



(43) International Publication Date
23 February 2017 (23.02.2017)

- (51) International Patent Classification:
H04R 3/00 (2006.01) *G10L 21/0216* (2013.01)
H04S 7/00 (2006.01)
- (21) International Application Number: PCT/FI2016/050560
- (22) International Filing Date: 10 August 2016 (10.08.2016)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 15181162.7 14 August 2015 (14.08.2015) EP
- (71) Applicant: NOKIA TECHNOLOGIES OY [FI/FI];
Karaportti 3, 02610 Espoo (FI).
- (72) Inventors: LEHTINIEMI, Arto; Lomarantatie 17, 33880
Lempäälä (FI). LEPPÄNEN, Jussi; Mielikinkatu 26,
33580 Tampere (FI). ERONEN, Antti; Rintamäenkatu 13
A 6, 33820 Tampere (FI).
- (74) Agents: NOKIA TECHNOLOGIES OY et al.; Ari
Aarnio, IPR Department, Karakaari 7, 02610 Espoo (FI).
- (81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,
DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR,
KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,
SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN,
TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ,
TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU,
TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,
DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,
LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,
SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,
GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— with international search report (Art. 21(3))

(54) Title: MONITORING

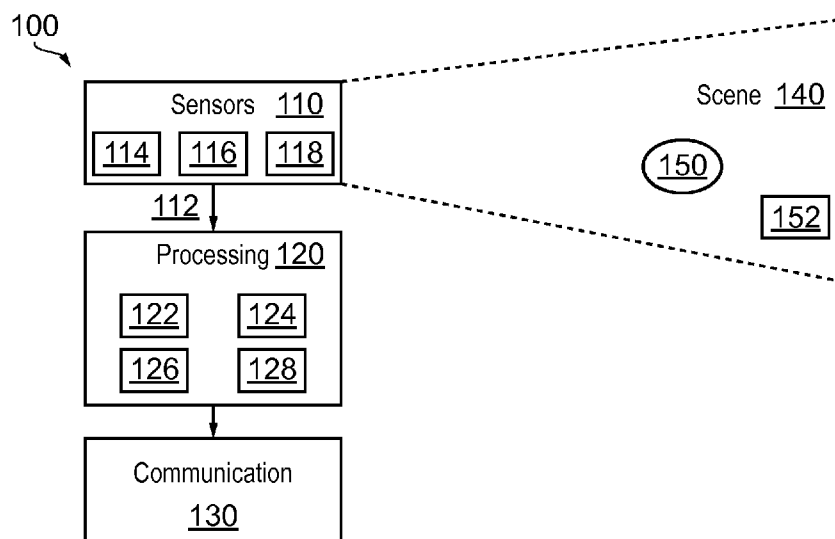


FIG. 1

(57) Abstract: A method comprising: processing received image data and/or received depth data to track an object moving in a monitored scene space; determining a bearing of the tracked moving object in the monitored scene space; and adaptively filtering outputs from a plurality of spatially diverse microphones in dependence upon the bearing of the tracked moving object in the monitored scene space to produce a combined audio signal.



1

TITLE

Monitoring

TECHNOLOGICAL FIELD

5

Embodiments of the present invention relate to monitoring a scene. In particular, they relate to automated monitoring of a scene.

BACKGROUND

10

Current monitoring systems may comprise one or more sensors connected via wires or wirelessly to a hub that stores or enables storage of the data recorded from the scene.

15 BRIEF SUMMARY

According to various, but not necessarily all, examples of the disclosure there is provided a method as claimed in any of claims 1-13.

20 According to various, but not necessarily all, examples of the disclosure there is provided an apparatus as claimed in claim 14.

According to various, but not necessarily all, examples of the disclosure there is provided an apparatus comprising: at least one processor; and

25 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 a method as claimed in any of claims 1-13.

30 According to various, but not necessarily all, examples of the disclosure there is provided computer program code that, when performed by at least one processor, causes a method of at least one of claims 1-13 to be performed.

2

According to various, but not necessarily all, examples of the disclosure there is provided a computer program that, when run on a computer, performs: a method as claimed in any of claims 1-13.

- 5 According to various, but not necessarily all, embodiments of the invention there is provided examples as claimed in the appended claims.

BRIEF DESCRIPTION

- 10 For a better understanding of various examples that are useful for understanding the brief description, reference will now be made by way of example only to the accompanying drawings in which:

- Fig 1 schematically illustrates an example of a system;
Fig 2 illustrates an example of a state machine;
15 Fig 3 illustrates an example of a processing module;
Fig 4 illustrates an example of a delivery mechanism for a computer program;
Fig 5 illustrates an example of operation of the system;
Fig 6 illustrates an example of a method that may be performed by the system;
Fig 7 illustrates an example of a bearing of a tracked object;
20 Fig 8 illustrates an example of a system for adaptively filtering outputs from microphones in dependence upon a bearing of a tracked moving object;
Figs 9A-9E illustrate an example of using the system to adaptively filter outputs from microphones in dependence upon a bearing of a tracked moving object; and
Figs 10A-10B illustrate an example of using the system to adaptively filter outputs from
25 microphones in dependence upon a bearing of a tracked moving object.

DETAILED DESCRIPTION

- The system 100 described is a system that monitors at least one scene and adaptively
30 filters outputs from a plurality of spatially diverse microphones in dependence upon a bearing of a tracked moving object in the scene to produce a combined audio signal.

- Fig 1 schematically illustrates a system 100 comprising: one or more sensors 110
configured to record sensor data 112 from a scene 140; a processing module 120
35 configured to process the sensor data 112 recorded from the scene 140 to recognise

3

automatically events that occur in the scene 140 and to automatically take decisions as a consequence of the recognition; and a communication module 130 configured to communicate, when a decision to communicate is taken by the processing module 120.

5

Some but not necessarily all of the events that are recognised may relate to an object 152 or a particular object 152 in the scene 140. An object may be an inanimate object, an animal, a person or a particular person 150 in the scene 140. In some but not necessarily all examples of the system 100, the system 100 is a real-time recognition system and the automatic recognition and decision, if any, occur substantially at the same time as the sensor data 112 is recorded.

10

The system 100 may be used for surveillance, monitoring or other applications. The system 100 may be used in businesses, public areas or at home.

15

The system 100 may enable a user to control monitoring, for example, when the user is a person 150 in the scene 140. The system 100 may enable a user to control recognition and/or a consequence of recognition, for example, when the user is a person 150 in the scene 140.

20

In some but not necessarily all examples, the system 100 may comprise a number of discrete apparatus. For example, the sensors 110 may be housed in one or more separate apparatus, the processing module 120 may be housed in one or more apparatus and the communication module 130 may be housed in one or more apparatus. Where a component or components of the system 100 are housed in discrete apparatus, those apparatus may be local or remote to each other and, where they are remote they may communicate, for example, over a network.

25

In some but not necessarily all examples, the system 100 may be housed in a single apparatus.

30

The sensors 110 are configured to record or enable recording of sensor data 112 from the scene 140.

A scene 140 may comprise static components that have, for example, fixed relative positions, such as for example static objects. These static objects have different static positions in the scene's three-dimensional space (scene space). A scene 140 may comprise moving components, such as for example a moving object. A moving object
5 has different positions in the scene space over time. Reference to 'the scene' in relation to different sensor data 112 or different times implies a continuity of those static components of the scene 140 that are sensed, it does not necessarily imply a continuity of some or all of the dynamic components although this may occur.

10 The recording of sensor data 112 may comprise only temporary recording, or it may comprise permanent recording or it may comprise both temporary recording and permanent recording. Temporary recording implies the recording of data temporarily. This may, for example, occur during sensing, occur at a dynamic memory, occur at a buffer such as a circular buffer, a register, a cache or similar. Permanent recording
15 implies that the data is in the form of an addressable data structure that is retrievable from an addressable memory space and can therefore be stored and retrieved until deleted or over-written, although long-term storage may or may not occur.

20 The sensors 110 may be configured to transduce propagating waves, such as sound waves and/or light waves, to electrical signals encoding the propagating wave data from the scene 140 as sensor data 112.

In some but not necessarily all examples, the sensors 110 are fixed in space relative to the scene space 140. In other examples, the sensors 110 are movable or moving
25 relative to the scene space 140.

In some, but not necessarily all embodiments, the sensors 110 are or comprise image sensors 114. An example of an image sensor 114 is a digital image sensor that is configured to operate as a camera. Such a camera may be operated to record static
30 images and/or video images.

In some, but not necessarily all embodiments, cameras may be configured in a stereoscopic or other spatially distributed arrangement so that the scene 140 is visually sampled from different perspectives. This may enable the creation of a three-

dimensional image and/or processing to establish depth, for example, via the parallax effect.

In some, but not necessarily all embodiments, the sensors 110 are or comprise audio sensors 116. An example of an audio sensor 116 is a microphone or microphones. Microphones may be configured in a stereoscopic or other spatially distributed arrangement such as a microphone array so that the scene 140 is audio sampled from different perspectives. This may enable three-dimensional spatial audio processing, that allows positioning of audio within the scene 140.

