



(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
03.03.2021 Bulletin 2021/09

(51) Int Cl.:
H04S 7/00 (2006.01)

(21) Application number: **20191021.3**

(22) Date of filing: **14.08.2020**

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(72) Inventors:
• **MATE, Sujeet Shyamsundar**
33720 Tampere (FI)
• **LEHTINIEMI, Arto**
33880 Lempäälä (FI)
• **LEPPÄNEN, Jussi**
33580 Tampere (FI)
• **ERONEN, Antti**
33820 Tampere (FI)

(30) Priority: **02.09.2019 FI 20195723**

(71) Applicant: **Nokia Technologies Oy**
02610 Espoo (FI)

(74) Representative: **Nokia EPO representatives**
Nokia Technologies Oy
Karakaari 7
02610 Espoo (FI)

(54) **RENDERING 2D VISUAL CONTENT RELATED TO VOLUMETRIC AUDIO CONTENT**

(57) There is provided rendering 2D visual content and related volumetric audio content. A method comprises determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content, aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume, rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

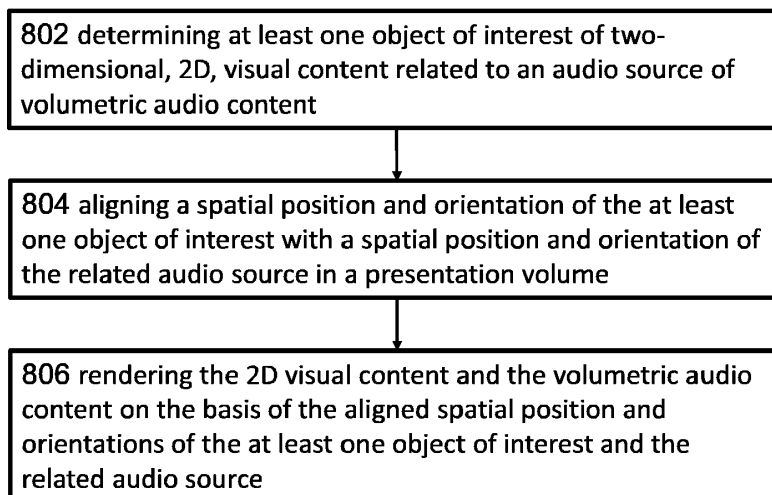


Fig. 8

Description**TECHNICAL FIELD**

5 **[0001]** The present invention relates to an apparatus, a method and a computer program for consuming 2D visual content related to volumetric audio content.

BACKGROUND

10 **[0002]** 2D visual content can be generated by mobile devices equipped with cameras and microphones. The generated 2D visual content can be consumed, e.g. viewed, on the same mobile device it is generated and/or the 2D visual content can be shared for consumption on another device such as another mobile device.

[0003] Volumetric video and audio data represent a three-dimensional scene with spatial audio, which can be used as input for virtual reality (VR), augmented reality (AR) and mixed reality (MR) applications. The user of the application can move around in the blend of physical and digital content, and digital content presentation is modified according to user's position and orientation. Most of the current applications operate in three degrees-of-freedom (3DoF), which means that head rotation in three axes yaw/pitch/roll can be taken into account. However, the development of VR/AR/MR applications is eventually leading to 6DoF volumetric virtual reality, where the user is able to freely move in a Euclidean space (x, y, z) and rotate his/her head (yaw, pitch, roll). 6DoF audio content provides rich and immersive experience of the audio scene.

20 **[0004]** 6DoF audio content can be consumed in conjunction with 2D visual content. If a visual scene provided by the 2D visual content and an audio scene provided by the 6DoF audio content are not aligned, audio sources can be heard by the user from different directions compared with directions, where the user can see visual counterparts of the audio sources.

25

SUMMARY

[0005] The scope of protection sought for various embodiments of the invention is set out by the independent claims. The embodiments, examples and features, if any, described in this specification that do not fall under the scope of the independent claims are to be interpreted as examples useful for understanding various embodiments of the invention.

30 **[0006]** According to a first aspect, there is provided a method comprising, determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume;

35 rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0007] According to a second aspect there is provided an apparatus comprising, means for determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; means for aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume; and means for rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

40 **[0008]** According to a third aspect there is provided a computer program comprising computer readable program code means adapted to perform at least the following: determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume; rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

50 **[0009]** According to a fourth aspect, there is provided an apparatus comprising at least one processor; and at least one memory including computer program code; the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus at least to:

determine at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content;

align a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume;

render the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and

orientations of the at least one object of interest and the related audio source.

[0010] According to a fifth aspect, there is provided a computer program according to an aspect embodied on a computer readable medium.

[0011] According to a sixth aspect, there is provided a non-transitory computer readable medium comprising program instructions stored thereon for performing at least the following: determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume; rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0012] According to one or more further aspects, embodiments according to the first, second, third, fourth, fifth and sixth aspect comprise one or more features of:

- obtaining information indicating spatial positions and orientations of one or more visual objects of the 2D visual content and related audio sources of the volumetric audio content; and aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source in the presentation volume on the basis of the obtained information
- wherein the information indicating spatial position and orientation is included in metadata of at least one of the 2D visual content and the volumetric audio content
- determining the at least one object of interest of the 2D visual content on the basis of:
 - a user input; or
 - metadata associated with the 2D visual content
- in response to determining a misalignment of the at least one object of interest with the related audio source, re-aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source
- wherein the re-aligning comprises at least one of:
 - adapting a visual zoom of the 2D visual content in the presentation volume;
 - moving a visual rendering plane of the 2D visual content in the presentation volume; and
 - adapting orientation of the 2D visual content in the presentation volume
- obtaining information indicating a permissible zooming of the 2D visual content; zooming the 2D visual content within the indicated permissible zooming for aligning the at least one object of interest with the related audio source
- determining an initial position of the 2D visual content in the presentation volume; and in response to a user input indicating a subsequent position of the 2D visual content, re-aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source in the presentation volume, and rendering the 2D visual content and the volumetric audio content using the re-aligned spatial position and orientation.
- modifying the volumetric audio content for reducing depth differences between audio sources; and rendering the 2D visual content and the volumetric audio content using the modified volumetric audio content
- wherein the 2D visual content is positioned as world locked content in the presentation volume.

[0013] Apparatuses according to some embodiments comprise at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform the above methods.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] For a more complete understanding of example embodiments of the present invention, reference is now made to the following descriptions taken in connection with the accompanying drawings in which

Fig. 1 illustrates an example of an application supporting consumption of volumetric audio content associated with 2D visual content, in accordance with at least some embodiments of the present invention;
 Figs. 2 and 3 that illustrate examples of 2D visual content and related volumetric audio content rendered in a presentation volume in accordance with at least some embodiments of the present invention;
 Fig. 4, 5 and 6 illustrate examples of spatial positions and orientations of 2D visual content related to volumetric

audio content in accordance with at least some embodiments of the present invention;

Fig. 7 illustrates determining a spatial position and orientation of 2D visual content in a presentation volume on the basis of user input in accordance with at least some embodiments of the present invention;

Fig. 8 illustrates an example of a method in accordance with at least some embodiments of the present invention;

Fig. 9 an example of a method in accordance with at least some embodiments of the present invention;

Fig. 10 illustrates an example of a method for maintaining alignment of 2D visual content and volumetric audio in an audio-visual scene in accordance with at least some embodiments of the present invention;

Fig. 11 illustrates an example of a method for controlling zooming of 2D visual content in accordance with at least some embodiments of the present invention;

Fig. 12 illustrates an example of a method for controlling a need for changing a resolution of 2D visual content in accordance with at least some embodiments of the present invention;

Fig. 13 illustrates an example of a method for rendering volumetric audio in conjunction with visual content for playback of an audio-visual scene in a presentation volume, in accordance with at least some embodiments of the invention;

Fig. 14 shows a system for capturing, encoding, decoding, reconstructing and viewing a three-dimensional scene, that is, for visual content and 3D audio digital creation and playback in accordance with at least some embodiments of the present invention;

Fig. 15 depicts example devices for implementing various embodiments;

Fig. 16 shows schematically an electronic device employing embodiments of the invention;

Fig. 17 shows schematically a user equipment suitable for employing embodiments of the invention; and

Fig. 18 shows an example of a system within which embodiments of the present invention can be utilized.

DETAILED DESCRIPTION OF SOME EXAMPLE EMBODIMENTS

[0015] In the following, several embodiments will be described in the context of rendering volumetric audio content associated with two-dimensional, 2D, visual content. It is to be noted, however, that while some of the embodiments are described relating to certain coding technologies, the invention is not limited to any specific volumetric audio or visual content technology or standard. In fact, the different embodiments have applications in any environment where volumetric audio may be consumed in conjunction with associated 2D visual content. Thus, applications including but not limited to general computer gaming, virtual reality, or other applications of digital virtual acoustics can benefit from the use of the embodiments.

[0016] In connection with two-dimensional, 2D, visual content related to an audio source of volumetric audio content, there is provided determining at least one object of interest of the 2D visual content. Spatial position and orientation of the at least one object of interest is aligned with a spatial position and orientation of the related audio source in a presentation volume. The 2D visual content and the volumetric audio content are rendered on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source. In this way, a user may experience the object of interest and related audio source from a uniform direction.

[0017] A presentation volume may be defined as a closed region in a volumetric scene, within which a user may be able to move and view the scene content with full immersion and with all physical aspects of the scene accurately represented. Defining a presentation volume in a volumetric presentation may be useful in (limited) 6 DoF environment where the presentation volume, from where the content can be immersively consumed, may need to be restricted and decided beforehand.

[0018] The scene content may comprise 2D visual content and related volumetric audio content. The 2D visual content may be rendered for playback of a visual scene on the basis of the rendered 2D visual content. The volumetric audio content may be rendered for playback of an audio scene on the basis of the rendered volumetric audio content. Both 2D visual content and related volumetric audio content may be rendered for playback of an audio-visual scene in a presentation volume.

[0019] An example of the presentation volume is a single contained closed volumetric area or region, multiple disjoint closed regions of presentation locations, or multiple volumetric areas or regions combined together. Presentation volume representation may be useful in MPEG-I activities e.g. 3DoF+ and upcoming 6DoF areas.

[0020] The presentation volume may be of any size, very small or very large area. Often the presentation volume might only have valid locations for consuming volumetric content in some parts of the volumetric scene. Overall, the shape of the viewing volume can be very complicated in a large 6 DoF scene.

[0021] Examples of two-dimensional, 2D, visual content comprise images and video that are rendered on a 2D plane for presentation to a user. The 2D visual content may be digital content coded in a content format. Examples of the content formats comprise at least computer-readable image file formats and computer-readable video file formats.

[0022] Fig. 1 illustrates an example of an application supporting consumption of volumetric audio content associated with 2D visual content. Functionalities of the application are illustrated in views A, B and C of the application, when the

application is used by a user. The application may be a gallery application for presenting the user a plurality of media items, M1,M2,M3,M4 102 for selection by the user. The gallery application may be displayed to the user on a user interface that also supports receiving input from the user.

[0023] In the view A, the media items are displayed in the gallery application for selection by the user. The media items may comprise 2D visual content and related volumetric audio content. Accordingly, by selecting one of the media items, the user may choose which of the 2D visual content and associated volumetric audio content he/she would like to consume.

[0024] In the view B, a user input 104, received by the gallery application for selecting a media item M2 is illustrated. In an example, the user may touch, e.g. long press, an area representing the media item M2 for selecting 2D visual content and related volumetric audio content. Other media items may be selected in a similar manner by the user.

[0025] In the view C, playback 106 of the selected 2D visual content and associated volumetric audio content in response to the selection of the media item in view B is illustrated. The playback comprises presenting the user an audio-visual scene formed by rendering the selected 2D visual content and related volumetric audio content in a presentation volume.

[0026] It is referred to Figs. 2 and 3 that illustrate examples of 2D visual content and related volumetric audio content rendered in a presentation volume 202 in accordance with at least some embodiments of the present invention. The 2D visual content and related volumetric audio content rendered may be rendered in response to a user 204 selecting a media item 206 from an application 208 supporting consumption of volumetric audio content associated with 2D visual content, in accordance with Fig. 1.

[0027] In an embodiment, the 2D visual content is rendered as world locked content in the presentation volume 202. In this way the 2D visual content may be positioned at a suitable distance from the user 204 in the presentation volume. The world locked 2D visual content should be understood to be positioned at a specific location within the presentation volume. It should be appreciated that the location of the world locked 2D visual content may be adapted to take into account a movement of the user or a change of the object of interest, OOI. In an example, adapting the location of the 2D visual content may comprise adapting a position and/or orientation of the 2D visual content. In an example, the world lock may be achieved by using an Augmented Reality, AR, tracking enabler such as AR Toolkit or AR core.

