamente correcto cuando el clasificador de materiales no produce probabilidades lo suficientemente altas para cualquiera de las clases predefinidas. Por ejemplo, una pelota de goma virtual hará más o menos el mismo sonido cuando golpee diferentes superficies del mundo real.

Un estándar de AR puede proporcionar sintaxis para un flujo de bits con audio, visual y metadatos para elementos virtuales para aumentar áreas físicas específicas. El estándar puede permitir al usuario interactuar con estos elementos virtuales en el entorno físico. Para tener una amplificación de audio de estas interacciones, el estándar puede proporcionar medios para transmitir muchos clips de efectos de sonido al decodificador sin que se rendericen al usuario a menos que se activen.

Un flujo de bits proporcionado puede incluir un metadato indicativo de una tabla de consulta multidimensional que asigna cada clip de efecto de sonido a una entrada en esa tabla de consulta. Cada dimensión de la tabla puede corresponder con un aspecto de una interacción entre un elemento virtual y un elemento físico. Cuando una interacción tiene aspectos (por ejemplo, detectados por clasificación y/o algoritmos lógicos) presentes en esta tabla de consulta

que identifican conjuntamente una determinada entrada de tabla, un clip de efecto de sonido asociado con esta entrada de tabla puede reproducirse en la posición de la interacción entre el elemento físico y virtual.

Se apreciará que la descripción anterior para mayor claridad ha descrito realizaciones de la invención con referencia a diferentes circuitos, unidades y procesadores funcionales. Sin embargo, será evidente que se puede utilizar cualquier distribución adecuada de funcionalidad entre diferentes circuitos funcionales, unidades o procesadores sin detrimento de la invención. Por ejemplo, la funcionalidad ilustrada para ser realizada por procesadores o controladores separados puede ser realizada por el mismo procesador o controladores. Por lo tanto, las referencias a unidades funcionales o circuitos específicos deben verse solo como referencias a medios adecuados para proporcionar la funcionalidad descrita en lugar de ser indicativas de una estructura u organización lógica o física estricta.

La invención puede implementarse en cualquier forma adecuada que incluye hardware, software, microprograma o cualquier combinación de estos. La invención puede implementarse opcionalmente al menos parcialmente como un software informático que se ejecuta en uno o más procesadores de datos y/o procesadores de señales digitales. Los elementos y componentes de una realización de la invención pueden ser implementados física, funcional y lógicamente de cualquier manera adecuada. De hecho, la funcionalidad puede implementarse en una sola unidad, en una pluralidad de unidades o como parte de otras unidades funcionales. Como tal, la invención puede ser implementada en una sola unidad o puede estar distribuida física y funcionalmente entre diferentes unidades, circuitos y procesadores.

Aunque la presente invención se ha descrito en relación con algunas realizaciones, no se pretende que se limite a la forma específica establecida en la presente memoria. Más bien, el ámbito de la presente invención está limitado únicamente por las reivindicaciones adjuntas. Además, aunque una característica puede parecer ser descrita en conexión con realizaciones particulares, un experto en la técnica reconocerá que varias características de las realizaciones descritas pueden combinarse de acuerdo con la invención. En las reivindicaciones, el término que comprende no excluye la presencia de otros elementos o pasos.

Además, aunque se enumeran individualmente, una pluralidad de medios, elementos, circuitos o pasos del procedimiento puede implementarse, por ejemplo, un solo circuito, unidad o procesador. Adicionalmente, aunque las características individuales pueden incluirse en diferentes reivindicaciones, estas pueden combinarse ventajosamente, y la inclusión en diferentes reivindicaciones no implica que una combinación de características no sea factible y/o ventajosa. Además, la inclusión de una característica en una categoría de reivindicaciones no implica una limitación a esta categoría, sino que indica que la característica es igualmente aplicable a otras categorías de reivindicaciones según corresponda. Además, el orden de las características en las reivindicaciones no implica ningún orden específico en el que deban trabajarse las características y, en particular, el orden de los pasos individuales en una reivindicación del procedimiento no implica que los pasos deban realizarse en este orden. Más bien, las etapas pueden realizarse en cualquier orden adecuado. Además, las referencias singulares no excluyen una pluralidad. Por lo tanto, las referencias a "un", "una", "primero", "segundo", etc. no excluyen una pluralidad. Los signos de referencia en las reivindicaciones se proporcionan meramente como un ejemplo aclaratorio y no se interpretarán como limitantes del ámbito de las reivindicaciones de ninguna manera.