In some, but not necessarily all embodiments, the sensors are or comprise depth sensors 118. A depth sensor 118 may comprise a transmitter and a receiver. The transmitter transmits a signal (for example, a signal a human cannot sense such as ultrasound or infrared light) and the receiver receives the reflected signal. Using a single transmitter and a single receiver some depth information may be achieved via measuring the time of flight from transmission to reception. Better resolution may be achieved by using more transmitters and/or more receivers (spatial diversity). In one example, the transmitter is configured to 'paint' the scene 140 with light, preferably invisible light such as infrared light, with a spatially dependent pattern. Detection of a certain pattern by the receiver allows the scene 140 to be spatially resolved. The distance to the spatially resolved portion of the scene 140 may be determined by time of flight and/or stereoscopy (if the receiver is in a stereoscopic position relative to the transmitter).

In these 'passive' or 'non-active' examples of depth sensing the person 150 or object 152 to be sensed are passive and merely reflect incident light or sound waves emitted by a transmitter. However, 'active' examples, which require activity at the sensed object, may additionally or alternatively be used. As an example, a person 150 may carry a positioning apparatus that is configured to position the positioning apparatus within the scene space 140. The positioning apparatus may for example measure motion of the positioning apparatus from a reference position using accelerometers. Gyroscopes may optionally be used to determine orientation. Additionally or alternatively, the positioning apparatus may enable positioning via triangulation (trilateration) by transmitting to multiple receivers and/or receiving from multiple transmitters.

In the illustrated example, but not necessarily all examples, the processing module 120 comprises a memory sub-module 122, a processing sub-module 124, a recognition sub-module 126, and a control sub-module 128. While the 'modules' are described and illustrated separately they may be, although they need not be, separate or combined in different combinations. For example, the processing sub-module 124, the recognition sub-module 126, and the control sub-module 128 may be performed by the same circuitry or under the control of the same computer program. Alternatively one or more of the processing sub-module 124, the recognition sub-module 126, and the control sub-module 128 may be performed by dedicated circuitry or a dedicated computer program. The sub-modules may be performed using software, dedicated hardware or a mix of programmed hardware and software.

The memory sub-module 122 may be used to store unprocessed sensor data and/or processed sensor data 112 (scene data), computer programs, scene space models and other data used by the processing module 120, although other sub-modules may have their own memories.

The processing sub-module 124 may be configured to process the sensor data 112 to determine scene data that is meaningful about the scene 140.

The processing sub-module 124 may be configured to perform image processing where the sensor data 110 comprises image data from a camera or cameras 114.

The processing sub-module 124 may be configured to perform audio processing where the sensor data 110 comprises audio data from a microphone or microphones 116.

The processing sub-module 124 may be configured to perform automatically one or more of the following tasks using the sensor data 112 to create scene data that has potential meaning for the scene 140:

use machine (computer) vision to perform one or more of:

detect a (moving or stationary) object or person,

classify a (moving or stationary) object or person, and/or

track a (moving or stationary) object or person;

use spatial analysis to perform one or more of:

7

position a (moving or stationary) object in the scene space using depth determination; and/or

create a map of the scene space; and/or

5 use behaviour analysis to describe an event that occurs in the scene 140 as a potentially meaningful symbol.

10 An example of image processing is 'histogram of gradient features' analysis which creates a distribution of intensity gradients or edge directions for an image. The image may be divided into small connected regions (cells), and for each cell, a histogram of gradient directions or edge orientations is created for the pixels within the cell. The combination of these histograms then represents a descriptor.

15 An example of audio processing is 'mel-frequency cepstral coefficients' determination, spatial audio processing using, for example, audio beamforming techniques, audio event recognition or classification, speaker recognition or verification or speech recognition.

20 Motion detection may be achieved, for example, using differencing with respect to a background model (background subtraction) or with respect to a preceding image (temporal differencing), or using some other approach such as optical flow analysis using a vector-based approach.

25 Object classification may be achieved, for example, using shape-based analysis and/or motion-based analysis.

30 Classification of a person may be classification that an object is human or classification that an object is a particular human (identification). Identification may be achieved using an attribute or a combination of attributes that uniquely identifies a person within the set of possible persons. Examples of attributes include: biometric features that are or may be particular to a person such as their face or their voice: their shape and size; their behaviour.

35 Object tracking may be achieved by labelling objects and recording the position in the scene 140 of the labelled object. The algorithm may need to deal with one or more of the following events: object entrance to the scene 140; object exit from the scene 140;

object re-entrance to the scene 140; object occlusion; object merge. How to deal with these events is known in the art.

Object tracking may be used to determine when an object or person changes. For example, tracking the object on a large macro-scale allows one to create a frame of reference that moves with the object. That frame of reference can then be used to track time-evolving changes of shape of the object, by using temporal differencing with respect to the object. This can be used to detect small scale human motion such as gestures, hand movement, facial movement. These are scene independent user (only) movements relative to the user.

The system may track a plurality of objects and/or points in relation to a person's body, for example one or more joints of the person's body. In some examples, the system may perform full body skeletal tracking of a person's body.

The tracking of one or more objects and/or points in relation to a person's body may be used by the system 100 in gesture recognition and so on.

Behaviour analysis requires describing an event that occurs in the scene 140 using a meaningful symbology. An event may be something that occurs at a spatio-temporal instance or it may be a spatio-temporal sequence (a pattern of spatio-temporal instances over time). An event may relate to motion of an object (or person) or interaction of a person and object.

In some, but not necessarily all implementations, an event may be represented by a putative symbol defined in terms of parameters determined from the machine (computer) vision analysis and/or the spatial analysis. These parameters encode some or more of what is happening, where it is happening, when it is happening and who is doing it.

The recognition sub-module 126 is configured to recognise a putative symbol encoding an event in the scene 140 as a meaningful symbol associated with a particular meaning.

The recognition sub-module 126 may be configured to recognize a putative symbol, defined in terms of parameters determined from the machine (computer) vision analysis and/or the spatial analysis, and produced by the processing sub-module 124 as having meaning. The recognition sub-module 126 may, for example, store or access
5 a database of meaningful reference symbols and may use a similarity test to determine whether a putative symbol is 'similar' to a meaningful symbol.

The recognition sub-module 126 may be configured as a machine (computer) inference engine or other recognition engine such as an artificial neural network or
10 clustering in the parameter space. The recognition sub-module may 126, in some examples, be trained, via supervised learning, to recognise meaningful symbols as similar to reference symbols.

The control sub-module 128 responds to the determination that a meaningful event
15 has occurred in the scene 140 in dependence on the event:

a) If the putative symbol is similar to a response symbol, the meaningful event is a 'response' event, and the control sub-module 126 performs a response action.

20 In some but not necessarily all examples the action performed may be programmed by a user. In some but not necessarily all examples the programming may occur via motion of the user within the scene.

b) If the putative symbol is similar to a user control symbol, the meaningful event is a
25 'user control' event, and the control sub-module 126 enables user control of monitoring and/or user control of the response.

For example, a user may be able to program or teach a meaningful response symbol that is added to the database of reference symbols. In some but not necessarily all
30 examples the programming may occur via motion of the user within the scene 140.

For example, a user may be able to program or teach an action performed when the putative symbol is matched to a response symbol.

10

The operation of the processing module 120 may be further understood from Fig 2 which illustrates a state machine 200 for the processing module 120. The state machine 200 has a monitoring state 210, a user control state 220 and an automatic response state 230.

5

In the monitoring state 210, the sensors 110 provide sensor data 112, and the processing sub-module 124 automatically processes the sensor data 112 (video and/or audio and/or depth) to create scene data that has potential meaning. The recognition sub-module 126 automatically processes the scene data to identify actual meaning, that is meaningful symbols, within the scene data.

10

The meaningful symbols are predetermined, that is actions have occurred that determine a symbol prior to recognition. However, 'predetermined' should not be considered to mean exact or fixed. The symbol used for similarity matching merely has prior determination, it may dynamically evolve or may be fixed.

15

If the recognition sub-module 126 determines that a meaningful event has occurred in the scene 140, the control sub-module 126 responds automatically depending on the event. If the meaningful event is a 'response' event, then the state machine 200 transitions to the response state 230 and the control sub-module 126 performs a response action. The response action may be associated with the response event. If the event is a 'user control' event, then the state machine 200 transitions to the user control state 220 and user control of monitoring and/or response is enabled. The enablement may be in a manner associated with that user control state 220.

20
25

A particular gesture user input may be used as a 'user control' event. A gesture user input is a gesture that has meaning to the system 100 as a user input.

30

In some examples, the system 100 may track one or more objects and/or points in relation to a person's body in gesture recognition. For example, the system 100 may perform full skeletal tracking of a person's body in gesture recognition.

Implementation of the processor module 120 or part of the processor module 120 may be as controller circuitry. The controller circuitry 120 may be implemented in hardware

alone, have certain aspects in software including firmware alone or can be a combination of hardware and software (including firmware).

5 As illustrated in Fig 3 the controller 120 may be implemented using instructions that enable hardware functionality, for example, by using executable computer program instructions 322 in a general-purpose or special-purpose processor 310 that may be stored on a computer readable storage medium (disk, memory etc) to be executed by such a processor 310.

10 The processor 310 is configured to read from and write to the memory 320. The processor 310 may also comprise an output interface via which data and/or commands are output by the processor 310 and an input interface via which data and/or commands are input to the processor 310.

15 The memory 320 stores a computer program 322 comprising computer program instructions (computer program code) that controls the operation of the processing module 120 when loaded into the processor 310. The computer program instructions, of the computer program 322, provide the logic and routines that enables the processing module to perform the methods discussed with reference to Figs 1, 2, 5, 6,
20 7, 8, 9 and/or 10. The processor 310 by reading the memory 320 is able to load and execute the computer program 322.