[0028] In accordance with at least some embodiments, the 2D visual content may comprise one or more visual objects, 1', 2',3', that may have one or more related audio sources, 1, 2, 3, in the volumetric audio content. At least one of the visual objects may be determined as the object of interest, OOI, 210 for aligning the 2D visual content and the volumetric audio content. Rendering the aligned 2D visual content and the volumetric audio content facilitates the user 204 to experience the object of interest and related audio source from a uniform direction.

[0029] According to an embodiment, an object of interest 210 of the 2D visual content is determined on the basis of a user input or metadata associated with the 2D visual content.

[0030] In an example, the 2D visual content may be world locked content that is displayed by a user device 212 on a viewfinder 218 of the user device. On the other hand the user device may be HMD or connected to HMD and the 2D visual content may be displayed by the HMD in see-through mode. The user may select one of the visual objects, 1',2',3' of the 2D visual content as the object of interest 210 by pointing on the visual object, for example by a touch operation or by a pointing device. The 2D visual content may be associated with metadata that comprises information identifying one or more visual objects in a visual scene rendered based on the 2D visual content, whereby the metadata may be used to determine the visual object selected by the user. In an example, the object of interest may be determined on the basis of the metadata associated with the 2D visual content, without necessarily any user input for indicating the object of interest. The metadata may comprise information identifying one or more visual objects of the 2D visual content and one of the identified visual objects in the metadata may be associated with information indicating the visual object as an object of interest.

[0031] The user may move from his initial position 214 in the presentation volume to another position 216. Movement of the user may cause a misalignment of the object of interest and related audio source, whereby the spatial position and orientation of the object of interest and the related audio source need to be aligned/re-aligned for rendering an audio-visual scene based on the 2D visual content and the volumetric audio content, where the user 204 may experience the object of interest and related audio source from a uniform direction.

[0032] Fig. 4, 5 and 6 illustrate examples of spatial positions and orientations of 2D visual content related to volumetric audio content in accordance with at least some embodiments of the present invention. The spatial positions and orientations may define spatial positions and orientations in an audio-visual scene rendered to a presentation volume. Accordingly, the spatial positions and orientations may define rendering positions and rendering orientations in a rendering space, i.e. the presentation volume. The volumetric audio content related to the 2D visual content 402 comprises audio sources 1, 2, 3, 4, 5, 6, 7 that have spatial positions in a presentation volume. Audio sources 1, 2, 3, 4 have related visual objects 1', 2',3',4' in the 2D visual content 402. A user 404 may be positioned in the presentation volume to consume an audio-visual scene rendered on the basis of the 2D visual content and the related volumetric audio content. The user experiences the visual objects and audio sources in directions 406 with respect to the user according to the

spatial positions and orientations of the visual objects and audio sources. At least one of the visual objects may be determined as an object of interest for aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in the presentation volume. In this way the 2D visual content and the volumetric audio content may be rendered on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source, whereby the user 404 positioned in the presentation volume may experience the visual object and audio from the related audio source from a uniform direction. It should be appreciated that optionally the volumetric audio content may comprise one or more audio sources, e.g. 5, 6, 7, that may not be related to visual objects. In an example, the audio sources 1, 2, 3, 4, 5, 6, 7 may be captured by a camera and microphones. Accordingly, the audio sources may be related to visual objects captured by the camera. In an example, audio objects from a person singing may be captured related to visual objects of the person formed by the camera. However, in case the camera does not capture visual objects of the person singing, e.g. due to limited field of view of the camera, audio objects may be still captured. However, in such a case the audio objects are not related to visual objects.

[0033] Referring to the example of Fig. 4, the spatial positions of the 2D visual content and related audio sources are illustrated in the view A on the left-hand side and a more detailed example of the 2D visual content is illustrated in view B on the right-hand side. The 2D visual content 402 may be positioned at spatial positions 408 that are at different distances from the user 404, while visual objects of the 2D visual content may be maintained aligned with related audio sources. Resolution and/or dimensions of the 2D visual content may be adapted for aligning one or more of the visual objects with the audio sources. In an example, a resolution and/or dimensions of the 2D visual content may be greater at a spatial position that is at distance that is closer to the user than at another spatial position that is at a distance that is further away from the user. The dimensions of the 2D visual content may be defined by a width and height of the 2D visual content.

[0034] In one example, referring to the example of Fig. 5, orientation of the 2D visual content 402 may be adapted for aligning one or more spatial orientations of the objects of interest, e.g. 3', with one or more orientations of the related audio sources, e.g. 3, in the presentation volume. In this way, when the user 404 moves 410 from one position in the presentation volume to a subsequent position, the user may experience the visual objects and related audio sources from a uniform direction. It should be appreciated that although, the 2D visual content is illustrated in Fig. 5 in two orientations, also further orientations are viable. In one example, the orientation of the 2D visual content may be adapted by rotating the 2D visual content. The rotation of the visual content may be caused by rotating a rendering plane of the 2D visual content, whereby the 2D visual content may be rendered on the plane that has been rotated for aligning at least one visual object, e.g. an object of interest, with related audio source.

[0035] In one example, referring to the example of Fig. 5, the 2D visual content may be positioned at spatial position, where at least one of the visual objects, 3', e.g. the visual object of interest, is aligned with the related audio source 3. In this example, the user may be moved 410 from one position to a subsequent position, where in both positions the visual object 3' and the related audio source are aligned. In the first position of the user, the 2D visual content may be positioned for example in one of the positions illustrated in Fig. 4, where the audio sources are aligned with related visual objects. In the subsequent position, the visual object and the related audio source may be co-located at the same spatial location, for example.

[0036] Referring to Fig. 6, a misalignment of the 2D visual content 402 and related volumetric audio is illustrated. View A of the Fig. 6 illustrates misalignment in the presentation volume and view B illustrates the misalignment. The misalignment by a skew caused by the movement 410 of the user 404. The 2D visual content has a spatial position and orientation in the presentation space arranged such that visual object 3' and related audio source 3 are co-located. At a first position of the user, the visual object of interest is item 2' that is aligned with related audio source 2. After the user has moved to a subsequent position, there is a skew, i.e. a difference between directions 406 of the audio source 2 and the related visual object 2' with respect to the user. In Fig. 6 the skew is illustrated by Angular Difference, AD, between the directions 406 of the object of interest 2' and the related audio source 2. Accordingly, the misalignment may be determined on the basis of the skew. It should be appreciated that in practise some skew may be permitted for the sake efficiency of implementation, at least if the skew cannot be perceived by human.

[0037] Fig. 7 illustrates determining a spatial position and orientation of 2D visual content in a presentation volume on the basis of user input in accordance with at least some embodiments of the present invention. 2D visual content 702 may be determined at an initial position in the presentation volume 704, for example as world locked content in accordance with Fig. 2. The 2D visual content may be displayed to a user 710 by a user device, for example on a viewfinder or by HMD in a see-through mode. In an example, the initial position of the 2D visual content may be at default distance from a user device 706. In response to a user input indicating a subsequent position and/or orientation of the 2D visual content in the presentation space, the spatial position and orientation of the at least one object of interest may be re-aligned with the spatial position and orientation of the related audio source, e.g. audio sources 1, 2, or 3, in the presentation volume, and the 2D visual content and the volumetric audio content may be rendered using the re-aligned spatial position and orientation. The object of interest may be determined on the basis of user input or metadata from the visual objects 1', 2', 3' of the 2D visual content as described with reference to Fig. 3 above.

[0038] In an example, the user device 706 may be predefined position and orientation in the presentation space of the volumetric audio scene and the 2D visual content may be at the initial position. The user 710 may enter user input on the user device 706 for indicating one or more subsequent positions and/or orientations for the 2D visual content. In an example, the user input may comprise gestures that cause traversing the presentation volume of the volumetric audio scene. For example, the gestures may comprise swiping up-down, left-right or an inverted U that may cause longitudinal movement, lateral movement and orientation change respectively, of the 2D visual content in the presentation space.

[0039] Fig. 8 illustrates an example of a method in accordance with at least some embodiments of the present invention. The method facilitates generating an audio-visual scene, where a user experiences the object of interest and related audio source from a uniform direction.

[0040] Phase 802 comprises determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content.

[0041] Phase 804 comprises aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume.

[0042] Phase 806 comprises rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0043] Since the spatial positions and orientations of the object of interest and the related audio source are aligned in phase 804, playback of the 2D visual content and the related volumetric audio content rendered in phase 806 causes generating an audio-visual scene, where the user experiences the object of interest and related audio source from a uniform direction.

[0044] Fig. 9 an example of a method in accordance with at least some embodiments of the present invention. The method facilitates rendering 2D visual content related to volumetric audio content.

[0045] Phase 902 comprises obtaining information indicating spatial positions and orientations of one or more visual objects of the 2D visual content and related audio sources of the volumetric audio content.

[0046] Phase 904 comprises aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source in the presentation volume on the basis of the obtained information.

[0047] In this way the obtained information may be used for aligning objects of interest for rendering the 2D visual content and the volumetric audio content for rendering an audio-visual scene.

[0048] In an embodiment, phase 902 comprises that the information indicating spatial position and orientation is included in metadata of at least one of the 2D visual content and the volumetric audio content. The metadata may be provided together with the 2D visual content and/or the volumetric audio content. Alternatively or additionally, the metadata may be provided separately from the 2D visual content and/or the volumetric audio content. For example, the metadata may be received from a capturing device or a network accessible service in response to a user input for playback of the 2D visual content and related volumetric audio content.

[0049] In an example, the phase 902 comprises that the obtained information comprises information indicating an initial position and 3D orientation of the volumetric audio scene to correspond with at least a subset of the 2D visual scene. In this way spatial positions of one or more audio sources of the volumetric audio scene may be determined with respect to the 2D visual scene in the audio-visual scene.

[0050] An example of the information obtained in phase 902 comprises an audio-visual correspondence data structure, where positions of objects of the 2D visual content in an audio scene and positions of the audio sources of the volumetric audio content in the audio scene may be defined. An example of the audio-visual correspondence data structure is as follows:

```
aligned(8) class AudioAlignmentStruct() {
    signed int(32) audio_scene_pos_x;
    signed int(32) audio_scene_pos_y;
    signed int(32) audio_scene_pos_z;
    signed int(32) audio_scene_yaw;
    signed int(32) audio_scene_pitch;
    signed int(32) audio_scene_roll;
} aligned(8) class VisualAlignmentStruct() {
    signed int(32) visual_scene_pos_x;
    signed int(32) visual_scene_pos_y; }
```

[0051] Sample syntax for audio-visual alignment using the audio-visual correspondence data structure is as follows:

```
aligned(8) class AudioVisualAlignmentStruct() {
    for(i=0; i<num_OOIs_in_2DVisual; i++){
        AudioAlignmentStruct();
    }
```

```

VisualAlignmentStruct();
} }

```

[0052] In one implementation example, for a static 2D visual content, the AudioAlignmentStruct() and the VisualAlignmentStruct() may be signaled in the header of the audio tracks representing the volumetric audio scene and the track or item for visual content respectively.

[0053] In case of dynamic scenes, the audio-visual alignment structure needs vary over time. Consequently, this information can be signaled as a time varying metadata track with 'cdsc' reference.

[0054] Fig. 10 illustrates an example of a method for maintaining alignment of 2D visual content and volumetric audio in an audio-visual scene in accordance with at least some embodiments of the present invention. The 2D visual content and volumetric audio content may have been aligned in accordance with the method of Fig. 9.

[0055] Phase 1002 comprises determining a misalignment of the at least one object of interest with related audio source.

[0056] In an example the misalignment may be determined on the basis of a skew, i.e. a difference between directions of the audio source and the visual object with respect to the user. It should be appreciated that in practise some skew may be permitted for the sake efficiency of implementation, at least if the skew cannot be perceived by human.

[0057] In an example the misalignment may be caused by a change of the object of interest, a movement of the object of interest and/or a movement of the user.

[0058] Phase 1004 comprises in response to determining the misalignment in phase 1002, re-aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source. In this way the skew may be reduced such that the user may experience the object of interest and the audio source from a uniform direction. It should be appreciated that once the re-aligning is performed, the 2D visual content and the volumetric audio content may be rendered on the basis of the re-aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0059] In an embodiment, phase 1004 comprises that the re-aligning comprises at least one of:

- adapting a visual zoom of the 2D visual content in the presentation volume;
- moving a visual rendering plane of the 2D visual content in the presentation volume; and
- adapting orientation of the 2D visual content in the presentation volume.

[0060] In an example, the misalignment may be determined in phase 1002, when the object of interest is changed, or the object of interest has moved, the re-aligning in phase 1004 may be performed by adapting the visual zoom and/or by moving the visual rendering plane. On the other hand, if the user has moved, the re-aligning may be performed by adapting the orientation of the 2D visual content.

