



(51) International Patent Classification:

G06T 7/70 (2017.01) *G06T* 7/564 (2017.01)

(21) International Application Number:

PCT/NL2018/050349

(22) International Filing Date:

29 May 2018 (29.05.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

2018998 30 May 2017 (30.05.2017) NL

(71) Applicant: SCISPORTS HOLDING B.V. [NL/NL]; Hengelosestraat 500, 7521 AN Enschede (NL).

(72) Inventors: BROUWER, Gijsbert; Javastraat 121A, 7512 ZE Enschede (NL). NIBBELKE, Vincent; Keppelerdijk 309, 7534 PA Enschede (NL). TON, Bram; Paulus Potterstraat 68, 7545 DX Enschede (NL).

(74) Agent: D'HALLEWEYN, Nele Veerle Trees Gertrudis; Arnold & Siedsma, Bezuidenhoutseweg 57, 2594 AC Den Haag (NL).

(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, DJ, DK, DM, DO,

DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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).

Published:

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

(54) Title: IMAGE PROCESSING METHOD AND SYSTEM FOR DETERMINING A POSITION OF AN OBJECT AND FOR TRACKING A MOVING OBJECT

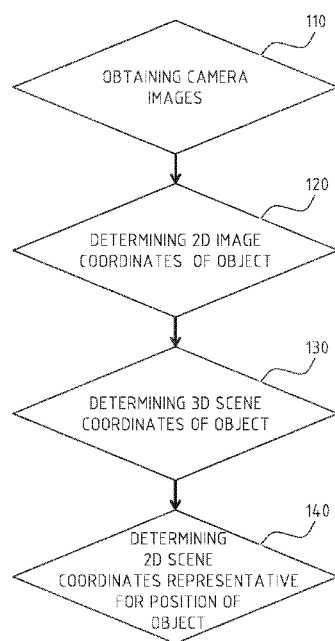


FIG. 1

(57) Abstract: Image processing method for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane, the method comprising: obtaining at least three contemporaneous camera images of the object within the scene, wherein the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene; determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images; determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D image coordinates; determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined 3D scene coordinates.



Image processing method and system for determining a position of an object and for tracking a moving object

Field of Invention

- 5 The field of the invention relates to image processing methods for determining a position of an object within a scene. Particular embodiments relate to the field of image processing methods for determining a position of an object by using multiple contemporaneous camera images. The field of the invention further relates to image processing systems for determining a position of an object within a scene and for tracking a moving object.

10

Background

A variety of techniques are known which perform (re)constructions of 3D objects based on a plurality of camera images. An example of such a technique is voxel carving.

In general two classes of techniques to construct 3D objects are known.

- 15 A first class of 3D construction techniques is concerned with accuracy and a second class of 3D construction techniques is concerned with speed.

Typically, the techniques of the first class are computationally intensive and time consuming.

On the other hand, techniques of the second class may be performed in real-time, but typically do not offer a sufficient amount of accuracy.

- 20 In prior art methods a trade-off needs to be made between spatial resolution and temporal resolution. Moreover, the described prior art methods are not suitable for use in real-time applications where a position of a 3D object needs to be accurately determined.

- 25 An example of an application which requires a fast and accurate determination of the position of an object is video tracking. In video tracking, one or more moving objects need to be located over time using one or more cameras. Video tracking is a time consuming process due to the amount of data that needs to be processed. The amount of required processing power may be further expanded when video tracking is combined with object recognition techniques.

- 30 The objective of video tracking is to determine a position of a 3D object in consecutive video frames. However, when the movement of the to be tracked object or objects cannot easily be predicted, i.e. when the object changes direction and/or orientation over time, prior art tracking methods are not capable of accurately determining the position of the object(s) in real-time.

Summary

- 35 The object of embodiments of the invention is to provide an image processing method for determining a position of an object within a scene which allows for an accurate and fast

determination of the position of the object. It is a further object of embodiment of the invention to provide an image processing method which allows that the position of the object can be accurately tracked in real-time.

- 5 According to a first aspect of the invention there is provided an image processing method for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane. The image processing method comprises the steps of:
- obtaining at least three contemporaneous camera images of the object within the scene, wherein
 - 10 the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene;
 - determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images;
 - determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D
 - 15 image coordinates; and
 - determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined 3D scene coordinates.