The system 100 may therefore comprise an apparatus 120 that comprises:

25 at least one processor 310; and at least one memory 320 including computer program code 322 the at least one memory 320 and the computer program code 322 configured to, with the at least one processor 310, cause the apparatus 120 at least to perform one or more of blocks 124, 126, 128 of Fig 1 and/or one or more of the blocks of Fig 6 and/or 8.

30 As illustrated in Fig 4, the computer program 322 may arrive at such an apparatus via any suitable delivery mechanism 324. The delivery mechanism 324 may be, for example, a non-transitory computer-readable storage medium, a computer program product, a memory device, a record medium such as a compact disc read-only memory (CD-ROM) or digital versatile disc (DVD), an article of manufacture that tangibly
35 embodies the computer program 322. The delivery mechanism may be a signal

configured to reliably transfer the computer program 322 such as a modulated electromagnetic wave or digitally encoded electrical signal. The apparatus 120 may propagate or transmit the computer program 322 as a computer data signal.

- 5 Although the memory 320 is illustrated as a single component/circuitry it may be implemented as one or more separate components/circuitry some or all of which may be integrated/removable and/or may provide permanent/semi-permanent/dynamic/cached storage.
- 10 Although the processor 310 is illustrated as a single component/circuitry it may be implemented as one or more separate components/circuitry some or all of which may be integrated/removable. The processor 310 may be a single core or multi-core processor.
- 15 References to 'computer-readable storage medium', 'computer program product', 'tangibly embodied computer program' etc. or a 'controller', 'computer', 'processor' etc. should be understood to encompass not only computers having different architectures such as single /multi- processor architectures and sequential (Von Neumann)/parallel architectures but also specialized circuits such as field-programmable gate arrays
- 20 (FPGA), application specific circuits (ASIC), signal processing devices and other processing circuitry. References to computer program, instructions, code etc. should be understood to encompass software for a programmable processor or firmware such as, for example, the programmable content of a hardware device whether instructions for a processor, or configuration settings for a fixed-function device, gate array or
- 25 programmable logic device etc.

As used in this application, the term 'circuitry' refers to all of the following:

- (a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry) and
- 30 (b) to combinations of circuits and software (and/or firmware), such as (as applicable):
(i) to a combination of processor(s) or (ii) to portions of processor(s)/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) to circuits, such as a microprocessor(s) or a portion of a microprocessor(s), that require software or firmware for operation, even if the software or firmware is not physically present.

5 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" would also cover an implementation of merely a processor (or multiple processors) or portion of a processor and its (or their) accompanying software and/or firmware. The term "circuitry" would also cover, for example and if applicable to the particular claim
10 element, a baseband integrated circuit or applications processor integrated circuit for a mobile phone or a similar integrated circuit in a server, a cellular network device, or other network device.

The blocks 124, 126, 128 illustrated in the Fig 1 and/or the blocks illustrated in Fig 6
15 and/or 8 may represent steps in a method and/or sections of code in the computer program 322. The illustration of a particular order to the blocks does not necessarily imply that there is a required or preferred order for the blocks and the order and arrangement of the block may be varied. Furthermore, it may be possible for some blocks to be omitted.

20 Fig 5 illustrates an example of operation of the system 100. Part a in the example of Fig 5 illustrates a scene 140. The scene 140 is a real-world scene. The scene 140 comprises a three-dimensional space and may therefore be considered a scene space 140. The scene space 140 is monitored by the system 100. The scene space may
25 therefore be considered a monitored scene space 140. In the example of Fig 5 the monitored scene space 140 is a room comprising four walls, a ceiling and a floor.

In some examples the monitored scene space 140 may comprise no boundaries. Additionally or alternatively in examples the monitored scene space 140 may comprise
30 a plurality of rooms that may or may not be spatially connected.

In the example of Fig 5 a person 150 is standing in the monitored scene space 140. The person 150 is performing at least one gesture user input 146 to define at least one computer-implemented virtual boundary 144. The gesture user input 146 may be

14

referred to as a gesture 146 and may be as described above with reference to Fig 2.
The gesture 146 may be considered a boundary indication gesture 146.

5 In the illustrated example the person 150 is performing a gesture 146 to define a single
computer-implemented virtual boundary 144.

The gesture 146 comprises a motion along a path 148 at a location in the monitored
scene space 140. In the example of Fig 5 the gesture 146 comprises the person 150
moving an object 152 in his hand along a path 148. In the example the path 148 is a
10 straight line across the monitored scene space 140.

In examples, the path 148 may comprise any motion in one spatial dimension, two
spatial dimensions or three spatial dimensions and may take any form.

15 Part c of the example of Fig 5 illustrates an overhead view of the monitored scene
space 140 in which the person 150 is illustrated making the gesture 146 to define a
computer-implemented virtual boundary 144.

20 As can be seen in parts a and c of the example of Fig 5, the gesture 146 starts at a
position having coordinates (x, y, z) in the scene 140 where x is measured from the
left wall, y is measured from the front wall and z is measured from the floor. The gesture
146 ends at a finishing position having coordinates (x', y' and z') (not illustrated in Fig
5).

25 In the example of Fig 5 the gesture 146 comprises the person 150 moving the object
152 from the starting position (x, y, z) along a path 148 which is a straight line across
the monitored scene space 140.

30 In the illustrated example the gesture 146 comprises motion in one spatial dimension
as the value of x changes along the path 148 and the values of y and z remain constant.

In other examples the values of x, y and/or z may vary in any way along the path 148
from any starting position in the monitored scene space 140.

15

Parts b and d of the example of Fig 5 illustrate a monitoring space 142. The monitoring space 142 comprises a virtual representation of the monitored scene space 140 created and used by the system 100. The monitoring space 142 may be formed, for example, from data received by the sensors 110, such as received image data and/or depth data.

Parts b and d illustrate respectively the same views as illustrated in parts a and c of Fig 5.

10 In some examples the system 100 may use information of the geometry of the monitored scene space 140 to form the monitoring space 142. For example, the system 100 may use a three dimensional model of the monitored scene space 140 in the formation of the monitoring space 142. The three dimensional model of the monitored scene space 140 may be created by the system 100 and/or may be received
15 by the system 100. Methods for creating a three dimensional model of the monitored scene space 140 are known in the art.

The system 100 may use the monitoring space 142 to monitor the scene space 140. For example, the system may use the monitoring space 142 to implement and monitor
20 at least one virtual boundary 144.

As indicated by the arrows between parts a and b and parts c and d of Fig 5 there is a correspondence between the monitoring space 142 and the monitored scene space 140. In some examples there may be a one to one mapping between the monitoring
25 space 142 and the monitored scene space 140.

In other examples there may be a scale invariant transformation between the monitored scene space 140 and the monitoring space 142.

30 The system 100 recognizes the gesture 146 performed by the person 150 and in response implements at least one virtual boundary 144 in the monitoring space 142. In some examples, recognizing the gesture 146 may comprise processing received data, such as image data. For example, the system 100 may analyze one or more video feeds.

35

16

In examples, at least part of the virtual boundary 144 is determined by the path 148 in the monitored scene space 140 and the at least part of the virtual boundary 144 is located in the monitoring space at a corresponding location equivalent to the location of the path 148 in the monitored scene space 140.

5

For example, at least part of the virtual boundary 144 may be implemented along the path 148 at a location in the monitoring space 142 equivalent to the location of the path 148 in the monitored scene space 140.

10 The gesture 146 may be an absolute gesture or tethered gesture as described above in relation to Fig 2.

As illustrated in parts b and d, in the example of Fig 5, the virtual boundary 144 is implemented from a point (x, y, z) in the monitoring space 142 along the path 144. The point (x, y, z) in the monitoring space is equivalent to point (x, y, z) in the monitored scene space 140.

15

The system 100 may extend the virtual boundary 144 beyond the path 148. The system 100 may use information of the geometry of the monitored scene space 140, such as a three dimensional model, to extend the virtual boundary 144.

20

In some examples, the system 100 may extend the virtual boundary 144 to at least one boundary in the monitoring space 142 corresponding to at least one physical boundary in the monitored scene space 140. For example, the system 100 may extend the virtual boundary 144 from a first wall to a second wall and/or from a floor to a ceiling (see Fig 6).

25

In the example of Fig 5, the dotted line in part d of the Fig illustrates the extension of the virtual boundary 144 to the left and right boundaries of the monitoring space 142 corresponding to the left and right walls of the monitored scene space 140.

30

In some examples a user may indicate a path 148 comprising part of a virtual boundary enclosing an area and the system 100 may automatically complete the virtual boundary 144 enclosing the area. For example, a user may draw a path 148 that is and/or

17

comprises a portion of a shape and the system 100 may complete the shape when implementing the virtual boundary 144.

5 As an example, a user may draw a path that is a semi-circle and the system 100 may automatically extend the virtual boundary 144 to comprise the complete circle.

In some examples a user may indicate to the system 100 that the virtual boundary 144 should be extended to complete a boundary enclosing an area using a gesture 146 and/or one or more other inputs.

10

In some examples the system 100 may infer from a drawn path 148 that a virtual boundary 144 should be extended to enclose an area and may extend the virtual boundary 144 automatically.

15 The system 100 may provide feedback to a user to confirm that the system 100 has recognized the gesture 146 and has implemented the virtual boundary 144. The feedback may comprise any suitable feedback, for example audio and/or visual and/or tactile feedback and so on. Additionally or alternatively the system 100 may send one or more messages to a device or devices to inform the user that the system 100 has
20 implemented the virtual boundary 144.