[0061] In an example of adapting the visual zoom in phase 1004, the visual zoom may be adapted on the basis of a skew, i.e. a difference between directions of the audio source and the visual object with respect to the user. Adapting the visual zoom provides that resolution of the 2D visual content may be increased, whereby the skew may be masked by the 2D visual content experienced by the user. Adapting the visual zoom on the basis of the skew is described with reference to the example of the skew described in Fig. 6 view B. The skew may be defined by an Angular Difference, AD, between a direction of the object of interest and a direction of the related audio source with respect to the user and the skew may be expressed by

$$AD = \sin \phi_1 - \sin \phi_2 \quad (1),$$

where ϕ_1 is the angle between a reference direction and the direction of the related audio source with respect to the user and ϕ_2 is the angle between the reference direction and the direction of the object of interest. The visual zoom may be adapted by rendering the 2D visual content using a resolution determined on the basis of a multiplier based on the AD and a current spatial rendering resolution of the 2D visual content. An example of the multiplier may be expressed by

$$\text{multiplier} = (1+AD) * \text{current_spatial_rendering_resolution} \quad (2)$$

[0062] In an example of moving the visual rendering plane in phase 1004, moving the visual rendering provides that the 2D visual content may be moved to another spatial position in the presentation volume. In one example, the 2D visual content may be positioned first at one of the positions illustrated in Fig. 4 by rendering the 2D visual content on a rendering plane positioned at one of the illustrated positions. If the object of interest is changed from 2' to 3', the 2D visual content may be moved to a position illustrated in Fig. 6, where the object of interest 3' is aligned with the related audio source. Accordingly, the rendering plane may be moved to the new position illustrated in Fig. 6 for positioning the

2D visual content at that position.

[0063] In an example of adapting orientation of the 2D visual content, the rendering orientation of the 2D visual content may be changed. The rendering orientation may be changed for example by adapting orientation of a visual rendering plane.

[0064] Fig. 11 illustrates an example of a method for controlling zooming of 2D visual content in accordance with at least some embodiments of the present invention.

[0065] Phase 1102 comprises obtaining information indicating a permissible zooming of the 2D visual content;

[0066] Phases 1104 comprises zooming the 2D visual content within the indicated permissible zooming for aligning the at least one object of interest with related audio source maintaining the 2D visual content within the indicated permissible zooming.

[0067] In an example, the zooming comprises that resolution of the 2D visual content is adapted, e.g. increased. In this way a user may experience the object of interest and related audio source from a uniform direction. Phase 1004 in fig. 10 describes an example of the zooming.

[0068] Phase 1106 comprises determining if zooming the 2D visual content further would be within the indicated permissible zooming. If the zooming would not be within the permissible zooming, the method may end 1108. In this way zooming the 2D visual content may be kept within the permissible zooming. This has the advantage that a possible loss in visual content quality caused by the zooming may be controlled. On the other hand, if the zooming would be within the indicated permissible zooming, the method may proceed to phase 1110. In phase 1110 it may be determined, if further zooming is needed. further zooming may be needed for example, if misalignment has not yet been compensated by the previous zooming. If further zooming is needed, the method may continue to phase 1104. If no further zooming is needed, e.g. there is no misalignment between the object of interest and related volumetric audio content, the method may proceed to end 1108.

[0069] Fig. 12 illustrates an example of a method for controlling a need for changing a resolution of 2D visual content in accordance with at least some embodiments of the present invention.

[0070] Phase 1202 comprises modifying the volumetric audio content for reducing depth differences between audio sources.

[0071] Phase 1204 comprises rendering the 2D visual content and the volumetric audio content using the modified volumetric audio content. Since the depth differences between audio sources are reduced in the volumetric audio content, the modified content facilitates rendering a non-pliable volumetric audio scene, whereby a need to change resolution of the 2D visual content may be reduced.

[0072] Non-pliable volumetric audio scene refers to audio representation which gives a spatial audio experience which responds to user movement with six degrees of freedom. However, the content representation is such that it cannot be modified to change the perceived user positions. For example, if the non-pliable audio content has objects 1 and 2 in positions (x1, y1) and (x2, y2), the audio source positions cannot be changed in order to make one or both of them aligned with the visual objects in 2D visuals. This inability to modify the volumetric audio scene, makes it essential to modify the rendering position/zoom/orientation of the 2D visuals.

[0073] Fig. 13 illustrates an example of a method for rendering volumetric audio in conjunction with visual content for playback of an audio-visual scene in a presentation volume, in accordance with at least some embodiments of the invention.

[0074] Phase 1302 comprises receiving 2D visual content and related volumetric audio content. In an example, the 2D visual content and the volumetric audio content may have one or more visual objects and one or more audio sources related to the visual objects. The 2D visual content and volumetric audio may be obtained by a user device.

[0075] Phase 1304 comprises determining a spatial position of the 2D visual content in a presentation volume. In an example the 2D visual content may be determined as a world locked content or with respect to a user device.

[0076] Phase 1306 and 1308 comprise determining at least one visual object of interest of the 2D visual content. A spatial position and orientation of the at least one visual object of interest may be determined in the presentation volume. In an embodiment, the 2D visual content is rendered as world locked content in the presentation volume.

[0077] In an example, phase 1306 may comprise determining coordinates of the object of interest in the 2D visual content, in a 2D plane and phase 1308 may comprise determining coordinates of the visual object of interest in the presentation volume, i.e. in a 3D space.

[0078] In an embodiment, phase 1307 comprises obtaining information indicating spatial positions and orientations of one or more visual objects of the 2D visual content and related audio sources of the volumetric audio content. The obtained information may be used for determining at least one visual object of interest of the 2D visual content in phase 1306.

[0079] In an embodiment, phase 1307 comprises that the information indicating spatial position and orientation is included in metadata of at least one of the 2D visual content and the volumetric audio content.

[0080] It should be appreciated that alternatively or additionally, phase 1307 comprises determining the information indicating spatial positions and orientations on the basis of content analysis of the 2D visual content and the volumetric

audio content.

[0081] Phase 1310 comprises determining a spatial position and orientation of the volumetric audio content in alignment with the 2D visual content. In an example, a direction of the determined at least one visual object of interest is aligned with a direction of an audio source related to the determined visual object. In this way the user experiences the visual object and audio source in a uniform direction.

[0082] Phase 1312 comprises rendering an audio-visual scene from the 2D visual content and the volumetric audio. In an example, the audio-visual scene is rendered on the basis of the aligned directions of the visual object of interest and the related audio source. In this way the user may experience the object of interest and related audio source from a uniform direction.

[0083] In at least some embodiments, phases 1314 to 1320 comprise determining a misalignment of the at least one object of interest with related audio source and re-aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source.

[0084] In an example, phase 1314 comprises determining an object of interest on the basis of: a user input; or metadata associated with the 2D visual content. The determined object of interest may be a different object of interest than the object of interest determined in phase 1306.

[0085] Phase 1316 comprises determining a misalignment with the object of interest determined in phase 1314 and the related audio source. Accordingly, when the object of interest and the related audio source are misaligned the user may experience the audio from a different direction than the direction, where the object of interest.

[0086] Phase 1318 comprises in response to determining the misalignment in phase 1316, re-aligning the spatial position and orientation of the at least one object of interest with the spatial position and orientation of the related audio source. In this way the user may experience the object of interest and related audio source from a uniform direction. In an embodiment, phase 1318 comprises at least one of:

- adapting a visual zoom of the 2D visual content in the presentation volume;
- moving a visual rendering plane of the 2D visual content in the presentation volume; and
- adapting orientation of the 2D visual content in the presentation volume.

[0087] Phase 1320 comprises rendering the 2D visual content and the volumetric audio content using the re-aligned spatial position and orientation. In this way the misalignment may be compensated and the user experiences the visual object and audio source in a uniform direction.

[0088] Fig. 14 shows a system for capturing, encoding, decoding, reconstructing and viewing a three-dimensional scheme, that is, for visual content and 3D audio digital creation and playback in accordance with at least some embodiments of the present invention. The visual content may be 2D images or 3D images, or the visual content may be 2D video or 3D video, for example. However, in the following description of the system of Fig. 14 the example of 3D video is used. The system is capable of capturing and encoding volumetric video and audio data for representing a 3D scene with spatial audio, which can be used as input for virtual reality (VR), augmented reality (AR) and mixed reality (MR) applications. The task of the system is that of capturing sufficient visual and auditory information from a specific scene to be able to create a scene model such that a convincing reproduction of the experience, or presence, of being in that location can be achieved by one or more viewers physically located in different locations and optionally at a time later in the future. Such reproduction requires more information that can be captured by a single camera or microphone, in order that a viewer can determine the distance and location of objects within the scene using their eyes and their ears. To create a pair of images with disparity, two camera sources are used. In a similar manner, for the human auditory system to be able to sense the direction of sound, at least two microphones are used (the commonly known stereo sound is created by recording two audio channels). The human auditory system can detect the cues, e.g. in timing and level difference of the audio signals to detect the direction of sound.

[0089] The system of Fig. 14 may consist of three main parts: image/audio sources, a server and a rendering device. A video/audio source SRC1 may comprise multiple cameras CAM1, CAM2, ..., CAMN with overlapping field of view so that regions of the view around the video capture device is captured from at least two cameras. The video/audio source SRC1 may comprise multiple microphones uP1, uP2, ..., uPN to capture the timing and phase differences of audio originating from different directions. The video/audio source SRC1 may comprise a high-resolution orientation sensor so that the orientation (direction of view) of the plurality of cameras CAM1, CAM2, ..., CAMN can be detected and recorded. The cameras or the computers may also comprise or be functionally connected to means for forming distance information corresponding to the captured images, for example so that the pixels have corresponding depth data. Such depth data may be formed by scanning the depth or it may be computed from the different images captured by the cameras. The video source SRC1 comprises or is functionally connected to, or each of the plurality of cameras CAM1, CAM2, ..., CAMN comprises or is functionally connected to a computer processor and memory, the memory comprising computer program code for controlling the source and/or the plurality of cameras. The image stream captured by the video source, i.e. the plurality of the cameras, may be stored on a memory device for use in another device, e.g. a viewer,

and/or transmitted to a server using a communication interface. It needs to be understood that although a video source comprising three cameras is described here as part of the system, another amount of camera devices may be used instead as part of the system.

[0090] It also needs to be understood that although microphones uP1 to uPN have been depicted along with cameras CAM1 to CAMN in Fig. 14 this does not need to be the case. For example, a possible scenario is that closeup microphones are used to capture audio sources at close proximity to obtain a dry signal of each source such that minimal reverberation and ambient sounds are included in the signal created by the closeup microphone source. The microphones co-located with the cameras can then be used for obtaining a wet or reverberant capture of the entire audio scene where the effect of the environment such as reverberation is captured as well. It is also possible to capture the reverberant or wet sound of single objects with such microphones if each source is active at a different time. Alternatively or in addition to, individual room microphones can be positioned to capture the wet or reverberant signal. Furthermore, each camera CAM1 through CAMN can comprise several microphones, such as two or 8 or any suitable number. There may also be additional microphone arrays which enable capturing spatial sound as first order ambisonics (FOA) or higher order ambisonics (HOA). As an example, a SoundField microphone can be used.

[0091] One or more two-dimensional video bitstreams and one or more audio bitstreams may be computed at the server SERVER or a device RENDERER used for rendering, or another device at the receiving end. The devices SRC1 and SRC2 may comprise or be functionally connected to one or more computer processors (PROC2 shown) and memory (MEM2 shown), the memory comprising computer program (PROGR2 shown) code for controlling the source device SRC1/SRC2. The image/audio stream captured by the device may be stored on a memory device for use in another device, e.g. a viewer, or transmitted to a server or the viewer using a communication interface COMM2. There may be a storage, processing and data stream serving network in addition to the capture device SRC1. For example, there may be a server SERVER or a plurality of servers storing the output from the capture device SRC1 or device SRC2 and/or to form a visual and auditory scene model from the data from devices SRC1, SRC2. The device SERVER comprises or is functionally connected to a computer processor PROC3 and memory MEM3, the memory comprising computer program PROGR3 code for controlling the server. The device SERVER may be connected by a wired or wireless network connection, or both, to sources SRC1 and/or SRC2, as well as the viewer devices VIEWER1 and VIEWER2 over the communication interface COMM3.