REIVINDICACIONES

1. Un aparato para generar una señal de audio de salida, el aparato que comprende:

un primer receptor (201) dispuesto para recibir una secuencia de imágenes en tiempo real de una escena del mundo real de un sensor de imagen, la secuencia de imágenes en tiempo real que comprende una secuencia de tramas de imágenes, cada trama de imagen que comprende al menos uno de los datos de imagen visual y los datos de imagen de profundidad;
un segundo receptor (209) dispuesto para recibir un conjunto de objetos de audio y metadatos para objetos de audio del conjunto de objetos de audio, los metadatos son indicativos de enlaces entre objetos de audio del conjunto de objetos de audio y características del material;
un generador de imágenes (203) dispuesto para generar una secuencia de imágenes de salida que comprende un objeto de imagen correspondiente a un objeto de escena virtual en la escena del mundo real;
un detector (205) dispuesto para detectar una interacción entre el objeto de escena virtual y un objeto de escena del mundo real de la escena del mundo real en respuesta a una detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real en un sistema de coordenadas tridimensional que representa la escena del mundo real;
un estimador (207) dispuesto para determinar una propiedad del material para el objeto de escena del mundo real en base a los datos de imagen comprendidos en las tramas de imagen de la secuencia de tramas de imagen;
un selector (211) dispuesto para seleccionar un primer objeto de audio del conjunto de objetos de audio en respuesta a la propiedad del material y las características del material vinculadas a objetos de audio del conjunto de objetos de audio;
un circuito de salida (213) dispuesto para generar la señal de audio de salida que comprende el primer objeto de audio; y
un circuito dispuesto para presentar la secuencia de imágenes de salida a un usuario.

2. El aparato de la reivindicación 1, en el que el estimador (207) se dispone para

determinar una región de imagen de interacción en al menos una trama de imagen de la secuencia de tramas de imagen, la región de imagen de interacción es una región de imagen de la al menos una trama de imagen en la que ocurre la interacción; y
determinar la propiedad del material para el objeto de escena en respuesta a los datos de imagen de la región de la imagen de interacción.

3. El aparato de la reivindicación 1 o 2, en el que el segundo receptor (209) se dispone para recibir los metadatos de un servidor remoto.

4. El aparato de cualquier reivindicación anterior, en el que los metadatos para al menos algunos objetos de audio comprenden indicaciones de enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena del mundo real y enlaces entre al menos algunos objetos de audio y características materiales de objetos de escena virtual; y en el que el selector (211) se dispone para seleccionar el primer objeto de audio en respuesta a la propiedad material y características materiales de objetos del mundo real vinculados al conjunto de objetos de audio y en respuesta a una propiedad material del objeto de escena virtual y características materiales de objetos de escena virtual vinculados al conjunto de objetos de audio.

5. El aparato de cualquier reivindicación anterior, en el que el selector (211) se dispone para seleccionar el primer objeto de audio en respuesta a una propiedad dinámica del objeto de escena virtual.

6. El aparato de cualquier reivindicación anterior, en el que el detector (205) se dispone para determinar una propiedad de la interacción y el selector (211) se dispone para seleccionar el primer objeto de audio en respuesta a la propiedad de la interacción.

7. El aparato de la reivindicación 6, en el que la propiedad de la interacción es al menos una propiedad seleccionada del grupo de:

una velocidad de la interacción;
una fuerza de una colisión entre el objeto de escena virtual y el objeto de escena del mundo real;
una elasticidad de una colisión entre el objeto de escena virtual y el objeto de escena del mundo real;
una duración de la interacción; y
una dirección de movimiento del objeto de escena virtual con relación al objeto de escena del mundo real.

8. El aparato de cualquier reivindicación anterior, en el que el selector (211) se dispone para seleccionar el primer objeto de audio en respuesta a una orientación del objeto virtual con relación al objeto de escena del mundo real.

9. El aparato de cualquier reivindicación anterior, en el que el estimador (207) se dispone para determinar una indicación de coincidencia para el objeto de escena del mundo real a al menos una primera categoría de una pluralidad de categorías de objetos; y para determinar la propiedad del material en respuesta a la indicación de coincidencia y las propiedades del material vinculadas a las categorías de objetos.

10. El aparato de la reivindicación 9 que comprende además un receptor de audio (215) para recibir una señal de audio de audio en tiempo real capturada en la escena del mundo real, y en el que el estimador se dispone para determinar la indicación de coincidencia en respuesta a la señal de audio.

11. El aparato de cualquier reivindicación anterior, en el que el selector (211) se dispone para seleccionar el primer objeto de audio como un objeto de audio predeterminado si no se detecta ningún objeto de audio para el cual se cumpla un criterio de selección.