Embodiments of the invention are based *inter alia* on the insight that by determining 2D scene

20 coordinates of the object which are representative for the position of the object within the scene, the position of the object can be determined faster as compared to prior art algorithms which track the entire 3D reconstructed object. This is because less processing resources are required for tracking relatively simple 2D scene coordinates as compared with tracking a relatively complex 3D reconstruction. Moreover, since the 2D scene coordinates which are representative for the position

25 of the object within the scene are based on the determined 3D scene coordinates, the 2D position can be accurately determined. By using at least three contemporaneous camera images of the object within the scene and since each of the at least three camera images is associated with a different respective predetermined camera location and orientation, the object is viewed from at least three different perspectives which allows the object to be more accurately constructed as compared to

30 only one or two views of the object are available. Determining 2D image coordinates of the object for only one or two camera images of the object may lead to certain aspects of the object getting lost, or the shape at a particular side of the object not being determined accurately. In other words, determining 3D scene coordinates of the object based the determined 2D image coordinates on only one or two camera images will be prone to errors and will be inaccurate. Therefore, the

35 accuracy of determining 3D scene coordinates of the object is vastly improved by determining 2D image coordinates of the object for at least three contemporaneous camera images which are

associated with three different predetermined camera locations and orientations with regard to the scene.

According to an embodiment determining 2D image coordinates of the object comprises

- 5 determining 2D pixel coordinates for each pixel representing the object; and determining 3D scene coordinates of the object comprises determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates.

- 10 By determining which pixels of the camera image represent the object and determining 2D pixel coordinates thereof, the object can efficiently be distinguished from other things in the camera image, such as a background. Likewise, by determining based on the 2D pixel coordinates 3D voxel coordinates of the 3D voxels representing the object, the object can efficiently be distinguished from a background within the scene in 3D.

- 15 According to an embodiment determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates comprises mapping the 2D pixel coordinates to 3D voxel coordinates.

- 20 This way processing resources can be efficiently allocated to the mapping of only the 2D pixels coordinates which are known to represent the object in stead of mapping all 2D pixel coordinates of each camera image regardless of the corresponding 2D pixel representing the object or not. By doing this, the image processing method can be executed more quickly as compared to prior art methods.

- 25 According to an embodiment, mapping the 2D pixel coordinates to 3D voxel coordinates comprises: determining for each pixel representing the object in a first camera image a set of candidate voxels within the scene which correspond with said respective pixel; and selecting from said set of candidate voxels a voxel representing the object, based on sets of candidate voxels corresponding to pixels representing the object in at least one camera image other than said first
30 camera image.

- This way it can be determined that a particular voxel represents the object when said particular voxel is a candidate voxel representing the object according to at least one pixel of each camera image of the at least three contemporaneous camera images. When a voxel is a candidate voxel to
35 represent the object according to pixels from only one or two video images of the at least three video images, it can be considered that that particular voxel does not represent the object.

According to an embodiment determining the set of candidate voxels for said respective pixel comprises determining a pixel projection line from said respective pixel to the scene and determining that a voxel is a candidate voxel when said voxel is situated substantially along said pixel projection line.

Each of the pixels in the camera image can be seen to correspond with a line of sight towards the scene. This line of sight may be represented by a pixel projection line from said camera position, more in particular from a particular pixel coordinate within a camera image, towards the scene.

More in particular, each of the 2D pixel coordinates representing the object can be seen to correspond with a line of sight or pixel projection line towards the object, the line of sight originating from the respective camera location. The set of candidate voxels within the scene which correspond with a particular pixel representing the object can be chosen along the corresponding line of sight or pixel projection line for that particular pixel. Furthermore, it can be determined that a particular voxel represents the object when said particular voxel is situated on or substantially along a pixel projection line originating from at least one pixel of each camera image of the at least three contemporaneous camera images. On the contrary, when a voxel is situated on or substantially along a pixel projection line originating from pixels of only one or two video images of the at least three video images, it can be considered that that particular voxel does not represent the object. Depending on a specific situation or condition wherein the image processing method according to the invention may be used, a desired accuracy may vary. Depending on the desired accuracy, error margins may be defined for determining whether a voxel is situated on or substantially along a pixel projection line. For example, an error margin may be defined to correspond with the dimensions of a voxel, two voxels, half a voxel, a quarter of a voxel etc. It is clear to the skilled person that any other error margins may be taken into account, depending on case specific conditions.

According to an embodiment determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates comprises dividing the scene into a voxel grid comprising a plurality of voxels, and determining for each voxel of the plurality of voxels whether said voxel represents the object based on the determined 2D pixel coordinates.

This way, 3D voxel coordinates can be determined quickly since the voxels that need to be evaluated in order to find out whether they represent the object or not, have a predetermined shape and a predetermined location within the voxel grid. In other words, for each voxel the 3D coordinates within the scene are known, and it only has to be determined which of the known 3D

coordinates correspond with voxels representing the object, which requires less computational resources as compared to determining a voxel size, shape and location during the step of determining whether a voxel represents the object.

5 According to an embodiment determining for each voxel of the plurality of voxels whether said voxel represents the object based on the determined 2D pixel coordinates comprises mapping corresponding 3D voxel coordinates of said voxel to mapped 2D image coordinates for each one of the at least three camera images, and determining whether said mapped 2D image coordinates correspond with 2D pixel coordinates representing the object.