[0092] For viewing and listening the captured or created video and audio content, there may be one or more reproduction devices REPROC1 and REPROC2. These devices may have a rendering module and a display and audio reproduction module, or these functionalities may be combined in a single device. The devices may comprise or be functionally connected to a computer processor PROC4 and memory MEM4, the memory comprising computer program PROG4 code for controlling the reproduction devices. The reproduction devices may consist of a video data stream receiver for receiving a video data stream and for decoding the video data stream, and an audio data stream receiver for receiving an audio data stream and for decoding the audio data stream. The video/audio data streams may be received from the server SERVER or from some other entity, such as a proxy server, an edge server of a content delivery network, or a file available locally in the viewer device. The data streams may be received over a network connection through communications interface COMM4, or from a memory device MEM6 like a memory card CARD2. The reproduction devices may have a graphics processing unit for processing of the data to a suitable format for viewing. The reproduction REPROC1 may comprise a high-resolution stereo-image head-mounted display for viewing the rendered stereo video sequence. The head-mounted display may have an orientation sensor DET1 and stereo audio headphones. The reproduction REPROC2 may comprise a display (either two-dimensional or a display enabled with 3D technology for displaying stereo video), and the rendering device may have an orientation detector DET2 connected to it. Alternatively, the reproduction REPROC2 may comprise a 2D display, since the volumetric video rendering can be done in 2D by rendering the viewpoint from a single eye instead of a stereo eye pair. The reproduction REPROC2 may comprise audio reproduction means, such as headphones or loudspeakers.

[0093] It needs to be understood that Fig. 14 depicts one SRC1 device and one SRC2 device, but generally the system may comprise more than one SRC1 device and/or SRC2 device.

[0094] The present embodiments relate to providing 2D visual and spatial audio in a 3D scene, such as in the system depicted in Fig. 14. In other words, the embodiments relate to consumption of volumetric or six-degrees-of-freedom (6DoF) audio in connection with 2D visual, and more generally to augmented reality (AR) or virtual reality (VR) or mixed reality (MR). AR/VR/MR is volumetric by nature, which means that the user is able to move around in the blend of physical and digital content, and digital content presentation is modified accordingly to user position & orientation.

[0095] It is expected that AR/VR/MR is likely to evolve in stages. Currently, most applications are implemented as 3DoF, which means that head rotation in three axes yaw/pitch/roll can be taken into account. This facilitates the audio-visual scene remaining static in a single location as the user rotates his head.

[0096] The next stage could be referred as 3DoF+ (or restricted/limited 6DoF), which will facilitate limited movement (translation, represented in Euclidean spaces as x, y, z). For example, the movement might be limited to a range of some tens of centimeters around a location.

[0097] The ultimate target is 6DoF volumetric virtual reality, where the user is able to freely move in a Euclidean space (x, y, z) and rotate his head (yaw, pitch, roll).

[0098] It is noted that the term "user movement" as used herein refers any user movement i.e. changes in (a) head orientation (yaw/pitch/roll) and (b) user position performed either by moving in the Euclidian space or by limited head movements. User can move by physically moving in the consumption space, while either sensors mounted in the environment track his location in outside-in fashion, or sensors co-located with the head-mounted-display (HMD) device track his location. Sensors co-located in a HMD or a mobile device mounted in an HMD can generally be either inertial sensors such as a gyroscope or image/vision based motion sensing devices.

[0099] Fig. 15 depicts example devices for implementing various embodiments. It is noted that capturing 2D visual content and related volumetric audio and rendering 2D visual content and related volumetric audio may be implemented either on the same device or different devices. The device performing the rendering may be referred to a rendering device 1502 and the device performing the capturing may be referred to a capturing device 1504. The capturing device may comprise at least a processor and a memory, and at least two microphones for capturing volumetric audio content and a camera for capturing 2D visual content, e.g. still images or video, operatively connected to the processor and memory. The capturing device may transmit at least part of the captured 2D visual content and related volumetric audio content to a rendering device, which renders the 2D visual content and related volumetric audio content for playback of an audio-visual scene in a presentation volume. Examples of the rendering device and capturing device comprise a user device for example a mobile device. According to at least some embodiments, the capturing device may generate information indicating spatial position and orientation of one or more visual objects of the captured 2D visual content and/or audio sources related to the visual objects. The information indicating spatial position and orientation of one or more visual objects of the captured 2D visual content and/or audio sources related to the visual objects may be included in metadata. The metadata may be transmitted to the rendering device together with the content, e.g. 2D visual content and/or related volumetric audio content, or the metadata may be transmitted to the rendering device separately. In one example the metadata may be transmitted to the rendering device on request by the rendering device. The metadata may be made available related to the captured the 2D visual content and related volumetric audio content on a server, where the rendering device may retrieve the metadata for rendering an audio-visual scene.

[0100] The devices shown in Figure 15 may operate according to the ISO/IEC JTC1/SC29/WG11 or MPEG (Moving Picture Experts Group) future standard called MPEG-I, which will facilitate rendering of audio for 3DoF, 3DoF+ and 6DoF scenarios. The technology will be based on 23008-3:201x, MPEG-H 3D Audio Second Edition. MPEG-H 3D audio is used for the core waveform carriage (encoding, decoding) in the form of objects, channels, and Higher-Order-Ambisonics (HOA). The goal of MPEG-I is to develop and standardize technologies comprising metadata over the core MPEG-H 3D and new rendering technologies to enable 3DoF, 3DoF+ and 6DoF audio transport and rendering.

[0101] The following describes in further detail suitable apparatus and possible mechanisms for implementing some embodiments. In this regard reference is first made to Figure 16 which shows a schematic block diagram of an exemplary apparatus or electronic device 50 depicted in Figure 17, which may incorporate a rendering device according to an embodiment.

[0102] The electronic device 50 may be a user device, for example a mobile terminal or user equipment of a wireless communication system. However, it would be appreciated that some embodiments may be implemented within any electronic device or apparatus which may require transmission of radio frequency signals.

[0103] The apparatus 50 may comprise a housing 30 for incorporating and protecting the device. The apparatus 50 further may comprise a display 32 in the form of a liquid crystal display. In other embodiments the display may be any suitable display technology suitable to display an image or video. The apparatus 50 may further comprise a keypad 34. In other embodiments any suitable data or user interface mechanism may be employed. For example the user interface may be implemented as a virtual keyboard or data entry system as part of a touch-sensitive display. The apparatus may comprise a microphone 36 or any suitable audio input which may be a digital or analogue signal input. The apparatus 50 may further comprise an audio output device which in some embodiments may be any one of: an earpiece 38, speaker, or an analogue audio or digital audio output connection. The apparatus 50 may also comprise a battery 40 (or in other embodiments the device may be powered by any suitable mobile energy device such as solar cell, fuel cell or clockwork generator). The term battery discussed in connection with the embodiments may also be one of these mobile energy devices. Further, the apparatus 50 may comprise a combination of different kinds of energy devices, for example a rechargeable battery and a solar cell. The apparatus may further comprise an infrared port 41 for short range line of sight communication to other devices. In other embodiments the apparatus 50 may further comprise any suitable short range communication solution such as for example a Bluetooth wireless connection or a USB/firewire wired connection.

[0104] The apparatus 50 may comprise a controller 56 or processor for controlling the apparatus 50. The controller 56 may be connected to memory 58 which in some embodiments may store both data and/or may also store instructions for implementation on the controller 56. The controller 56 may further be connected to codec circuitry 54 suitable for carrying out coding and decoding of audio and/or video data or assisting in coding and decoding carried out by the controller 56.

[0105] The apparatus 50 may further comprise a card reader 48 and a smart card 46, for example a universal integrated circuit card (UICC) reader and UICC for providing user information and being suitable for providing authentication information for authentication and authorization of the user at a network.

[0106] The apparatus 50 may comprise radio interface circuitry 52 connected to the controller and suitable for generating wireless communication signals for example for communication with a cellular communications network, a wireless communications system or a wireless local area network. The apparatus 50 may further comprise an antenna 59 connected to the radio interface circuitry 52 for transmitting radio frequency signals generated at the radio interface circuitry 52 to other apparatus(es) and for receiving radio frequency signals from other apparatus(es).

[0107] In some embodiments, the apparatus 50 comprises a camera 42 capable of recording or detecting imaging.

[0108] With respect to Fig. 18, an example of a system within which embodiments of the present invention can be utilized is shown. The system 10 comprises multiple communication devices which can communicate through one or more networks. The system 10 may comprise any combination of wired and/or wireless networks including, but not limited to a wireless cellular telephone network (such as a GSM (2G, 3G, 4G, LTE, 5G), UMTS, CDMA network etc.), a wireless local area network (WLAN) such as defined by any of the IEEE 802.x standards, a Bluetooth personal area network, an Ethernet local area network, a token ring local area network, a wide area network, and the Internet.

[0109] For example, the system shown in Fig. 18 shows a mobile telephone network 11 and a representation of the internet 28. Connectivity to the internet 28 may include, but is not limited to, long range wireless connections, short range wireless connections, and various wired connections including, but not limited to, telephone lines, cable lines, power lines, and similar communication pathways.

[0110] The example communication devices shown in the system 10 may include, but are not limited to, an electronic device or apparatus 50, a combination of a personal digital assistant (PDA) and a mobile telephone 14, a PDA 16, an integrated messaging device (IMD) 18, a desktop computer 20, a notebook computer 22, a tablet computer. The apparatus 50 may be stationary or mobile when carried by an individual who is moving. The apparatus 50 may also be located in a mode of transport including, but not limited to, a car, a truck, a taxi, a bus, a train, a boat, an airplane, a bicycle, a motorcycle or any similar suitable mode of transport.

[0111] Some or further apparatus may send and receive calls and messages and communicate with service providers through a wireless connection 25 to a base station 24. The base station 24 may be connected to a network server 26 that allows communication between the mobile telephone network 11 and the internet 28. The system may include additional communication devices and communication devices of various types.

[0112] The communication devices may communicate using various transmission technologies including, but not limited to, code division multiple access (CDMA), global systems for mobile communications (GSM), universal mobile telecommunications system (UMTS), time divisional multiple access (TDMA), frequency division multiple access (FDMA), transmission control protocol-internet protocol (TCP-IP), short messaging service (SMS), multimedia messaging service (MMS), email, instant messaging service (IMS), Bluetooth, IEEE 802.11, Long Term Evolution wireless communication technique (LTE) and any similar wireless communication technology. Yet some other possible transmission technologies to be mentioned here are high-speed downlink packet access (HSDPA), high-speed uplink packet access (HSUPA), LTE Advanced (LTE-A) carrier aggregation dual-carrier, and all multi-carrier technologies. A communications device involved in implementing various embodiments of the present invention may communicate using various media including, but not limited to, radio, infrared, laser, cable connections, and any suitable connection. In the following some example implementations of apparatuses utilizing the present invention will be described in more detail.

[0113] According to an embodiment, an apparatus comprises means for determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content, means for aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume, and means for rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0114] In an alternative embodiment, an apparatus comprises at least one processor; and at least one memory including computer program code; the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus at least to perform: determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content;

[0115] aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume;

[0116] rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0117] In an embodiment there is provided a computer program comprising computer readable program code means adapted to perform at least the following: determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume; rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of

the at least one object of interest and the related audio source.

[0118] In an embodiment there is provided a non-transitory computer readable medium comprising program instructions stored thereon for performing at least the following: determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content; aligning a spatial position and orientation of the at least one object of interest with a spatial position and orientation of the related audio source in a presentation volume; rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and orientations of the at least one object of interest and the related audio source.

[0119] A memory may be a computer readable medium that may be non-transitory. The memory may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor-based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The data processors may be of any type suitable to the local technical environment, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multi-core processor architecture, as non-limiting examples.

[0120] Embodiments may be implemented in software, hardware, application logic or a combination of software, hardware and application logic. The software, application logic and/or hardware may reside on memory, or any computer media. In an example embodiment, the application logic, software or an instruction set is maintained on any one of various conventional computer-readable media. In the context of this document, a "memory" or "computer-readable medium" may be any media or means that can contain, store, communicate, propagate or transport the instructions for use by or in connection with an instruction execution system, apparatus, or device, such as a computer.

[0121] Reference to, where relevant, "computer-readable storage medium", "computer program product", "tangibly embodied computer program" etc., or a "processor" or "processing circuitry" etc. should be understood to encompass not only computers having differing architectures such as single/multi-processor architectures and sequencers/parallel architectures, but also specialised circuits such as field programmable gate arrays FPGA, application specific circuits ASIC, signal processing devices and other devices. References to computer readable program code means, computer program, computer instructions, computer code etc. should be understood to express software for a programmable processor firmware such as the programmable content of a hardware device as instructions for a processor or configured or configuration settings for a fixed function device, gate array, programmable logic device, etc.

[0122] In general, the various embodiments may be implemented in hardware or special purpose circuits or any combination thereof. While various aspects may be illustrated and described as block diagrams or using some other pictorial representation, it is well understood that these blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

[0123] Embodiments may be practiced in various components such as integrated circuit modules. The design of integrated circuits is by and large a highly automated process. Complex and powerful software tools are available for converting a logic level design into a semiconductor circuit design ready to be etched and formed on a semiconductor substrate.

[0124] Programs, such as those provided by Synopsys, Inc. of Mountain View, California and Cadence Design, of San Jose, California automatically route conductors and locate components on a semiconductor chip using well established rules of design as well as libraries of pre stored design modules. Once the design for a semiconductor circuit has been completed, the resultant design, in a standardized electronic format (e.g., Opus, GDSII, or the like) may be transmitted to a semiconductor fabrication facility or "fab" for fabrication.