12. El aparato de cualquier reivindicación anterior, en el que al menos una trama de imagen comprende datos de imagen en profundidad y en el que el estimador (207) se dispone para determinar la propiedad del material para el objeto de escena del mundo real en respuesta a una detección de que al menos parte de una región de imagen de la al menos una trama de imagen que representa el objeto de escena del mundo real tiene un nivel de confianza para los datos de imagen en profundidad que no excede un umbral.

13. Un procedimiento para generar una señal de audio de salida, el procedimiento que comprende:

recibir una secuencia de imágenes en tiempo real de una escena del mundo real de un sensor de imagen, la secuencia de imágenes en tiempo real que comprende una secuencia de tramas de imágenes, cada trama de imagen que comprende al menos uno de los datos de imagen visual y los datos de imagen de profundidad;
recibir un conjunto de objetos de audio y metadatos para objetos de audio del conjunto de objetos de audio, los metadatos son indicativos de enlaces entre objetos de audio del conjunto de objetos de audio y características del material;
generar una secuencia de imágenes de salida que comprende un objeto de imagen correspondiente a un objeto de escena virtual en la escena del mundo real;
detectar una interacción entre el objeto de escena virtual y un objeto de escena del mundo real de la escena del mundo real en respuesta a una detección de una proximidad entre el objeto de escena virtual y el objeto de escena del mundo real en un sistema de coordenadas tridimensional que representa la escena del mundo real;
determinar una propiedad del material para el objeto de escena del mundo real en base a los datos de imagen comprendidos en las tramas de imagen de la secuencia de tramas de imagen;
seleccionar un primer objeto de audio del conjunto de objetos de audio en respuesta a la propiedad del material y las características del material vinculadas a objetos de audio del conjunto de objetos de audio;
generar la señal de audio de salida que comprende el primer objeto de audio; y
presentar la secuencia de imágenes de salida al usuario.

14. Un producto de programa de ordenador que comprende medios de código de programa de ordenador adaptados para realizar todas las etapas de la reivindicación 13 cuando dicho programa se ejecuta en un ordenador.