10 This way processing resources can be efficiently allocated by determining that 3D voxel coordinates do not represent the object when mapped 2D image coordinates of those 3D voxel coordinates do not correspond with 2D pixel coordinates representing the object in one of the at least three obtained camera images. By allocating or assigning resources this way, and by avoiding
15 mapping of 3D voxels coordinates to 2D image coordinates of the other camera images when it is already determined that the 3D voxel coordinates do not correspond with a voxel representing the object, the described image processing method is faster as compared to prior art methods by using processing resources in a more efficient way.

20 According to an embodiment the image processing method further comprises dividing the scene in a plurality of subspaces, wherein each subspace of the plurality of subspaces is comprised between a bottom subplane and an upper subplane, which subplanes are substantially parallel with the bottom plane and the upper plane, and wherein determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates is executed consecutively for
25 each subspace of the plurality of subspaces. In an alternative embodiment, determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates is executed in parallel for each subspace of the plurality of subspaces.

This way, the distance between a bottom subplane and an upper subplane may be determined based
30 on knowledge or assumptions about the object, e.g. regarding the dimensions of the object. Each subspace of the plurality of subspaces may be chosen to comprise an equal volume or alternatively the volumes of different subspace may vary. The order of the subspaces for which the step of determining 3D voxel coordinates is performed may also be determined based on knowledge or assumptions about the object. In an exemplary embodiment the step of determining 3D voxel
35 coordinates may first be performed for a lower subspace closest to the bottom plane of the scene, and subsequently for a higher subspace located above the lower subspace. Alternatively, the step of

determining 3D voxel coordinates may first be performed for a middle subspace between a lower subspace and a higher subspace. This may for example be beneficial when it is suspected, e.g. based on prior knowledge or assumptions on the shape of the object, that there are more voxels representing the object within the middle subspace as compared to the lower subspace or the higher subspace. It is clear to the skilled person that instead of dividing the scene in a plurality of subspaces, the processing method may comprise dividing a part of the scene in a plurality of subspaces. For example, a particular part of the scene may be determined to be more relevant to reconstruct or track a particular object based on knowledge or assumption regarding shape and/or dimensions of the object. Consequently such particular part of the scene can be divided in a plurality of subspaces to enhance efficiency as compared to dividing the entire scene in a plurality of subspaces.

According to an embodiment determining 2D scene coordinates of the object which are representative for the position of the object within the scene comprises projecting the determined 3D scene coordinates onto the bottom plane.

This way an accurate 2D position of the object can be obtained based on the determined 3D scene coordinates, especially when the object is located or positioned on the bottom plane. Moreover, by projecting the determined 3D scene coordinates on the bottom plane the 2D position of the object that is acquired this way allows for an efficient and accurate way of tracking the object within the scene.

In an alternative embodiment, the determined 3D scene coordinates may be projected on a projection plane which is substantially parallel to the bottom plane, but which is located above or below the bottom plane, when the object is located or positioned above or below the bottom plane, respectively.

According to an embodiment determining 2D scene coordinates of the object which are representative for the position of the object within the scene comprises determining a center of the projected determined 3D scene coordinates.

This way eventual required tracking resources may further be limited by narrowing down the size of the information or coordinates being representative for the position of the object. Determining a center of the projected determined 3D scene coordinates may comprise determining a smaller surface within boundaries of the projected determined 3D scene coordinates which is located substantially in the middle of the projected determined 3D scene coordinates. Alternatively, determining a center of the projected determined 3D scene coordinates may comprise determining

a one-dimensional (1D) point within boundaries of the projected determined 3D scene coordinates, which 1D point is located substantially in the middle of the projected determined 3D scene coordinates.

- 5 According to an embodiment, the image processing method further comprises the steps of:
- dividing the scene or a part thereof in at least a first subspace and a second subspace, wherein the first subspace is comprised between a first bottom subplane and a first upper subplane, and wherein the second subspace is comprised between a second bottom subplane and a second upper subplane, which subplanes are substantially parallel with the bottom plane and the upper plane;
 - 10 - determining for each of the at least first subspace and second subspace, whether determined 3D scene coordinates are present in the respective subspace; and
 - determining 2D scene coordinates of the object which are representative for the position of the object within the scene if determined 3D scene coordinates are present in each subspace of the at least first subspace and second subspace.

15

This way a noise cancelling step is built into the image processing method. When the object is expected or suspected to be present in at least the first subspace and the second subspace, it can be determined that when 3D scene coordinates allegedly representing the object are only present in the first subspace or in the second subspace, the 3D scene coordinates actually represent noise,

- 20 background or something else, but not the object. Depending on the expected or estimated nature, size and/or shape of the object, more specific decisions can be made in determining the at least first and second subspace. In other words, the image processing method can be adapted or fine-tuned in respect of a variety of objects of which the position is to be determined or which objects are to be tracked.