The system 100, may process received data to generate a response event when there is, relative to the at least one virtual boundary 144, a change in a portion of the monitored scene space 140. For example, the system 100 may process data received
25 by the sensors 110, such as audio and/or image data, to generate the response event. The system may analyze one or more audio and/or video feeds to generate the response event. In some examples processing the received data to generate the response event may be considered monitoring the virtual boundary 144.

30 In examples the same sensor, such as a camera, may be used in recognition of the gesture 146 defining the at least one virtual boundary 144 and for monitoring the at least one virtual boundary 144.

In some examples, processing received data to generate the response event may
35 comprise processing received data to monitor the virtual boundary 144 to detect

18

activity across the virtual boundary 144 and/or activity within a threshold of the virtual boundary 144. For example, the system 100 may process audio and/or image and/or depth data to monitor the virtual boundary 144 to detect activity across the virtual boundary 144 or within a threshold of the virtual boundary 144.

5

The system 100 may allow a user to verify that the one or more virtual boundaries 144 have been implemented without generating a response event. For example, the a user may define a virtual boundary 144 using a gesture 146 at a location in the monitored scene space 140 and may then interact with the location in the monitored scene space 140 to verify that the virtual boundary 144 is “in place”. In examples the user may move at least a part of his body, such as his hand, through the location in the monitored scene space and the system 100 may provide feedback to confirm that the user has touched and/or crossed the virtual boundary 144. Any suitable feedback may be provided, for example, audio and/or visual and/or tactile feedback. Additionally or alternatively the system 100 may send at least one message to a device or devices of the user.

10

15

In examples the system 100 may begin to process received data to generate the response event after the user has verified the virtual boundary 144. For example, the system 100 will not generate an alert and/or increment a counter until after the user has verified the virtual boundary 144.

20

The response event may cause or change spatial audio processing, for example. The response event may, for example, change adaptive filtering of outputs from a plurality of spatially diverse microphones to change a combined audio signal that is produced. This change may be, for example, to effect directional gain or not to effect directional gain. This change may cause directional gain to enhance audio produced by a tracked object or produced by an object other than the tracked object.

25

A user of the system 100, such as person 150, may configure the response event. For example, a user may perform one or more additional gestures 146 that are recognized by the system 100 to configure the response event. In addition or alternatively the user may provide input to the system 100 to configure the response event using a device such as a tablet or mobile phone.

30

35

The system 100, in some examples, may process received data to generate a response event when there is, relative to the virtual boundary 144, a change in a portion of the monitored scene space 140 in relation to any object 152. For example, the system may generate a response event when any object 152 crosses, touches and/or comes within a threshold of the virtual boundary 144. Such virtual boundaries 144 may be considered general virtual boundaries 144.

In other examples, the system may process received data to generate a response event when there is, relative to the virtual boundary 144, a change in a portion of the monitored scene space 140 in relation to one or more user specified objects 152.

For example, after performing the gesture 146 to define at least one virtual boundary 144, a user 150 of the system 100 may perform a gesture 146 to indicate to the system 100 one or more objects 152. For example, the user 150 may point to the one or more objects 152 using his hand or a mobile device and/or may position the object or objects in a viewfinder of a device.

In other examples, the gesture 146 indicating the one or more objects may be performed before the gesture 146 defining the virtual boundary 144.

The system 100 may recognize the gesture 146 indicating the one or more objects 152 and in response the system 100 may generate a response event only when there is a change in a portion of the monitored scene space 140, relative to the virtual boundary 144, in relation to the indicated objects 152. For example, the system 100 may generate a response event only when one of the indicated objects 152 crosses, touches and/or comes within a threshold of the virtual boundary 144. Such virtual boundaries 144 may be considered object specific virtual boundaries.

With reference to the example of Fig 5, the person 150 may perform the gesture 146 to define the virtual boundary 144 and the system 100 may recognize the gesture 144 and implement the virtual boundary 144.

The person 150 may then perform a further gesture, such as pointing to himself or another person or object. The system recognizes the further gesture and monitors the virtual boundary 144 only in relation to the pointed to object. For example, the system

20

100 may generate a response event when that object crosses, touches and/or comes within a threshold of the virtual boundary 144 but not when any other object 150 crosses, touches and/or comes within threshold of the virtual boundary 144.

5 In some examples a virtual boundary 144 may be conditionally monitored and/or conditionally enabled/disabled. For example, the system 100 may process received data to generate a response event when there is, relative to the virtual boundary 144, a change in a portion of the monitored scene space 140 when one or more criteria are satisfied but does not process received data to generate the response event otherwise.

10

The criteria may be predefined and/or may be configured by a user using a gesture and/or one or more additional inputs.

15

In examples any criterion/criteria may be used. For example, any temporal and/or physical criteria may be used. In examples, a virtual boundary 144 may be conditionally monitored dependent upon time of day and/or day of week and/or week of month and/or month of year and/or independence upon the presence of one or more objects 152 in the monitored scene space 140.

20

Fig 6 illustrates an example of a method 400 performed by the system 100 to produce a combined audio signal 421.

At block 402, the system 100 processes received image data and/or received depth data to track an object moving in a monitored scene space.

25

At block 404, the system 100 determines a bearing of the tracked moving object in the monitored scene space.

30

At block 406, the system 100 adaptively filters outputs from a plurality of spatially diverse microphones in dependence upon the bearing of the tracked moving object in the monitored scene space to produce a combined audio signal.

The blocks 400, 402, 404 may occur in real-time to produce in real-time the combined audio output signal.

35

21

Fig 7 illustrates a tracked moving object 150 in the monitored scene space 140. A coordinate system is also illustrated. In this example, a spherical coordinate system is used to define a bearing 401 of the tracked moving object 150 in the monitored scene space from an origin O. The origin O is located at the plurality of spatially diverse microphones. The bearing 401 in spherical coordinates is (θ', ϕ') , where θ' is a value of θ , a polar angle, between 0 and 180 degrees and ϕ' is a value of ϕ , an azimuthal angle, between 0 and 360 degrees.

The system 100 may be configured to adaptively filter outputs from the plurality of spatially diverse microphones in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space 140 to produce a combined audio signal that has, for example, directional gain 403 enhancing audio produced by the tracked object 150. The combined audio output signal may, for example, have directional gain 403 in a target direction (θ'', ϕ'') where θ'' is a value of θ , the polar angle, between 0 and 180 degrees and ϕ'' is a value of ϕ , the azimuthal angle, between 0 and 360 degrees.

The target direction (θ'', ϕ'') is determined by the bearing (θ', ϕ') of the tracked moving object 150. The target direction (θ'', ϕ'') and the bearing (θ', ϕ') may be aligned (the same), as illustrated in the example of Fig 7 where the target direction (θ'', ϕ'') is also a bearing (θ', ϕ') of the tracked moving object 150, or the target direction (θ'', ϕ'') and the bearing (θ', ϕ') may not be aligned (different).

Fig 8 illustrates an example of the system 100.

In this example, the processing module 120 is configured to process received image data 112 and/or received depth data 112 to track an object 150 moving in a monitored scene space 140. The processing module 120 is configured to determine a bearing 401 of the tracked moving object 150 in the monitored scene space 140. The processing module 120 is configured adaptively filter 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space 140 to produce a combined audio signal 421. The combined audio output signal 421 may be produced in real-time. The system 100 is therefore multiple input 413, single output 421 (MISO).

In the example of Fig 8, a beamformer module 420 is configured to create independent filter coefficients (f_i) 414 for each of the outputs (x_i) 413 from the plurality of spatially diverse microphones 412. Each filter 416 filters the output (x_i) 413 from a particular microphone 412 to produce a filtered output 417 for each microphone 412. For example, each filter 416 may be a finite impulse response (FIR) filter, defined by coefficients f_i 414, that filters the output (x_i) 413 from a particular microphone 412 to produce a filtered output 417 for each microphone 412. The filtered outputs 417 of the plurality of spatially diverse microphones 412 are summed to produce the combined audio output signal S 421.

The beamformer module 420 may control the directional gain of the combined audio output signal S 421. For example, particular filter coefficients (f_i) 414 may be used for each of the outputs (x_i) 413 from the microphones 412 to achieve a desired gain. For example, filter coefficients (f_i) 414 may be selected to produce omni-directional gain or different filter coefficients (f_i) 414 may be selected to produce gain in a target direction.

For example, the filter coefficients (f_i) 414 may be created in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space 140 to produce a combined audio output signal (S) 421 that has directional gain 403 in a target direction.

The above described beamformer module 420 is a filter-sum beamformer module. In some examples, the filter-sum beamformer module may be a delay-sum beamformer module. A delay-sum beamformer uses filter coefficients (f_i) 414 to add a particular delay to each of the outputs (x_i) 413 from the plurality of spatially diverse microphones 412 in the time domain.