[0125] As used in this application, the term "circuitry" may refer to one or more or all of the following:

- (a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry) and
- (b) combinations of hardware circuits and software, such as (as applicable):

- (i) a combination of analog and/or digital hardware circuit(s) with software/firmware and
- (ii) any portions of hardware processor(s) with software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile phone or server, to perform various functions) and

- (c) hardware circuit(s) and or processor(s), such as a microprocessor(s) or a portion of a microprocessors), that requires software (e.g., firmware) for operation, but the software may not be present when it is not needed for operation.

[0126] This definition of circuitry applies to all uses of this term in this application, including in any claims. As a further example, as used in this application, the term circuitry also covers an implementation of merely a hardware circuit or processor (or multiple processors) or portion of a hardware circuit or processor and its (or their) accompanying software

and/or firmware. The term circuitry also covers, for example and if applicable to the particular claim element, a baseband integrated circuit or processor integrated circuit for a mobile device or a similar integrated circuit in server, a cellular network device, or other computing or network device.

[0127] Although the above examples describe some embodiments operating within a user device, it would be appreciated that embodiments as described above may be implemented as a part of any apparatus comprising a circuitry in which 2D visual content and volumetric audio may be rendered. Thus, for example, embodiments may be implemented in a mobile phone, in a computer such as a desktop computer or a tablet computer.

[0128] The foregoing description has provided by way of exemplary and non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those skilled in the relevant arts in view of the foregoing description, when read in conjunction with the accompanying drawings and the appended claims. However, all such and similar modifications of the teachings of this invention will still fall within the scope of this invention.

Claims

1. A method comprising:

determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content;
aligning a spatial position and/or orientation of the at least one object of interest with a spatial position and/or orientation of the related audio source in a presentation volume;
rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and/or orientations of the at least one object of interest and the related audio source.

2. The method according to claim 1, comprising:

obtaining information indicating spatial positions and/or orientations of one or more visual objects of the 2D visual content and related audio sources of the volumetric audio content; and aligning the spatial position and/or orientation of the at least one object of interest with the spatial position and/or orientation of the related audio source in the presentation volume on the basis of the obtained information.

3. The method according to claim 2, wherein the information indicating spatial position and/or orientation is included in metadata of at least one of the 2D visual content and the volumetric audio content.

4. The method according to any of claims 1 to 2, comprising:

determining the at least one object of interest of the 2D visual content on the basis of:

a user input; or
metadata associated with the 2D visual content.

5. An apparatus comprising:

means for determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content;
means for aligning a spatial position and/or orientation of the at least one object of interest with a spatial position and/or orientation of the related audio source in a presentation volume; and
means for rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and/or orientations of the at least one object of interest and the related audio source.

6. The apparatus according to claim 5, comprising:

means for obtaining information indicating spatial positions and/or orientations of one or more visual objects of the 2D visual content and related audio sources of the volumetric audio content; and
means for aligning the spatial position and/or orientation of the at least one object of interest with the spatial position and/or orientation of the related audio source in the presentation volume on the basis of the obtained information.

7. The apparatus according to claim 6, wherein the information indicating spatial position and/or orientation is included

in metadata of at least one of the 2D visual content and the volumetric audio content.

8. The apparatus according to any of claims 5 to 6, comprising:
means for determining the at least one object of interest of the 2D visual content on the basis of:

a user input; or
metadata associated with the 2D visual content.

9. The apparatus according to any of claims 5 to 8, comprising:
means for, in response to determining a misalignment of the at least one object of interest with the related audio source, re-aligning the spatial position and/or orientation of the at least one object of interest with the spatial position and/or orientation of the related audio source.

10. The apparatus according to claim 9,
wherein the means for re-aligning comprises at least one of:

- means for adapting a visual zoom of the 2D visual content in the presentation volume;
- means for moving a visual rendering plane of the 2D visual content in the presentation volume; and
- means for adapting orientation of the 2D visual content in the presentation volume.

11. The apparatus according to any of claims 5 to 10, comprising:

means for obtaining information indicating a permissible zooming of the 2D visual content; and
means for zooming the 2D visual content within the indicated permissible zooming for aligning the at least one object of interest with the related audio source.

12. The apparatus according to any of claims 5 to 11, comprising:

means for determining an initial position of the 2D visual content in the presentation volume; and
means for, in response to a user input indicating a subsequent position of the 2D visual content, re-aligning the spatial position and/or orientation of the at least one object of interest with the spatial position and/or orientation of the related audio source in the presentation volume, and
means for rendering the 2D visual content and the volumetric audio content using the re-aligned spatial position and/or orientation.

13. The apparatus according to any of claims 5 to 12, wherein the audio source comprises at least two audio sources and the apparatus further comprises:

means for modifying the volumetric audio content for reducing depth differences between audio sources; and
means for rendering the 2D visual content and the volumetric audio content using the modified volumetric audio content.

14. The apparatus according to any of claims 5 to 13, wherein the 2D visual content is rendered as world locked content in the presentation volume.

15. A computer program comprising computer program instructions that, when executed by processing circuitry, cause:

determining at least one object of interest of two-dimensional, 2D, visual content related to an audio source of volumetric audio content;
aligning a spatial position and/or orientation of the at least one object of interest with a spatial position and/or orientation of the related audio source in a presentation volume; and
rendering the 2D visual content and the volumetric audio content on the basis of the aligned spatial position and/or orientations of the at least one object of interest and the related audio source.