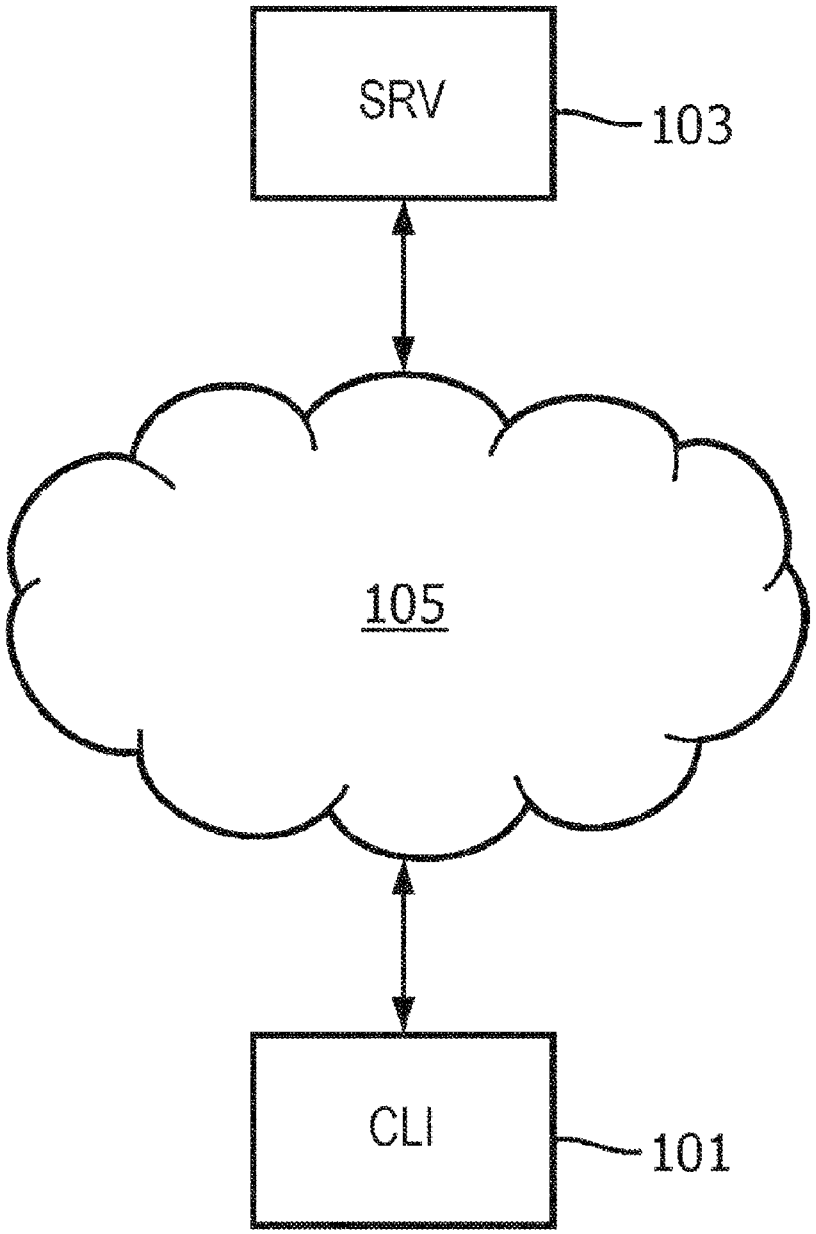

Figura 1

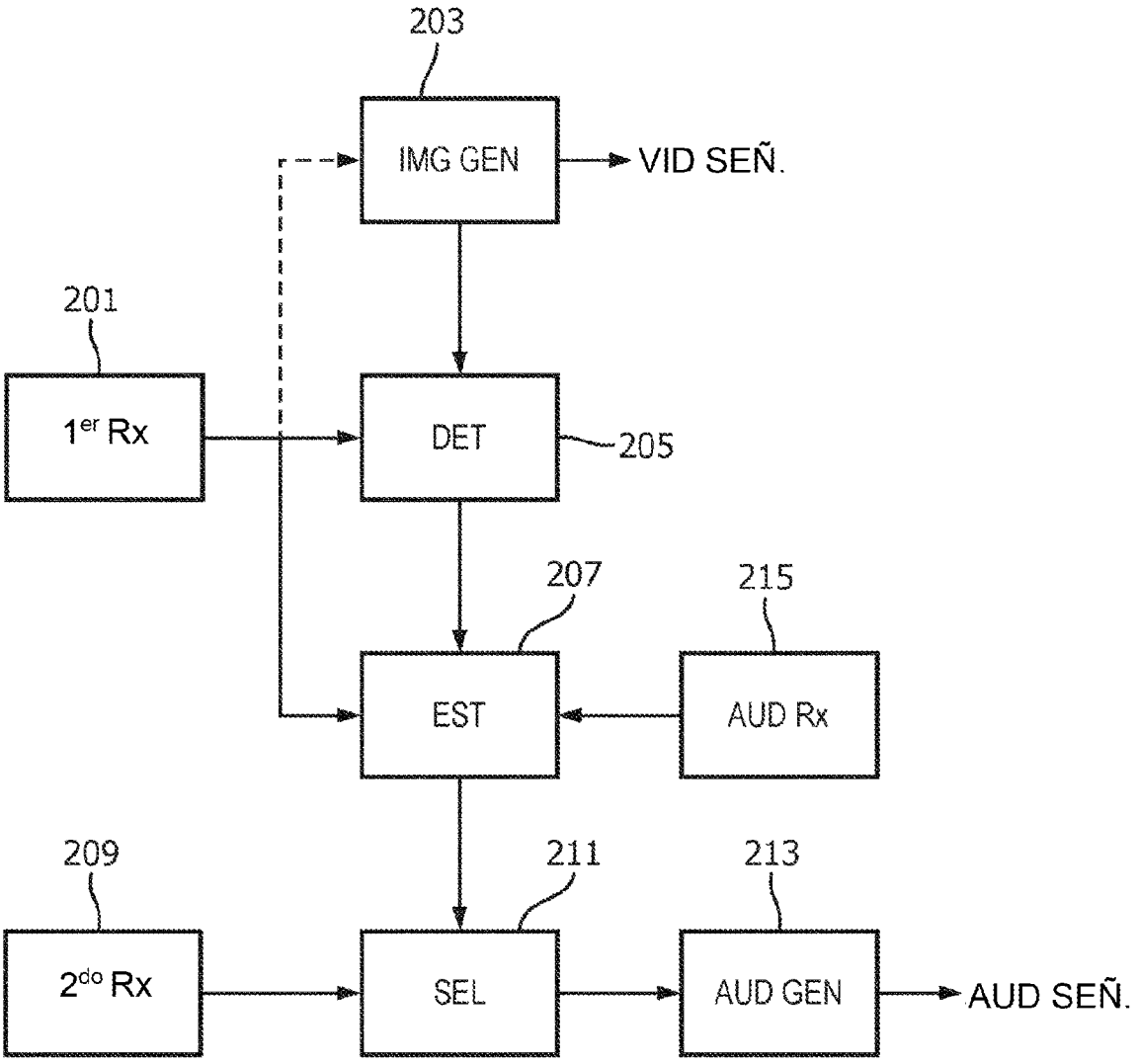
203
IMG GEN
VID SEÑ.
201
1er Rx
DET
205
207
215
EST
AUD Rx
209
211
213
2do Rx
SEL
AUD GEN
AUD SEÑ.