If the plurality of spatially diverse microphones 412 (labeled i , where $i = 1, 2, \dots, N$) are aligned along the unit vector $\underline{d}$ with regular separation D , then the scalar separation between an origin at the first $i=1$ microphone 412 and the i th microphone is $D \cdot i$ and the vector separation $\underline{S}_i$ between the first $i=1$ microphone 412 and the i th microphone is $D \cdot i \cdot \underline{d}$. If the directional gain 403 is in a target direction having unit vector $\underline{t}$, then the path length difference L_i at the i th filter compared to the origin ($i=1$) is the component of vector separation $\underline{S}_i$ aligned along the target direction unit vector $\underline{t}$ which is given by the vector dot (scalar) product of the vector separation $\underline{S}_i$ and the target direction unit

23

vector $\underline{t}$ ($L_i = \underline{S}_i \cdot \underline{t} = D_i \cdot \underline{d}_i \cdot \underline{t}$, where ' $\cdot$ ' is the vector dot (scalar) product). The phase difference ϕ_i introduced by the path length difference L_i at the i th filter compared to the origin ($i=1$) is the path length difference L_i divided by the wavelength λ and multiplied by 2π radians ($\phi_i = L_i \cdot 2\pi / \lambda = \underline{S}_i \cdot \underline{t} \cdot 2\pi / \lambda = D_i \cdot \underline{d}_i \cdot \underline{t} \cdot 2\pi / \lambda$). The filter f_i may be $\exp(j \cdot \phi_i)$. Therefore the phase delay introduced by the i th filter to the corresponding output 413 from the corresponding microphone 412 is proportional to $(D_i \cdot \underline{d}_i \cdot \underline{t})$, where ' $\cdot$ ' is the vector dot (scalar) product. The filter f_i is $\exp(j \cdot D_i \cdot \underline{d}_i \cdot \underline{t} \cdot 2\pi / \lambda)$, where λ is the wavelength of the sound. The target direction $\underline{t}$ of the directional gain 403 is dependent upon the bearing 401 of the tracked moving object 150.

10

Generalizing to an arbitrary origin, if each of the plurality of spatially diverse microphones 412 is instead at a distance D_i along a unit vector $\underline{d}_i$, and the directional gain 403 is in a target direction having unit vector $\underline{t}$, then the phase difference (relative to the arbitrary origin) introduced by each filter i to the corresponding output 413 from the corresponding microphone 412 is $D_i \cdot \underline{d}_i \cdot \underline{t} \cdot 2\pi / \lambda$, where ' $\cdot$ ' is the vector dot (scalar) product. The phase difference (relative to an origin at the first $i=1$ microphone 412) introduced by each filter i to the corresponding output 413 from the corresponding microphone 412 is $[D_i \cdot \underline{d}_i \cdot \underline{t} - D_1 \cdot \underline{d}_1 \cdot \underline{t}] \cdot 2\pi / \lambda$, where ' $\cdot$ ' is the vector dot (scalar) product. The filter f_i is $\exp(j \cdot [D_i \cdot \underline{d}_i \cdot \underline{t} - D_1 \cdot \underline{d}_1 \cdot \underline{t}] \cdot 2\pi / \lambda)$, where λ is the wavelength of the sound. The target direction $\underline{t}$ of the directional gain 403 is dependent upon the bearing 401 of the tracked moving object 150.

20

In some but not necessarily all examples, the system 100 may additionally comprise a post-filter 424 for filtering the combined audio output signal (S) 421 that has directional gain 403 in a target direction. The post-filter 424 may, for example, be an adaptive post filter that reduces noise.

25

The plurality of spatially diverse microphones 412 may be configured, for example, as a fixed array in one, two or three dimensions. The plurality of spatially diverse microphones 412 may be in some but not necessarily all examples internal microphones of a device that are permanently integrated as hardware of the device as opposed to externally attached or interconnected as peripherals.

30

Figs 9A to 9E illustrate some examples of how the system 100, in use, adaptively filters outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon

35

24

the bearing 401 of the tracked moving object 150 in the monitored scene space 140 to produce a combined audio signal 421.

5 In these examples, the system 100 adaptively filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space 140 relative to computer-implemented virtual boundaries 144 in a monitoring scene space 142 that corresponds to the monitored scene space 140 as previously described.

10 In these illustrated examples, different portions of the monitored scene space are demarcated by different computer-implemented virtual boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140. It should be understood that the number, shape and other attributes of the boundaries are merely illustrative and that any one or more portions of the monitored scene space
15 may be demarcated in any suitable way by one or more different computer-implemented virtual boundaries 144 in the monitoring scene space 142.

A first portion A of the monitored scene space is demarcated by a first computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142
20 corresponding to the monitored scene space 140.

A second portion B of the monitored scene space is demarcated by a second computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140.
25

A third portion C of the monitored scene space is demarcated by a third computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140.

30 A fourth portion D of the monitored scene space is demarcated by a fourth computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140.

In this example, the fourth portion D shares boundaries 144 with each of the first portion
35 A, the second portion B and the third portion C. That is the first portion A, the second

25

portion B and the third portion C form isolated islands separated by the fourth portion D.

5 The system 100 detects movement of the tracked object 150 and can determine whether the tracked object 150 is in the first, second, third or fourth portion and where it is located that portion.

10 As illustrated in Fig 9A, when the tracked object 150 is in the first portion A, the system 100 adaptively filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space to produce a combined audio signal 421. The combined audio signal (S) 421 has directional gain 403 in a target direction determined by the bearing (θ' , ϕ') of the tracked moving object 150. The target direction (θ'' , ϕ'') and the bearing (θ' , ϕ') are not aligned. The target direction (θ'' , ϕ'') is not the bearing (θ' , ϕ') of the tracked moving object 150 but is a predetermined fixed bearing 161 within the first portion A. The target direction (θ'' , ϕ'') does not change while the tracked moving object 150 moves within the first portion A.

20 As illustrated in Figs 9B and 9D, when the tracked object 150 is in the fourth portion D, the system 100 adaptively filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space to produce a combined audio signal 421. The combined audio signal S 421 has, for example, omni-directional gain 403.

25 As illustrated in Figs 9C, when the tracked object 150 is in the second portion B, the system 100 adaptively filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space to produce a combined audio signal 421. The combined audio signal (S) 421 has directional gain 403 in a target direction determined by the bearing (θ' , ϕ') of the tracked moving object 150. The target direction (θ'' , ϕ'') and the bearing (θ' , ϕ') are aligned (the same). The target direction (θ'' , ϕ'') is also a bearing (θ' , ϕ') of the tracked moving object 150 and changes 161 as the tracked moving object 150 moves within the second portion B.

30

26

As illustrated in Fig 9E, when the tracked object 150 is in the third portion C, the system 100 adaptively filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearing 401 of the tracked moving object 150 in the monitored scene space to produce a combined audio signal 421. The combined audio signal (S) 421 has directional gain 403 in a target direction determined by the bearing (θ' , ϕ') of the tracked moving object 150. The target direction (θ'' , ϕ'') and the bearing (θ' , ϕ') are not aligned. The target direction (θ'' , ϕ'') is not the bearing (θ' , ϕ') of the tracked moving object 150 but is, for example, a predetermined fixed bearing 161 outside the third portion C or is a bearing of a different tracked moving object. The target direction (θ'' , ϕ'') in this example does not change while the tracked moving object 150 moves within the third portion C.

The movement between the portions A, B, C, D across the virtual computer implemented boundaries 144 can be detected as previously described and the adaptive filtering described above for each portion may be a response event as previously described. The adaptive filtering consequentially changes in dependence upon detecting movement of the tracked object 150 from one portion of the monitored scene space to another portion of the monitored scene space across at least one computer-implemented virtual boundary 144.

As previously described the computer-implemented virtual boundaries may be predefined by a user creating portions A, B, C, D.

A user may also define for each portion how the filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 depends upon the bearing 401 of the tracked moving object 150 in the monitored scene space.

For example, a user may define whether or not, the combined audio signal (S) 421 has directional gain 403 or omni-directional gain.

For example, a user may define how the directional gain 403 in a target direction is determined by the bearing (θ' , ϕ') of the tracked moving object 150.

For example, the user may define whether or not the target direction (θ'' , ϕ'') and the bearing (θ' , ϕ') are aligned.

For example, a user may define, when the target direction (θ'' , φ'') and the bearing (θ' , φ') are not aligned, characteristics of the target direction (θ'' , φ'') such as whether it is fixed or variable and its value.

5

In the example illustrated in Figs 9A and 9C the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 has directional gain configured to enhance audio produced by the tracked object 150.

10 In the example illustrated in Figs 9D the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 has directional gain configured to enhance audio produced other than by the tracked object 150.

From the foregoing it will be appreciated that the processing module 120 may process
15 received data 112 to generate a response event when there is, relative to the at least one virtual boundary 144, a change in a portion of the monitored scene space 142; and

in response to the response event, change the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 to produce a combined audio
20 output signal 421. The processing of received data 112 to generate a response event comprises processing received image data and/or received depth data to track the object 150 moving in a monitored scene space and monitor the tracked object in relation to the at least one virtual boundary to detect movement across the at least one virtual boundary and/or movement within a threshold of the at least one virtual
25 boundary. Processing the received data to generate a response event may be conditional, for example, only performed if at least one condition is satisfied and not performed if the at least one condition is not satisfied.

Figs 10A and 10B illustrate some examples of how the system 100, in use, adaptively
30 filters 416 outputs 413 from a plurality of spatially diverse microphones 412 in dependence upon the bearings of tracked moving objects 150A, 150B in the monitored scene space 140 to produce a combined audio signal 421.

In these examples, the system 100 adaptively filters 416 outputs 413 from a plurality
35 of spatially diverse microphones 412 in dependence upon the bearing of the tracked

moving object 150A or the bearing of the tracked moving object 150B relative to computer-implemented virtual boundaries 144 in the monitoring scene space 142.

5 A first portion A of the monitored scene space is demarcated by a first computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140. A second portion B of the monitored scene space is demarcated by a second computer-implemented virtual boundary or boundaries 144 in the monitoring scene space 142 corresponding to the monitored scene space 140.