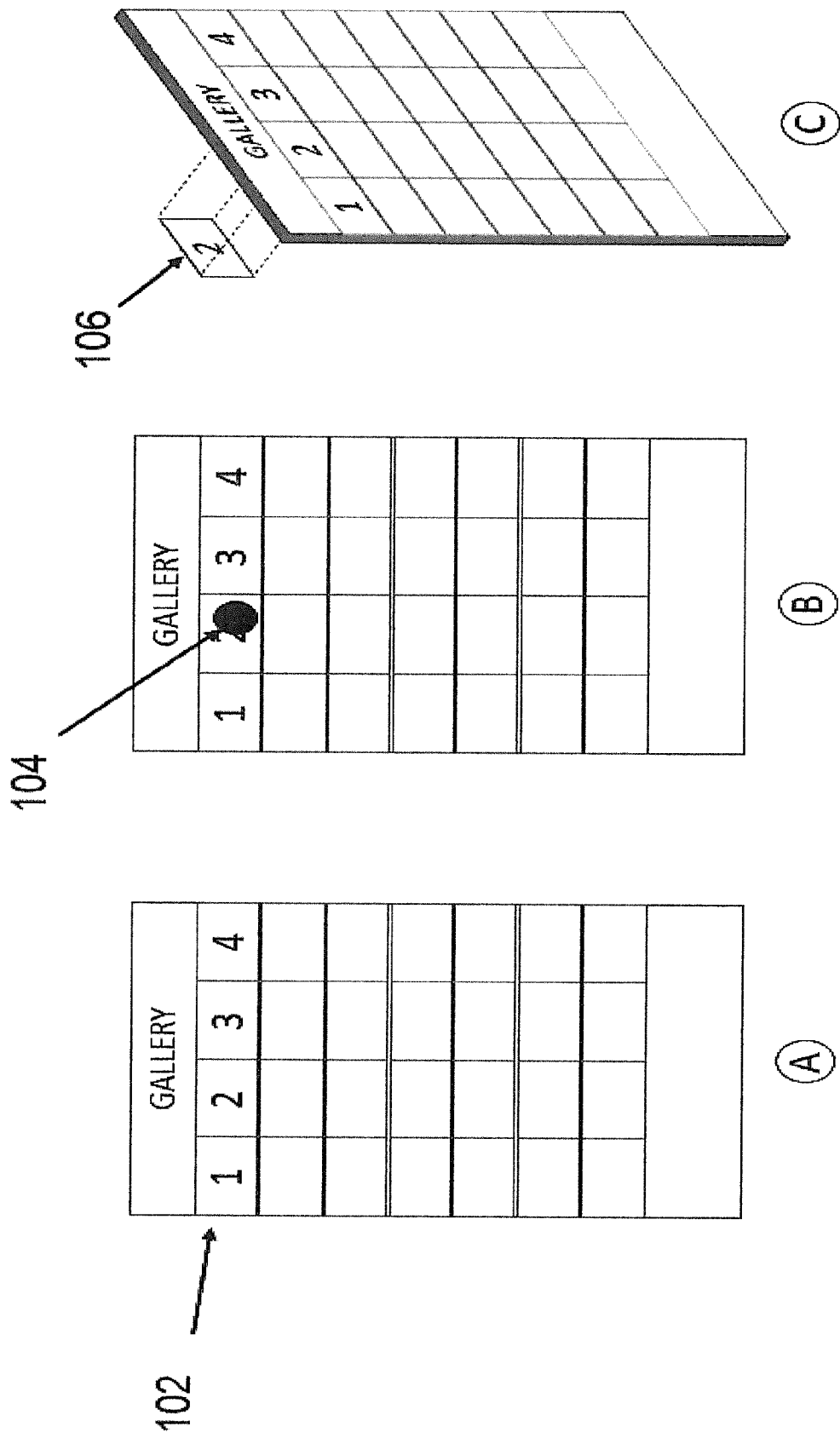


Fig. 1

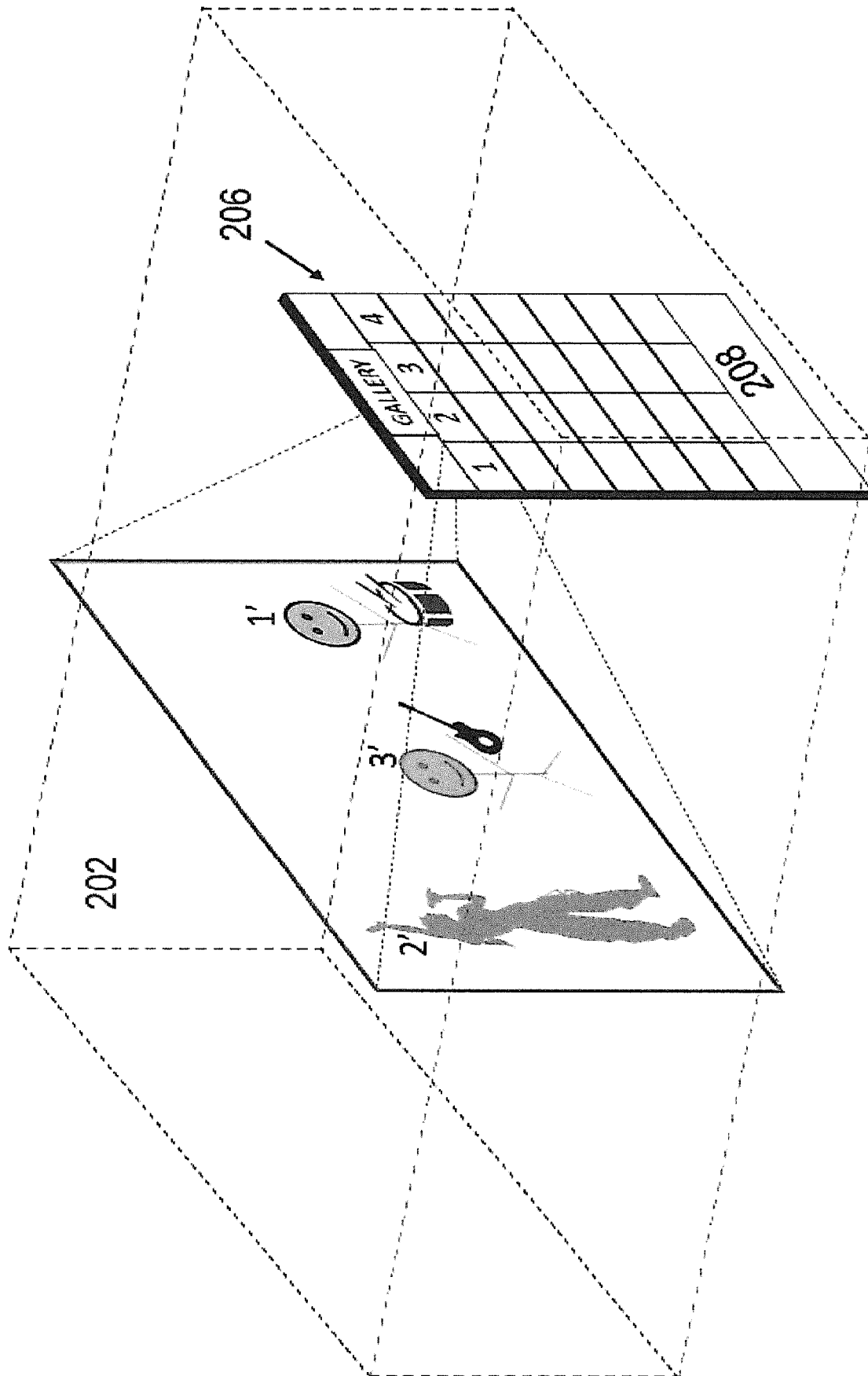


Fig. 2

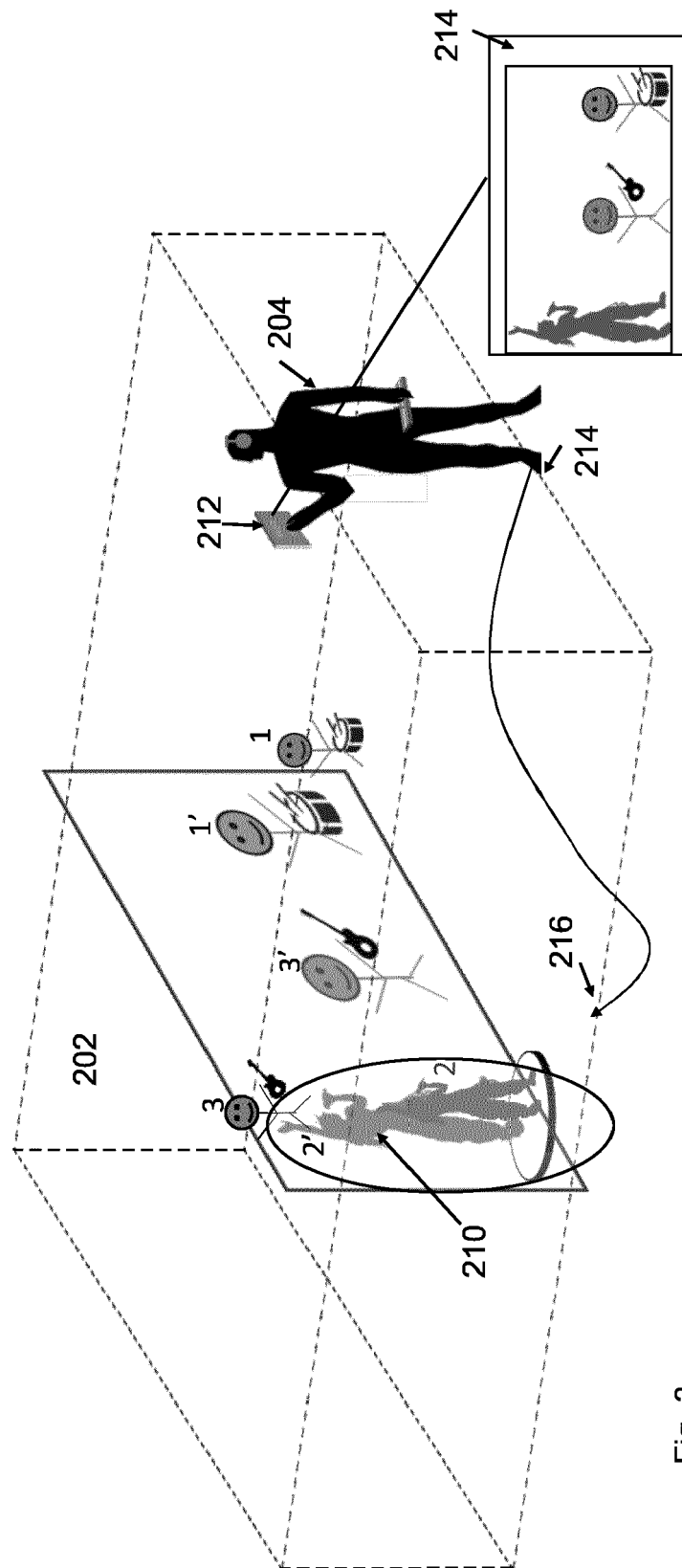
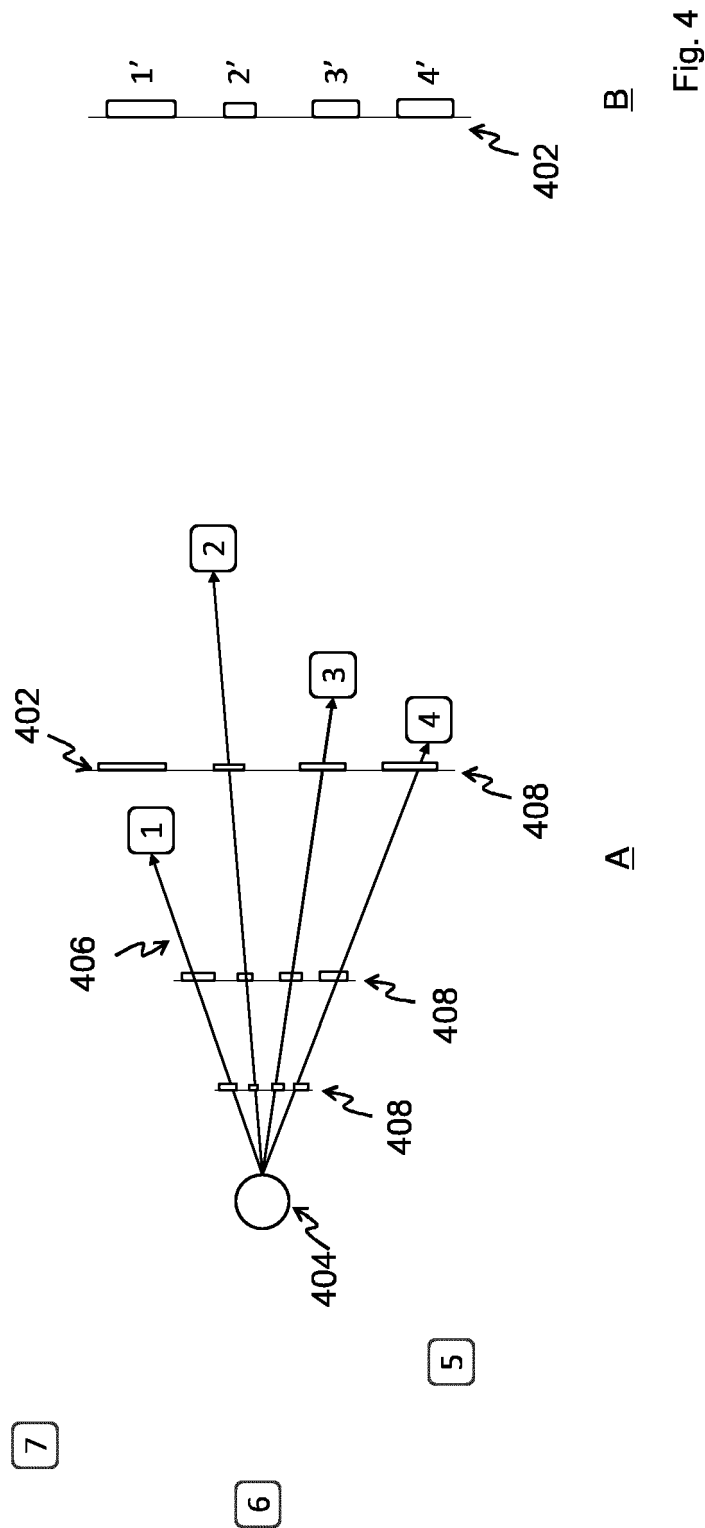