Figura 2

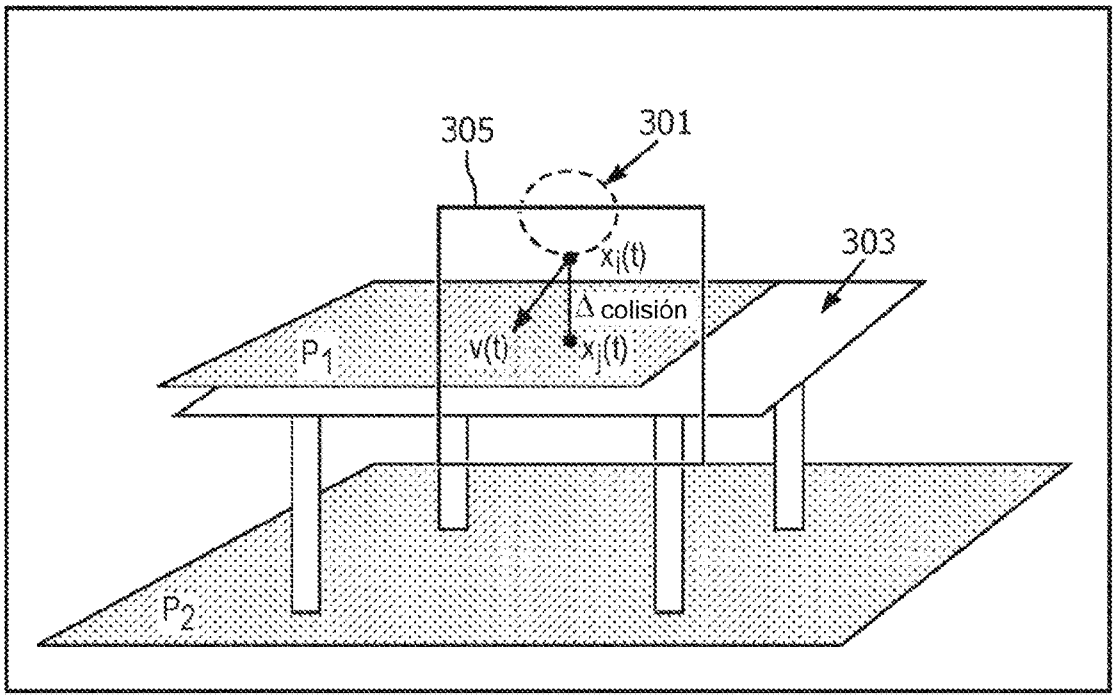

Figura 3

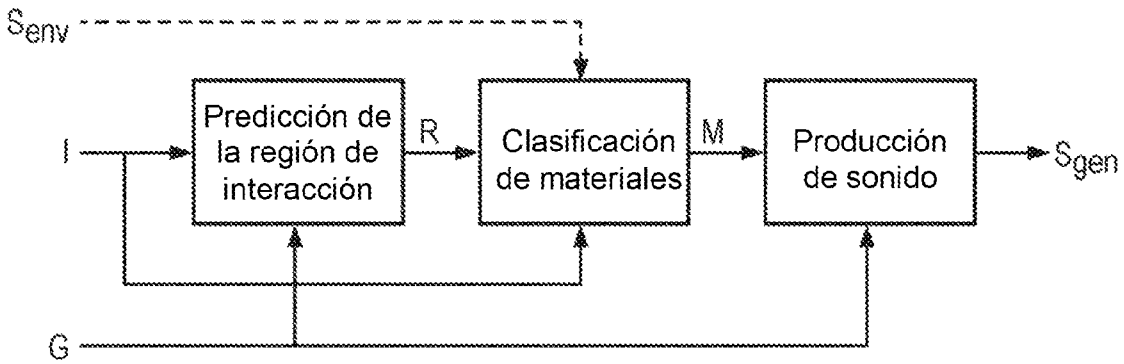

Figura 4