10

The system 100 simultaneously tracks more than one object. The system 100 simultaneously tracks object 150A and object 150B.

15

When the object 150A is in the first portion A, the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 has directional gain configured to enhance audio produced by the tracked object 150A not by the tracked object 150B.

20

When the object 150A is in the second portion B, the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412 has directional gain configured to enhance audio produced by the tracked object 150B not by the tracked object 150A.

25

It will be appreciated that different conditions when satisfied can cause changes in the adaptive filtering 416 of outputs 413 from the plurality of spatially diverse microphones 412. For example satisfaction of such a condition may be based on a bearing of only the tracked object 150A, based on a bearing of only the tracked object 150B or based on bearings of the tracked object 150A and the tracked object 150B. The condition may change based on a bearing of only the tracked object 150A, based on a bearing of only the tracked object 150B or based on bearings of the tracked object 150A and the tracked object 150B.

30

The term 'comprise' is used in this document with an inclusive not an exclusive meaning. That is any reference to X comprising Y indicates that X may comprise only one Y or may comprise more than one Y. If it is intended to use 'comprise' with an

exclusive meaning then it will be made clear in the context by referring to “comprising only one..” or by using “consisting”.

5 In this brief description, reference has been made to various examples. The description of features or functions in relation to an example indicates that those features or functions are present in that example. The use of the term ‘example’ or ‘for example’ or ‘may’ in the text denotes, whether explicitly stated or not, that such features or functions are present in at least the described example, whether described as an example or not, and that they can be, but are not necessarily, present in some of or all 10 other examples. Thus ‘example’, ‘for example’ or ‘may’ refers to a particular instance in a class of examples. A property of the instance can be a property of only that instance or a property of the class or a property of a sub-class of the class that includes some but not all of the instances in the class. It is therefore implicitly disclosed that a features described with reference to one example but not with reference to another 15 example, can where possible be used in that other example but does not necessarily have to be used in that other example.

Although embodiments of the present invention have been described in the preceding paragraphs with reference to various examples, it should be appreciated that 20 modifications to the examples given can be made without departing from the scope of the invention as claimed.

Features described in the preceding description may be used in combinations other than the combinations explicitly described.

25 Although functions have been described with reference to certain features, those functions may be performable by other features whether described or not.

Although features have been described with reference to certain embodiments, those 30 features may also be present in other embodiments whether described or not.

Whilst endeavoring in the foregoing specification to draw attention to those features of the invention believed to be of particular importance it should be understood that the Applicant claims protection in respect of any patentable feature or combination of

30

features hereinbefore referred to and/or shown in the drawings whether or not particular emphasis has been placed thereon.

I/we claim:

5

CLAIMS

1. A method comprising:
processing received image data and/or received depth data to track an object moving
in a monitored scene space;
5 determining a bearing of the tracked moving object in the monitored scene space; and
adaptively filtering outputs from a plurality of spatially diverse microphones in
dependence upon the bearing of the tracked moving object in the monitored scene
space to produce a combined audio signal.
- 10 2. A method as claimed in claim 1, wherein the combined audio output signal has
directional gain enhancing audio produced by the tracked object.
3. A method as claimed in claim 1 or 2, wherein the combined audio output signal has
directional gain in a target direction.
- 15 4. A method as claimed in claim 3, wherein the target direction is determined by the
bearing of the tracked moving object and/or wherein the target direction is a bearing of
the tracked moving object.
- 20 5. A method as claimed in claim 3 or 4 comprising:
defining at least a first portion of the monitored scene space and a second portion of
the monitored scene space separated by at least one computer-implemented virtual
boundary;
detecting movement of the tracked object from the first portion of the monitored scene
25 space to the second portion of the monitored scene space across the at least one
computer-implemented virtual boundary;
in dependence upon detecting movement of the tracked object from the first portion of
the monitored scene space to the second portion of the monitored scene space across
the at least one computer-implemented virtual boundary, changing the target direction
30 from a first direction being one of one or more target directions associated with the first
portion of the monitored scene space to a second direction being one of the one or
more target directions associated with the second portion of the monitored scene
space, wherein the second direction is different from the first direction
and/or

32

defining a plurality portions of the monitored scene space demarcated by computer-implemented virtual boundaries;

detecting movement of the tracked object from an originating portion of the monitored scene space to a destination portion of the monitored scene space across one or more

5 computer-implemented virtual boundaries;

in dependence upon detecting movement of the tracked object from an originating portion of the monitored scene space to a destination portion of the monitored scene space across one or more computer-implemented virtual boundaries, changing the target direction to point towards a predetermined bearing within the originating portion

10 of the monitored scene space to a predetermined bearing within the destination portion of the monitored scene space.

6. A method as claimed in claim 5 further comprising providing for user predefinition of one or more computer-implemented virtual boundaries and/or user predefinition of one
15 or more target directions.

7. A method as claimed in any preceding claim, wherein adaptively filtering outputs from a plurality of spatially diverse microphones in dependence upon the bearing of the tracked moving object in the monitored scene space to produce a combined audio
20 output signal comprises using a filter-sum beamformer or delay-sum beamformer configured in dependence upon the bearing of the tracked moving object in the monitored scene space.

8. A method as claimed in any preceding claim, comprising: determining whether the adaptive filtering of outputs from the plurality of spatially diverse microphones uses
25 common filtering for the outputs or different filtering for the outputs.

9. A method as claimed in any preceding claim, comprising determining whether the adaptive filtering of outputs from the plurality of spatially diverse microphones has
30 directional gain configured to enhance audio produced by the tracked object or directional gain configured to enhance audio produced other than by the tracked object and/or

determining whether the adaptive filtering of outputs from the plurality of spatially diverse microphones has directional gain configured to enhance audio produced by

the tracked object or directional gain configured to enhance audio produced by a different object.

5 10. A method as claimed in any preceding claim, wherein the processing of the received image data and/or the received depth data to track an object moving in a monitored scene space; the determining a bearing of the tracked moving object in the monitored scene space; and the adaptive filtering of outputs from the plurality of spatially diverse microphones in dependence upon the bearing of the tracked moving object in the monitored scene space, occur in real-time to produce in real-time the
10 combined audio output signal.

11. A method as claimed in any preceding claim, wherein the plurality of spatially diverse microphones form a uniformly spaced array and/or form a fixed array and/or are internal microphones of a device.

15

12. A method as claimed in any preceding claim, comprising: processing received data to generate a response event when there is, relative to at least one virtual boundary, a change in a portion of the monitored scene space;
in response to the response event, changing the adaptive filtering of outputs from the
20 plurality of spatially diverse microphones to produce a combined audio output signal.

13. A method as claimed in claim 12, comprising: recognizing at least one gesture to define the at least one computer-implemented virtual boundary in a monitoring space, wherein there is a correspondence between the monitoring space and the monitored
25 scene space;
causing implementation of the at least one virtual boundary in the monitoring space corresponding to the monitored scene space.

14. An apparatus comprising means for performing the method of at least one of claims
30 1 to 13.

15. Computer program code, that when performed by at least one processor, causes at least the method of at least one of claims 1 to 13 to be performed.