Fig. 3



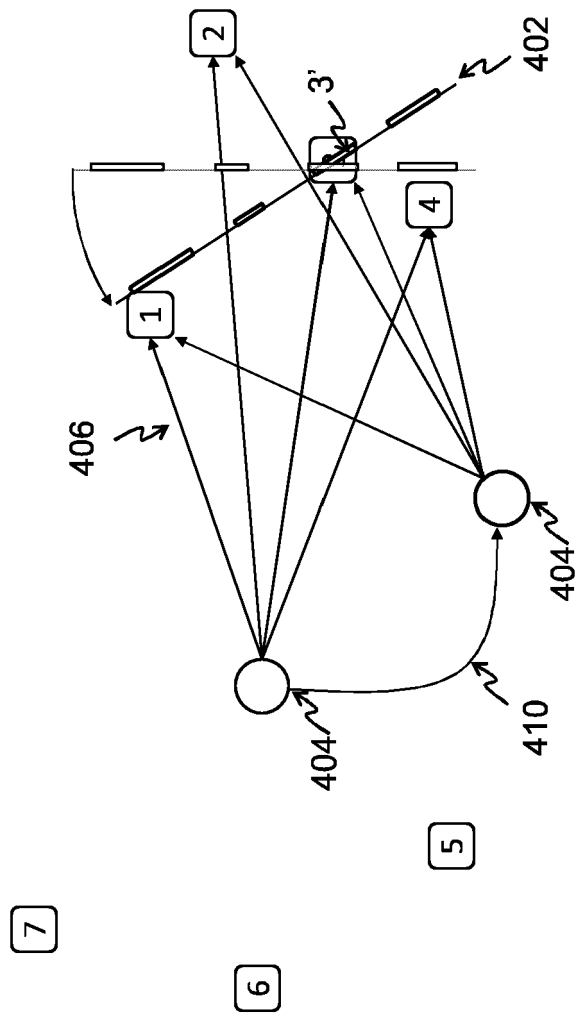


Fig. 5

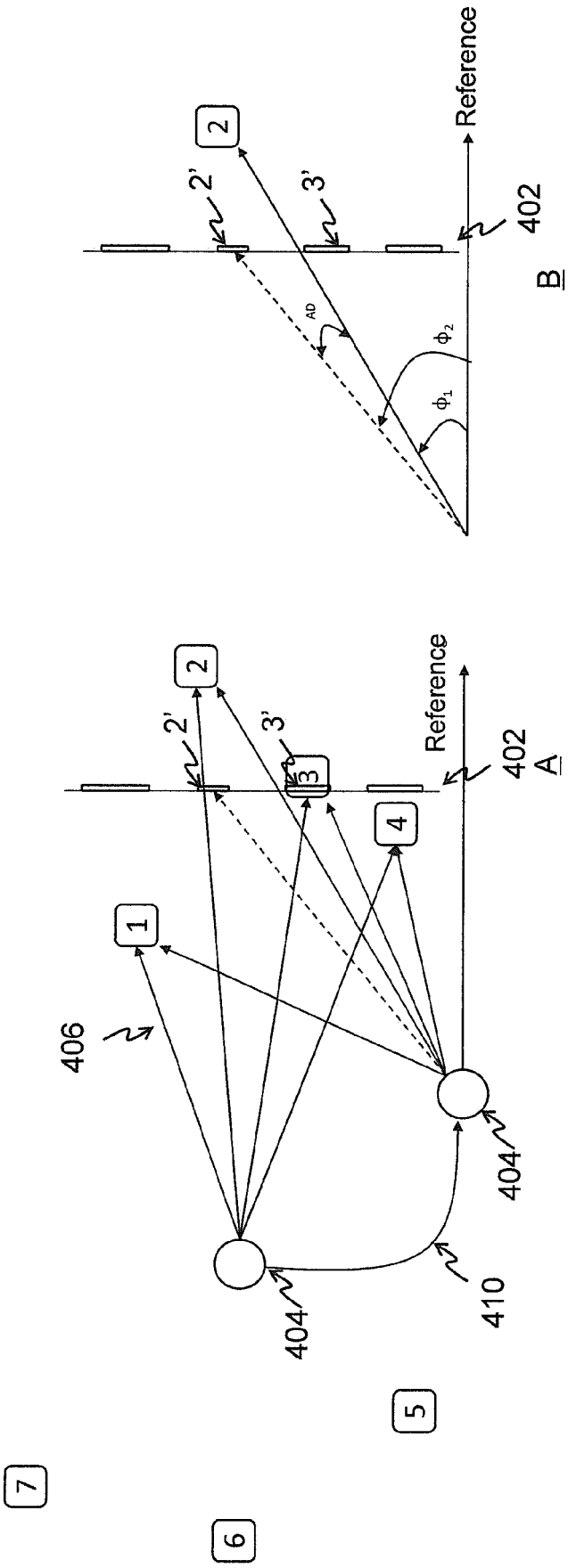


Fig. 6

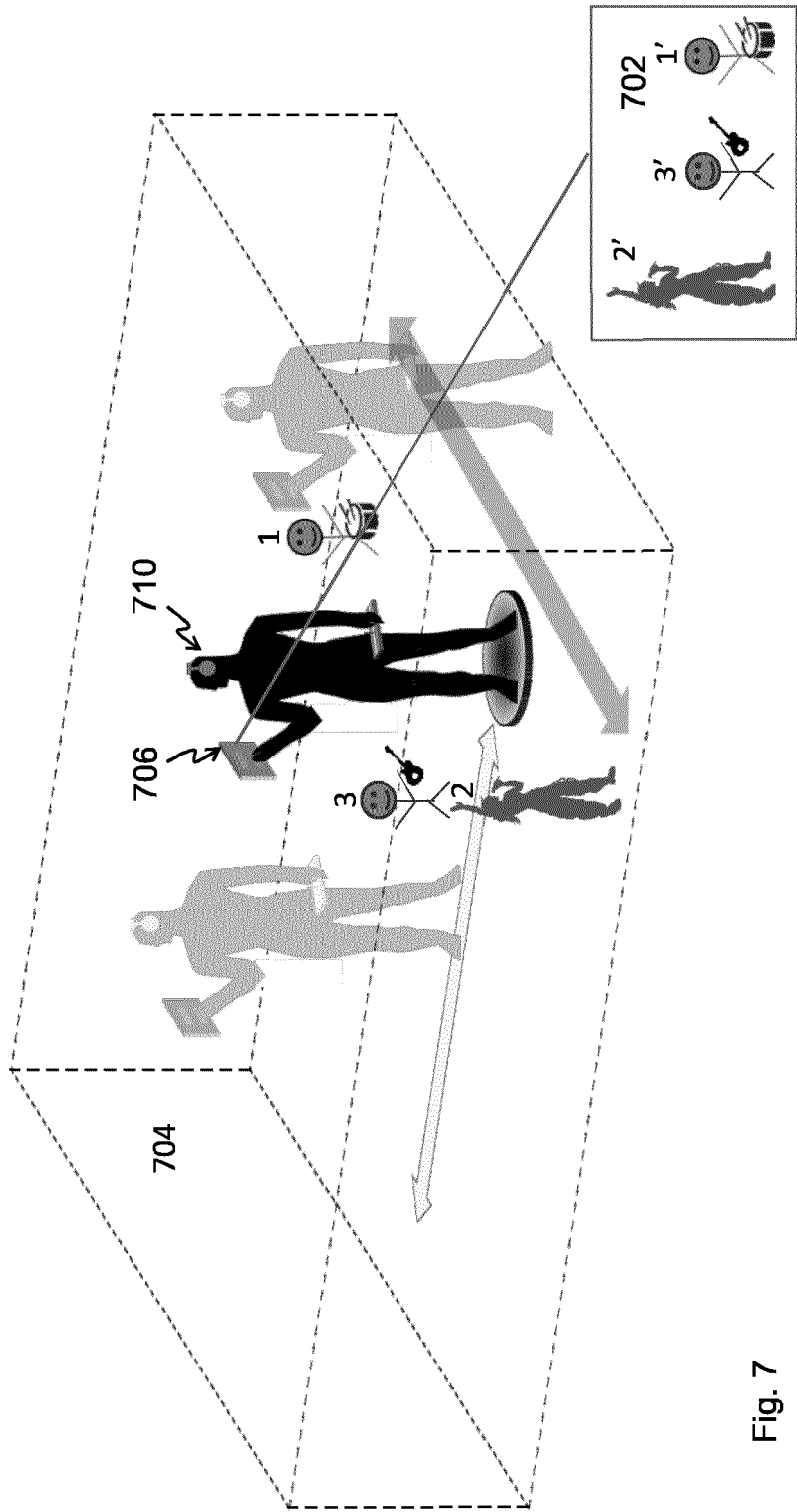


Fig. 7

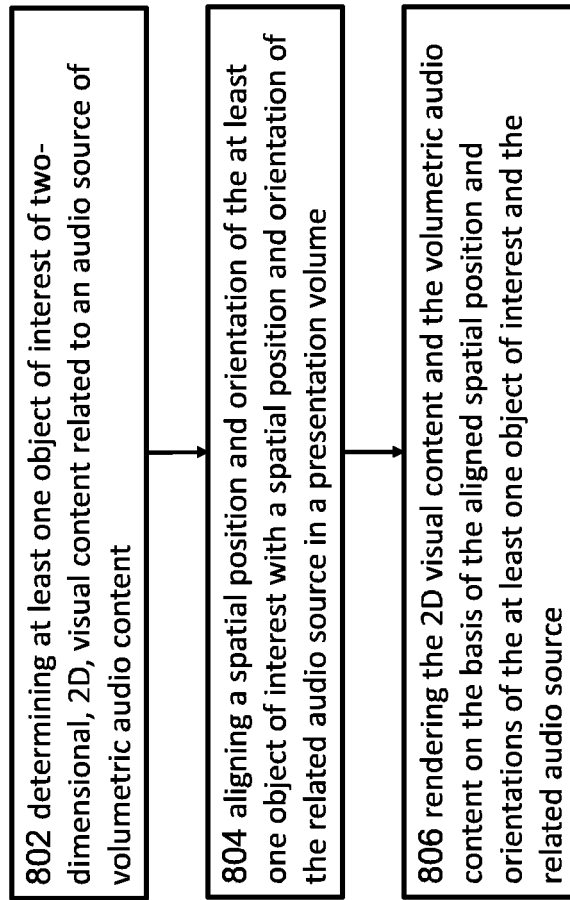


Fig. 8

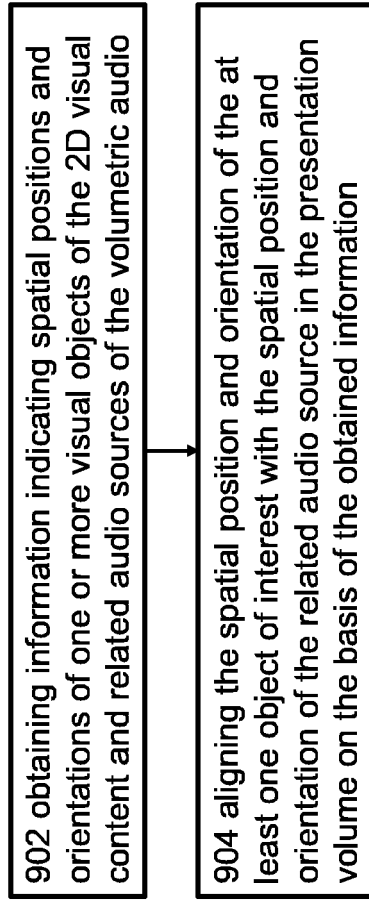


Fig. 9

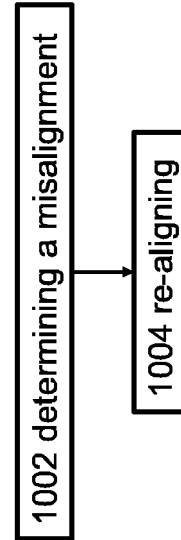


Fig. 10

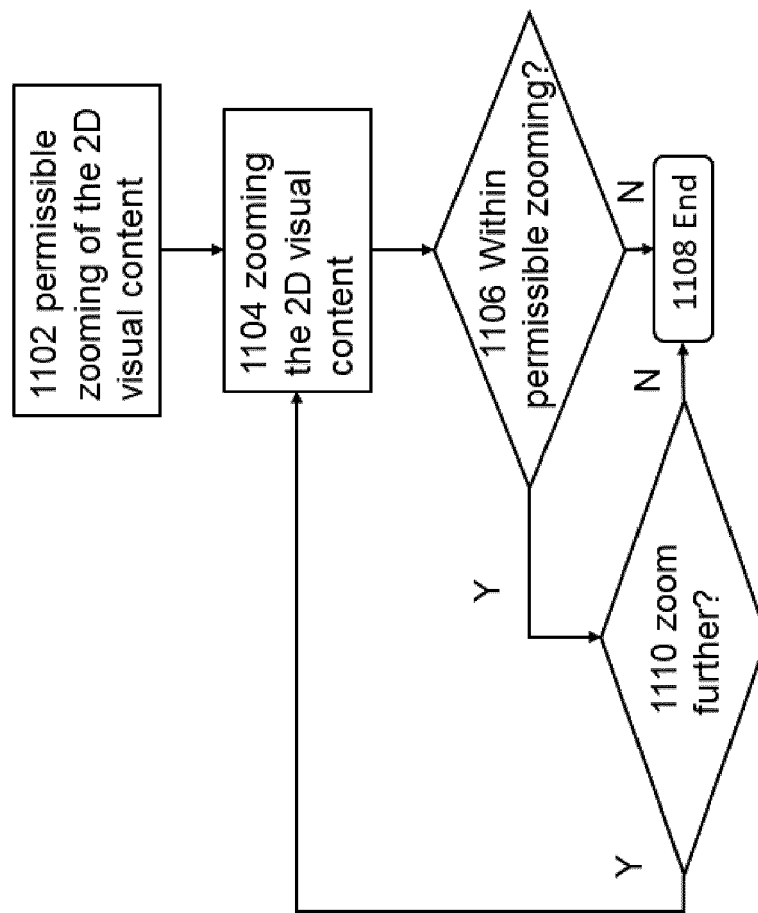


Fig. 11

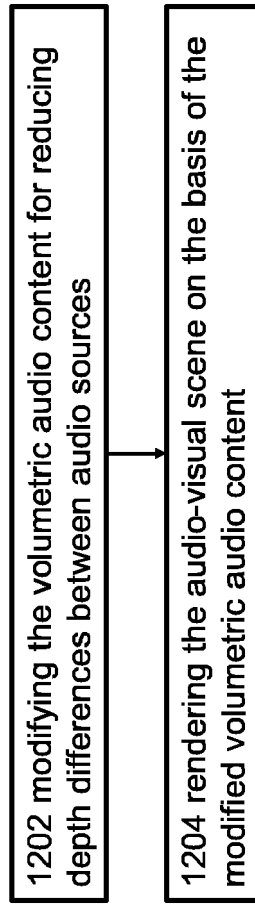


Fig. 12

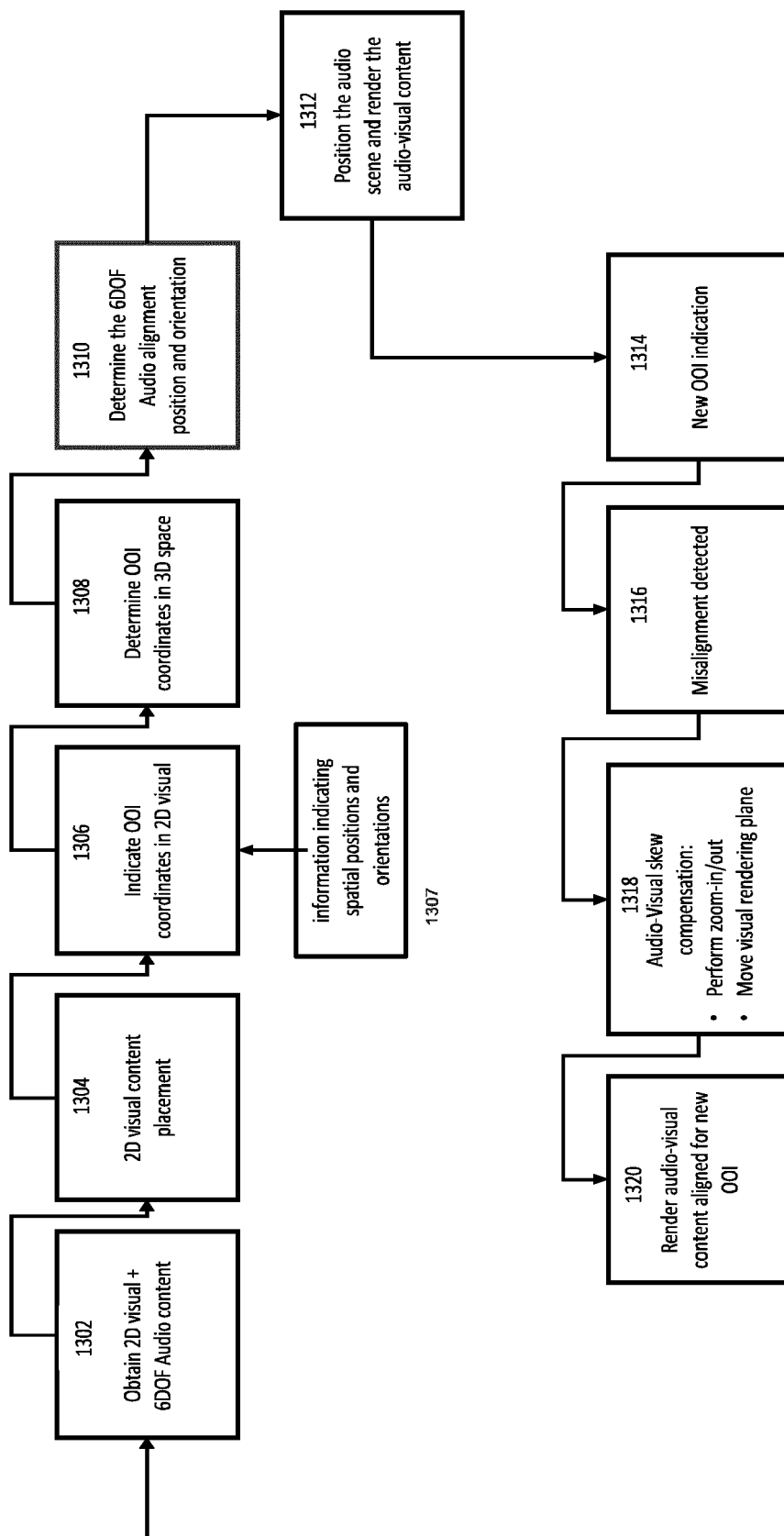


Fig. 13

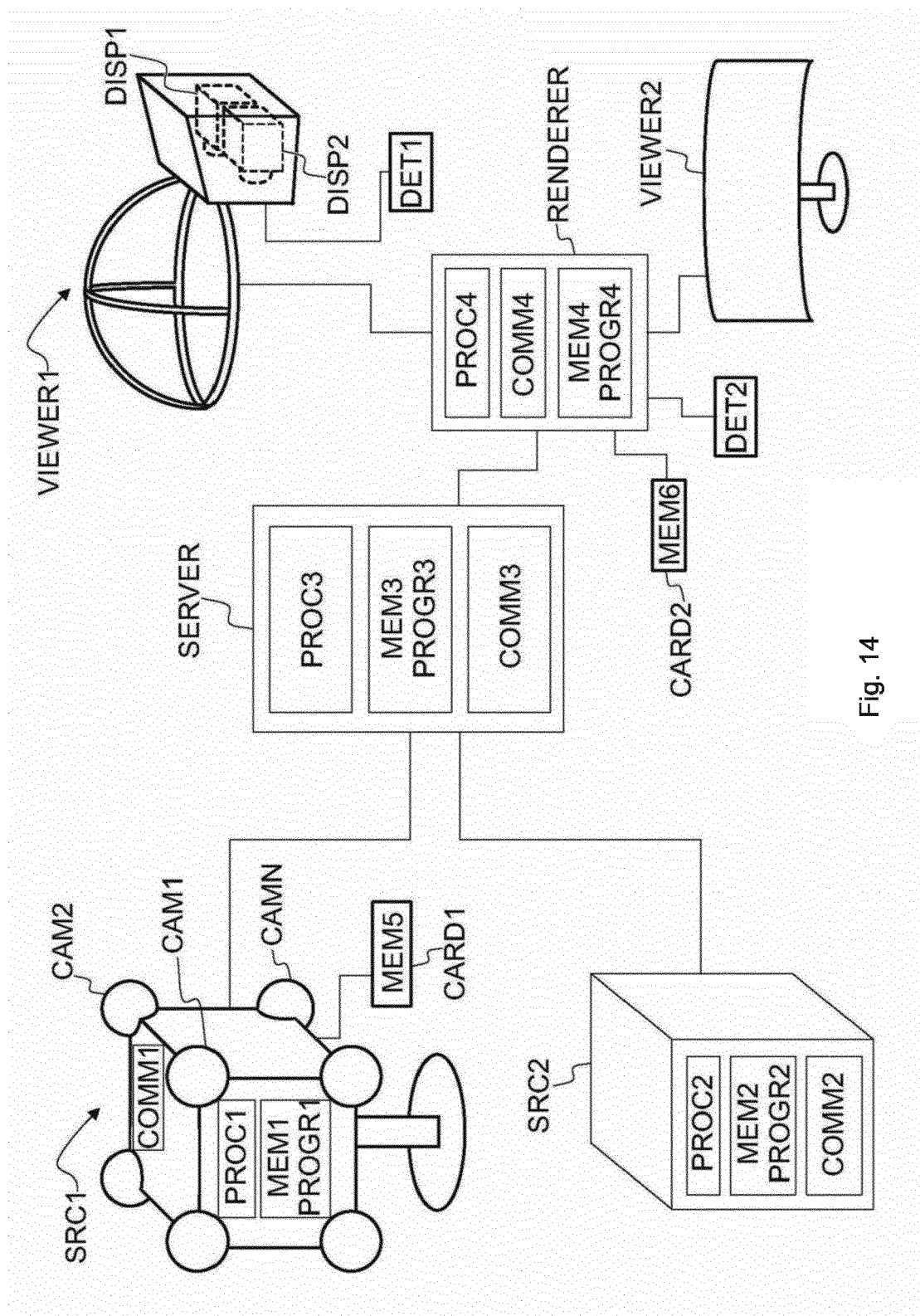


Fig. 14

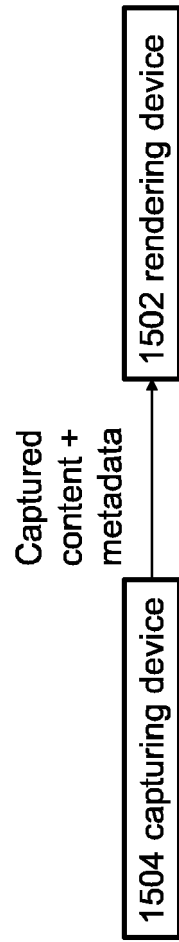


Fig. 15

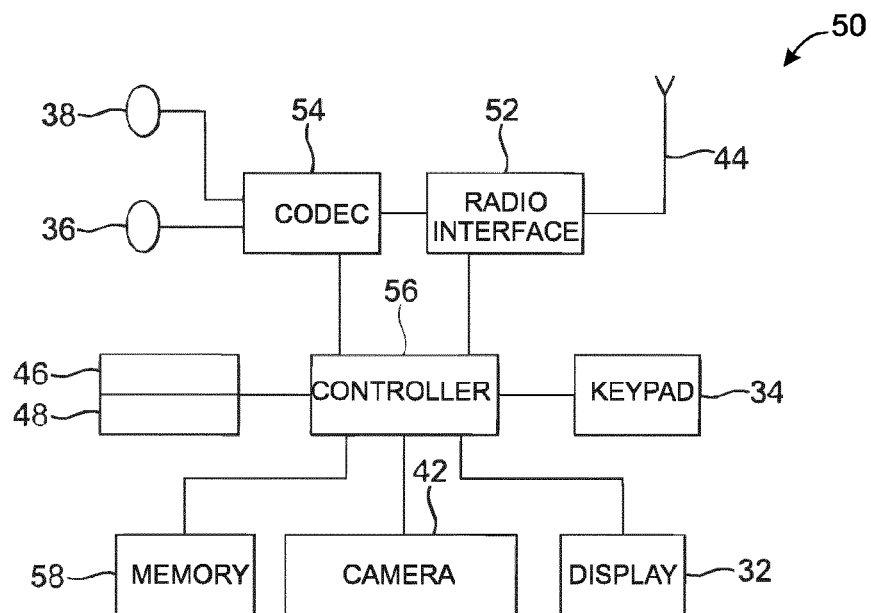


Fig. 16

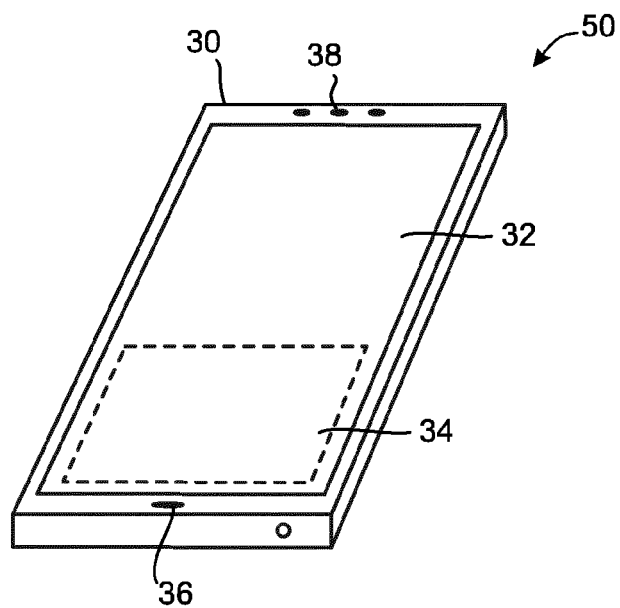


Fig. 17

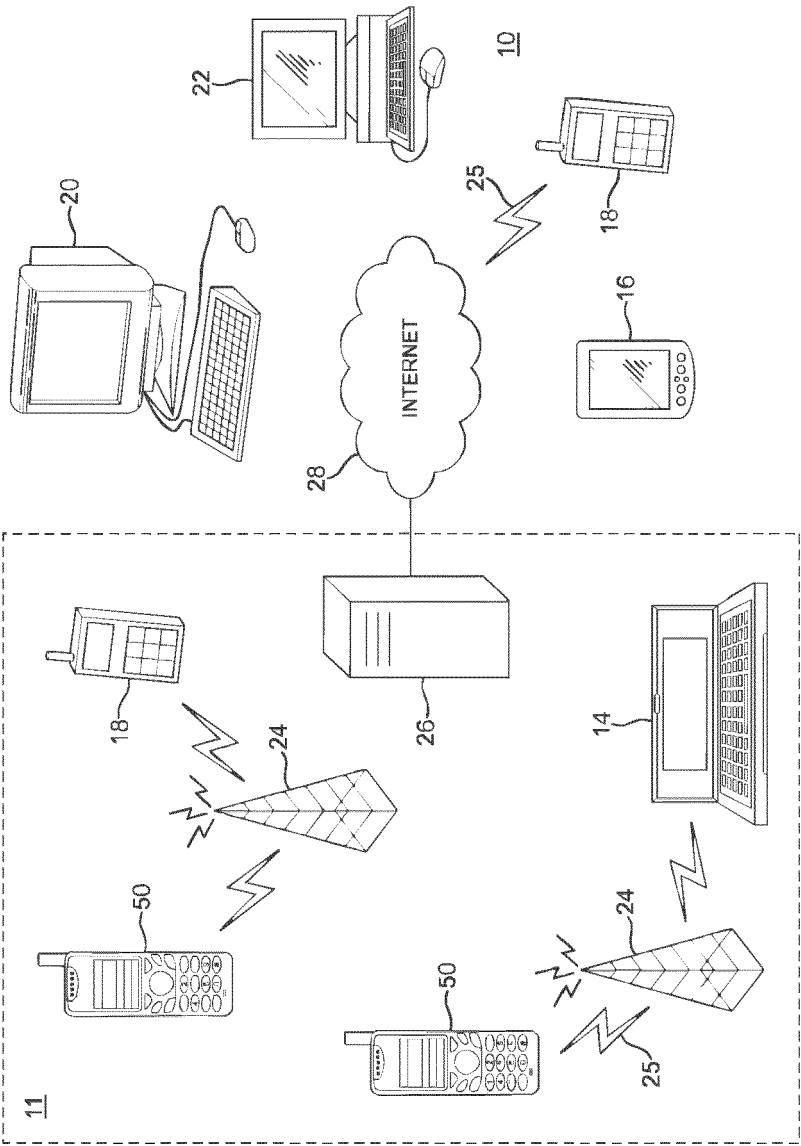


Fig. 18



EUROPEAN SEARCH REPORT

Application Number
EP 20 19 1021

5

10

15

20

25

30

35

40

45

50

55

1

EPO FORM 1503 03.82 (P04C01)

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	WO 2019/121864 A1 (KONINKLIJKE KPN NV [NL]; TNO [NL]) 27 June 2019 (2019-06-27) * figures 1-4, 6 * * page 2, lines 10-26 * * page 6, lines 36-39 * * page 12, lines 12-37 * * page 13, lines 31-39 * * page 14, line 38 - page 15, line 13 * * page 16, lines 24-30 *	1-15	INV. H04S7/00
A	WO 2018/211166 A1 (NOKIA TECHNOLOGIES OY [FI]) 22 November 2018 (2018-11-22) * figure 9 * * page 4, line 35 - page 5, line 6 * * page 8, line 21 - page 9, line 3 * * page 9, lines 5-15 *	10-13	
A	WO 2019/141900 A1 (NOKIA TECHNOLOGIES OY [FI]) 25 July 2019 (2019-07-25) * figures 2A, 2B, 7B * * page 23, lines 1-9 *	10,11,14	
			TECHNICAL FIELDS SEARCHED (IPC)
			H04S
The present search report has been drawn up for all claims			
Place of search The Hague		Date of completion of the search 27 November 2020	Examiner Mendoza Lopez, Jorge
CATEGORY OF CITED DOCUMENTS X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document		T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons & : member of the same patent family, corresponding document	

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 20 19 1021

5

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

27-11-2020

10

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2019121864 A1	27-06-2019	CN 111466124 A	28-07-2020
		EP 3729829 A1	28-10-2020
		WO 2019121864 A1	27-06-2019

WO 2018211166 A1	22-11-2018	EP 3625977 A1	25-03-2020
		US 2018338213 A1	22-11-2018
		WO 2018211166 A1	22-11-2018

WO 2019141900 A1	25-07-2019	CN 111630879 A	04-09-2020
		EP 3741139 A1	25-11-2020
		US 2020344563 A1	29-10-2020
		WO 2019141900 A1	25-07-2019

15

20

25

30

35

40

45

50

55

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82