1/5

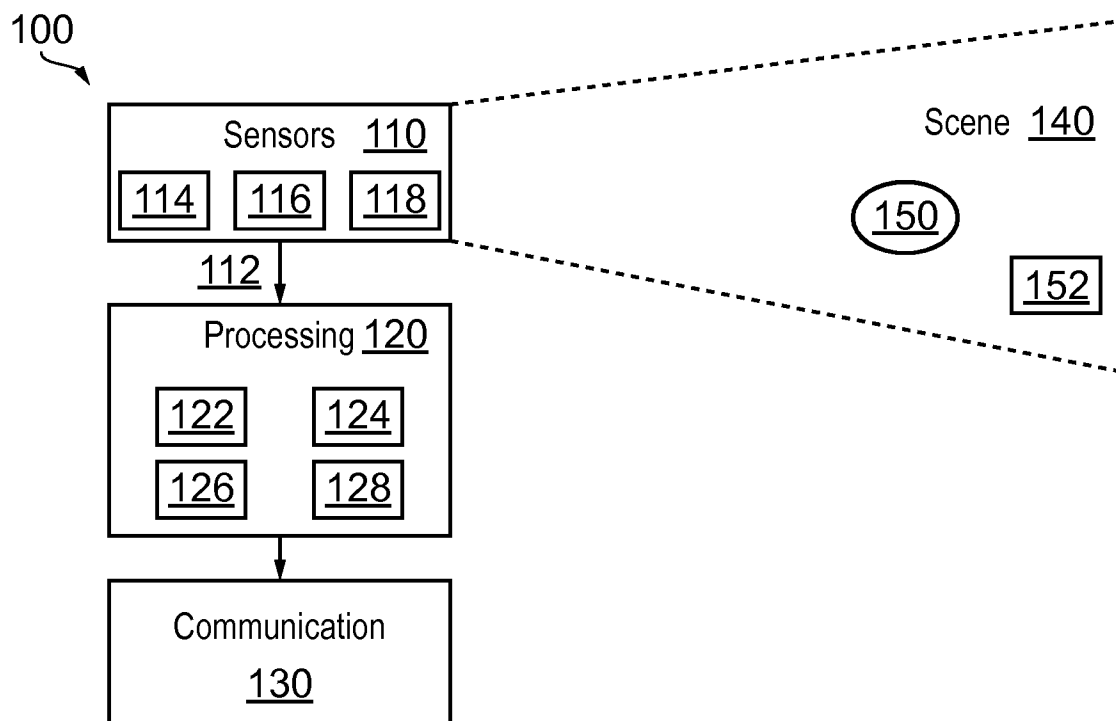


FIG. 1

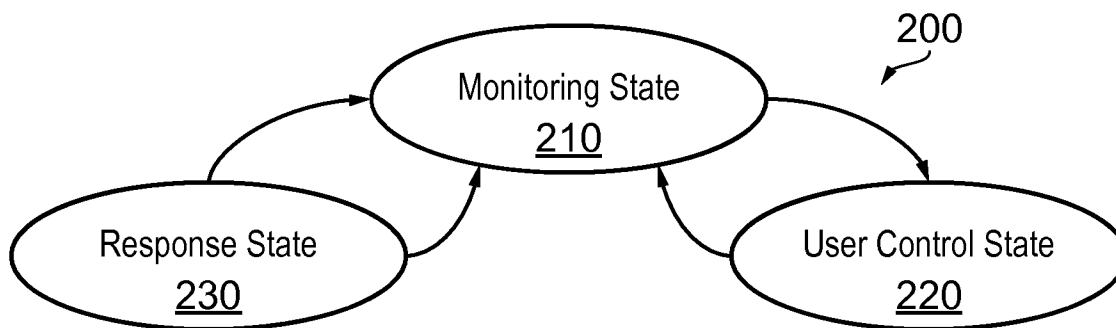


FIG. 2

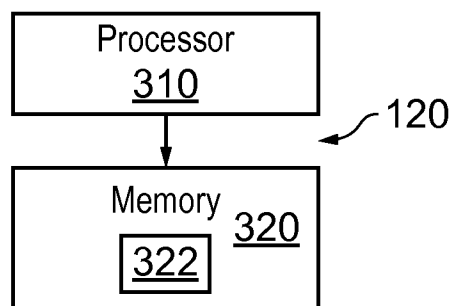


FIG. 3

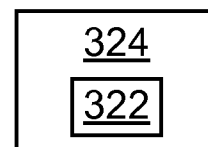


FIG. 4

2/5

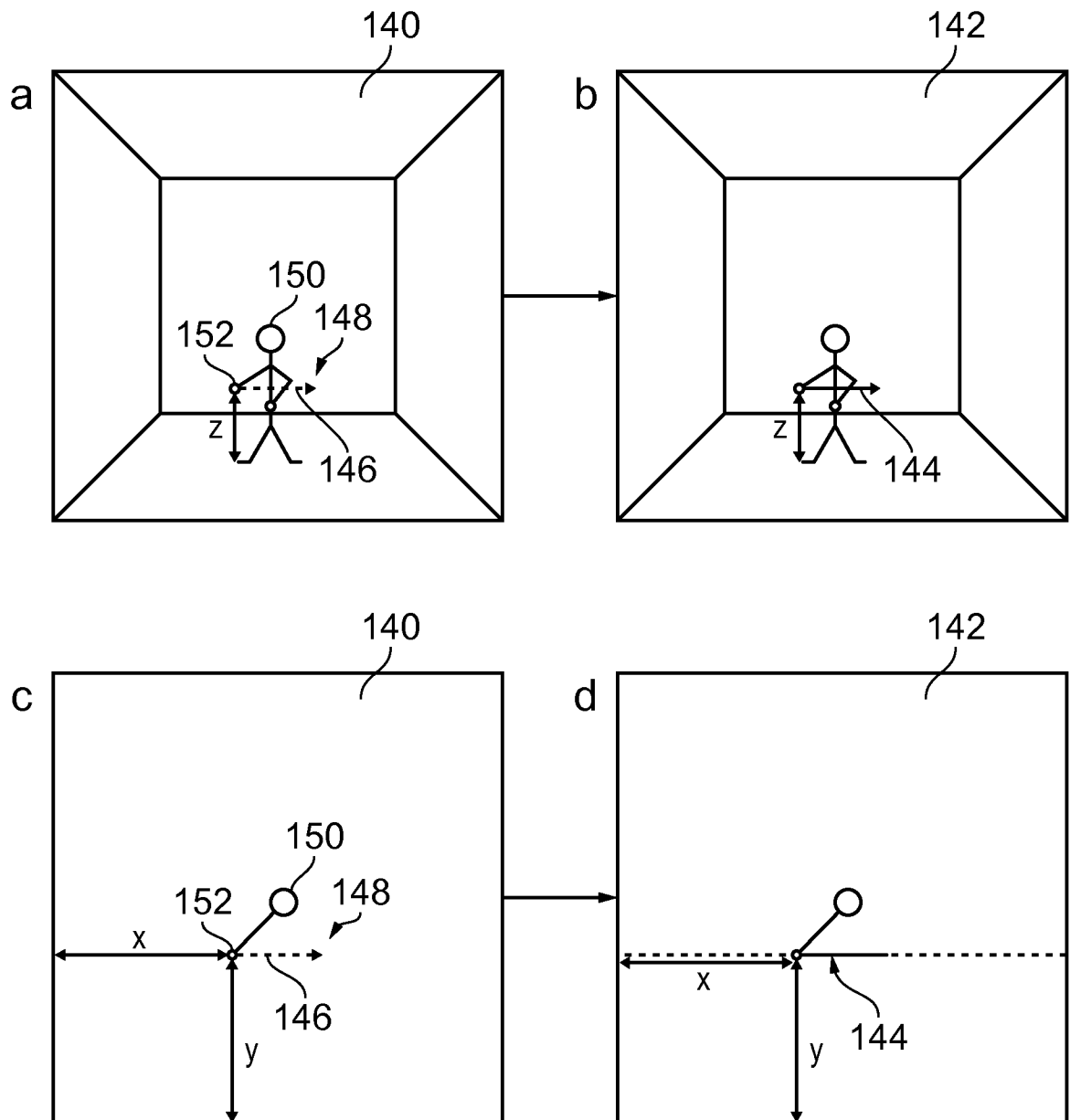


FIG. 5

3/5

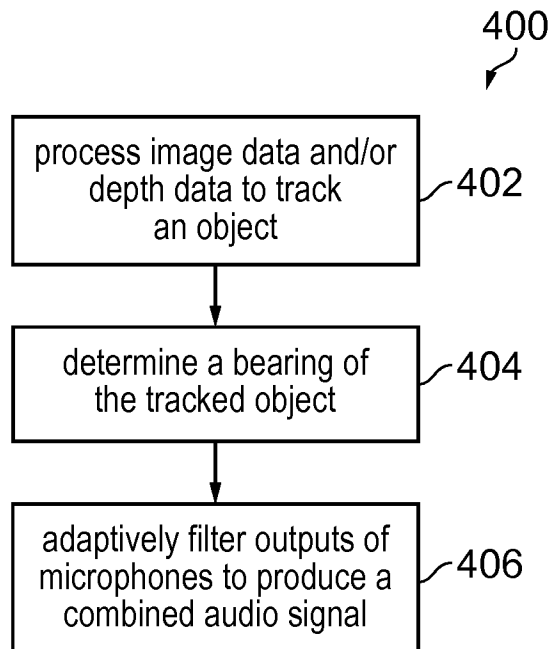


FIG. 6

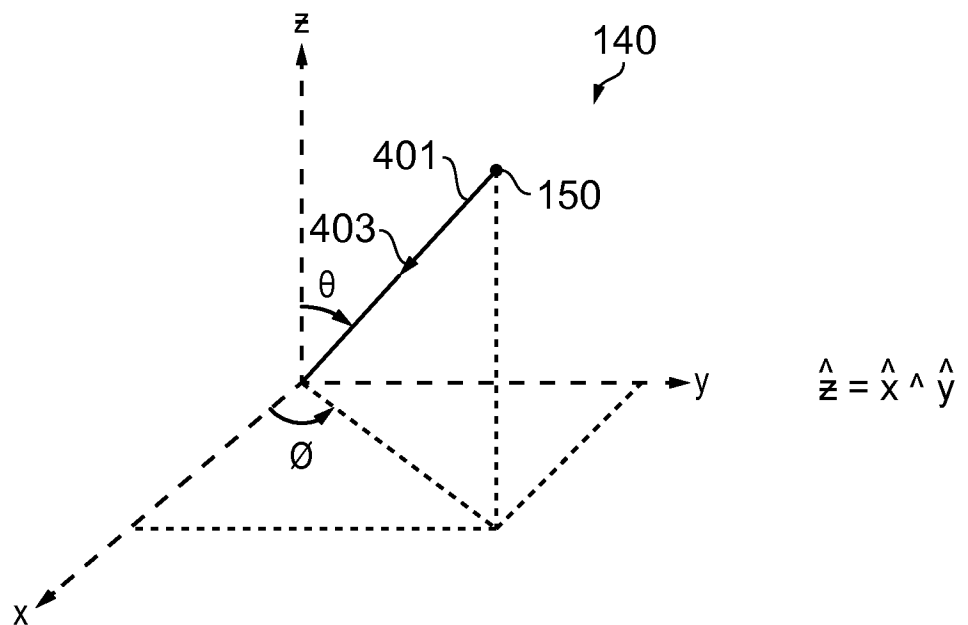


FIG. 7

4/5

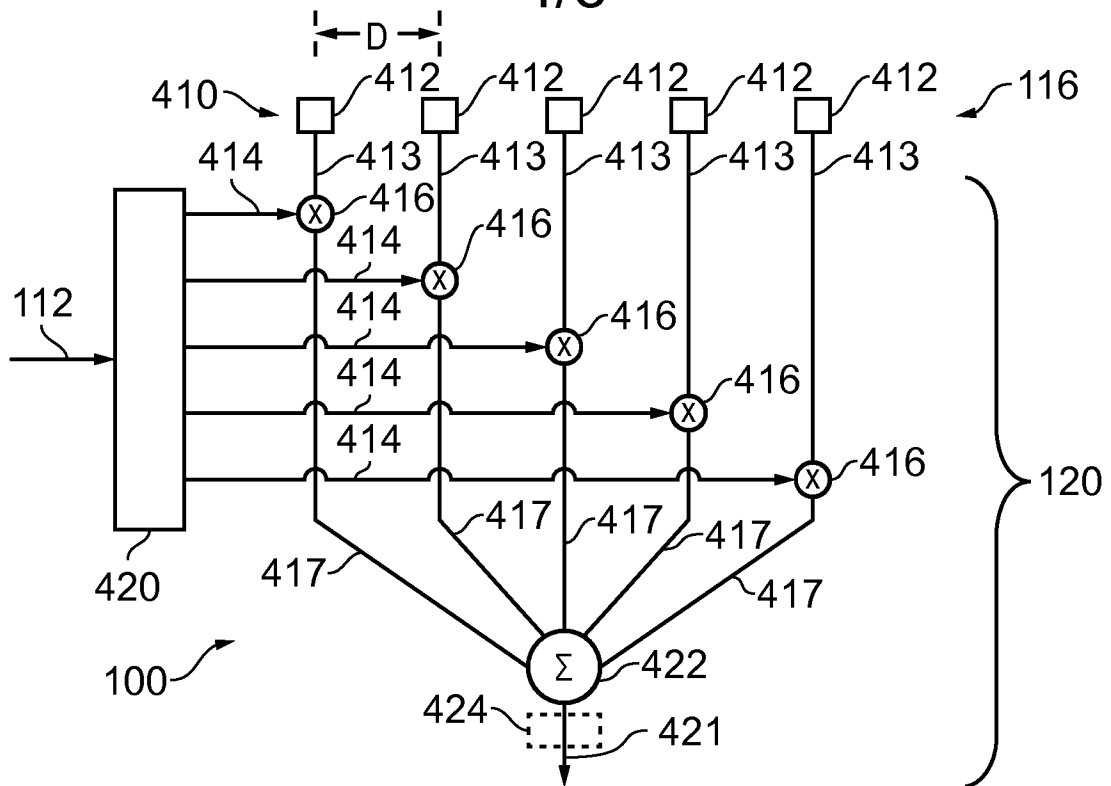


FIG. 8

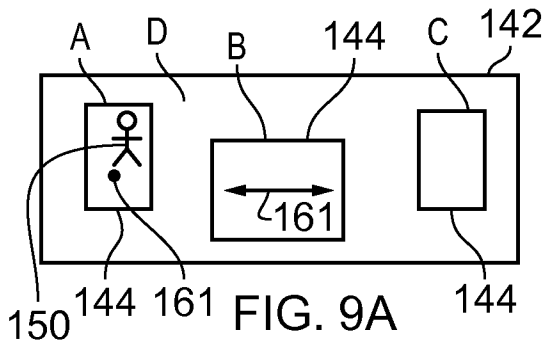


FIG. 9A

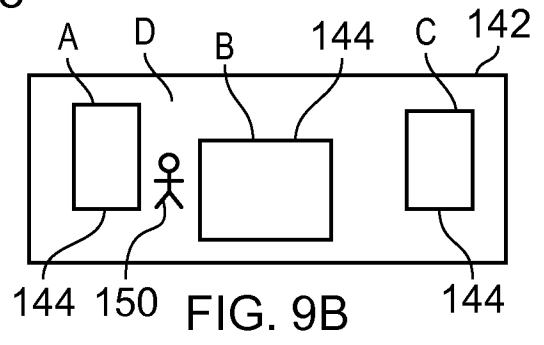


FIG. 9B

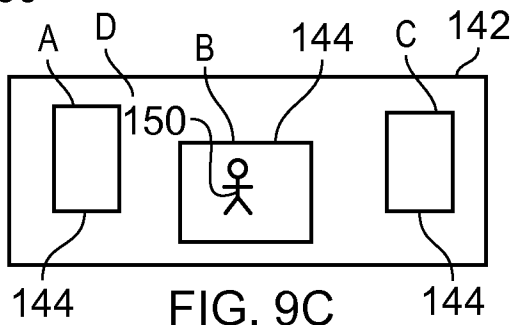


FIG. 9C

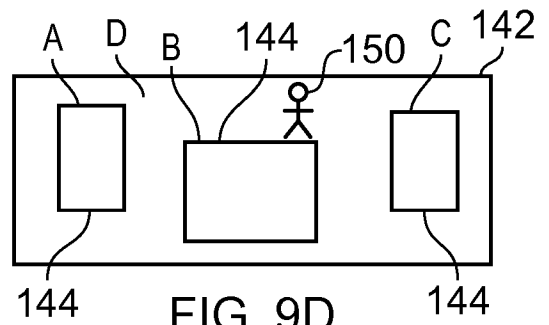


FIG. 9D

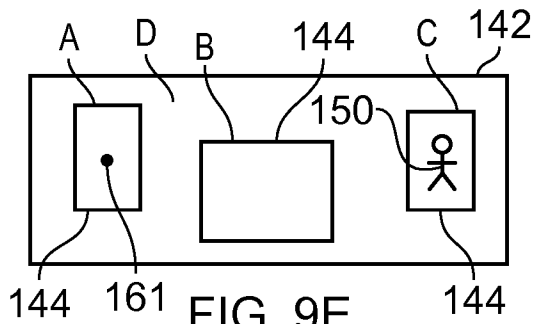
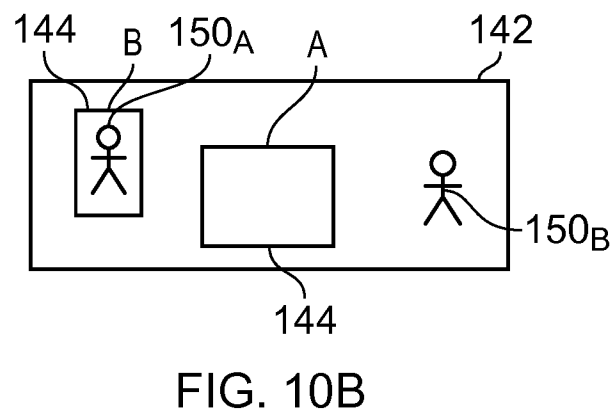
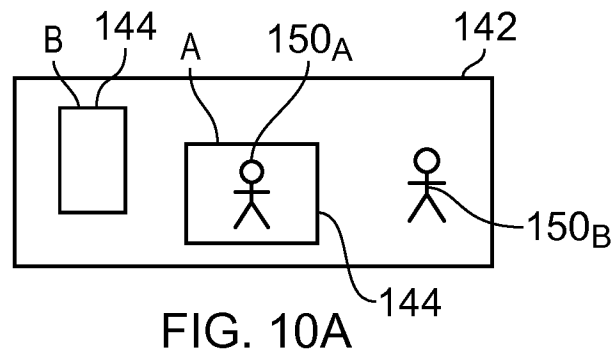


FIG. 9E

5/5



INTERNATIONAL SEARCH REPORT

International application No
PCT/FI2016/050560

A. CLASSIFICATION OF SUBJECT MATTER

INV. H04R3/00
ADD. H04S7/00 G10L21/0216

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04R H04S G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/054943 A1 (ZAD ISSA SYAVOSH [US] ET AL) 26 February 2015 (2015-02-26)	1-4, 7-11,14, 15
Y	abstract paragraph [0004] paragraph [0029] paragraph [0037] - paragraph [0039] paragraph [0052] - paragraph [0056] paragraph [0065] paragraph [0068] - paragraph [0069] paragraph [0095] - paragraph [0096] paragraph [0102] paragraph [0106] - paragraph [0107] paragraph [0123] figures 5,6,11,13,14 ----- -/-	5,6,12, 13



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

30 September 2016

Date of mailing of the international search report

10/10/2016

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Fülöp, István

INTERNATIONAL SEARCH REPORT

International application No
PCT/FI2016/050560

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/155703 A1 (HERNANDEZ-ABREGO GUSTAVO [US] ET AL) 21 June 2012 (2012-06-21) abstract paragraph [0042] - paragraph [0043] figure 6	1-4, 14, 15
Y	----- EP 2 884 492 A1 (SAMSUNG ELECTRONICS CO LTD [KR]) 17 June 2015 (2015-06-17) abstract paragraph [0025] - paragraph [0033] figures 3-6 -----	5, 6, 12, 13

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/FI2016/050560

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015054943 A1	26-02-2015	CN 105637894 A	01-06-2016
		EP 3036913 A1	29-06-2016
		KR 20160045083 A	26-04-2016
		US 2015054943 A1	26-02-2015
		WO 2015026748 A1	26-02-2015

US 2012155703 A1	21-06-2012	CN 103404169 A	20-11-2013
		US 2012155703 A1	21-06-2012
		WO 2012082379 A1	21-06-2012

EP 2884492 A1	17-06-2015	CN 104714734 A	17-06-2015
		EP 2884492 A1	17-06-2015
		KR 20150068112 A	19-06-2015
		US 2015162019 A1	11-06-2015