- 25 It is clear to the skilled person that in stead of dividing the scene in a plurality of subspaces, the processing method may comprise dividing a part of the scene in a plurality of subspaces. In other words, at least a first and second subspace are determined within the scene, wherein the at least first and second subspace may or may not constitute the entire scene. For example, a particular part of the scene may be determined to be more relevant to reconstruct or track a particular object based
- 30 on prior knowledge or assumptions regarding shape and/or dimensions of the object. Such prior knowledge or assumptions regarding the object may for example be acquired based on a machine learning method, more in particular a deep learning method. Consequently such particular part of the scene can be divided in a plurality of subspaces to enhance efficiency as compared to dividing the entire scene in a plurality of subspaces. Moreover, it is clear to the skilled person that the at
- 35 least first and second subspace may or may not be mutually adjacent subspaces.

The skilled person understands that when a higher amount of subspaces is determined, the method may comprise determining 2D scene coordinates of the object which are representative for the position of the object within the scene if determined 3D scene coordinates are present in at least 60%, preferably at least 70%, more preferably at least 80% and most preferably at least 90% of the determined subspaces. For example, when 10 subspaces are determined, determining 2D scene coordinates of the object which are representative for the position of the object within the scene may be done when the determined 3D scene coordinates are present in at least 6, 7, 8 or 9 subspaces, respectively.

- 10 According to an embodiment the first bottom subplane coincides with the bottom plane and the first upper subplane coincides with the second bottom subplane.

This may be particularly beneficial to determine 2D scene coordinates of an object which is located or expected to be located on the bottom plane, and/or which extends or is expected to extend, perpendicularly from the bottom plane, at least until into the second subspace. It is however clear to the skilled person that the at least first and second subspace do not need to be adjacent spaces.

According to an embodiment each predetermined camera location and orientation corresponds with a predetermined position and orientation with respect to a scene based coordinate system, respectively.

This way, when the position and orientation of each of the cameras is known with respect to the scene, a mapping between 2D image coordinates and 3D scene coordinates may also be predetermined and a correspondence between 2D image coordinates and 3D scene coordinates may for example be stored in a look-up table.

According to an embodiment obtaining each one of the at least three contemporaneous camera images of the object within the scene is performed by at least three corresponding cameras.

This way, the at least three corresponding cameras can be set-up in correspondence with each other and synchronized with each other to provide camera images. It is preferred that the at least three cameras are fixed cameras in order to prevent motion artefacts in comparison as to the use of one or multiple moving cameras.

According to an embodiment, the image processing method further comprises a step of calibrating the at least three cameras, prior to obtaining the at least three contemporaneous camera images of the object within the scene.

- 5 This way correspondence can be created between locations within the scene, more in particular 3D scene coordinates within the scene, and locations within the camera images, more in particular 2D image coordinates within the camera images. This correspondence can then be used later on for 3D scene coordinates of the object based on determined 2D image coordinates of the camera image.
- 10 According to a preferred embodiment, calibrating the at least three cameras comprises minimizing a reprojection error of the cameras.

- This way, by minimizing the reprojection error, a more accurate correspondence between locations within the scene, more in particular 3D scene coordinates within the scene, and locations within the camera images, more in particular 2D image coordinates within the camera images can be
- 15 determined. It is clear for the skilled person that other known ways of calibrating the at least three cameras or improving a calibration of the at least three cameras may be used.

- According to an embodiment determining 2D image coordinates of the object for each one of the at least three camera images comprises determining which pixels represent the object in each one of
- 20 the at least three camera images.

- This way, computational resources may be allocated efficiently by only assigning the resources to map 2D image coordinates which correspond to pixels representing the object, in stead of mapping
- 25 all 2D image coordinates, irrespective of whether they correspond to pixels representing the object or not.

- According to an embodiment determining which pixels represent the object in each one of the at least three camera images comprises performing an image segmentation method, e.g. a background
- 30 subtraction, for each one of the at least three camera images.

- This way it can be efficiently determined which pixels represent the object since only pixels representing the object will remain in the camera image after an image segmentation has been carried out. An exemplary method for performing an image segmentation is background
- 35 subtraction. The skilled person is aware of different ways of performing an image segmentation and/or a background subtraction.

According to an embodiment determining which pixels represent the object in each one of the at least three camera images comprises at least one of noise filtering and shadow filtering based on a predetermined threshold.

5

This way the accuracy of the image processing method can be improved. By performing noise filtering and/or shadow filtering in an early stage, i.e. while determining which pixels of the camera images represent the object, these filtering techniques can be applied in 2D images, which requires less computational resources as compared to performing these or similar filtering techniques within a 3D scene or voxel space.

10

According to an embodiment determining which pixels represent the object in each one of the at least three camera images comprises a blob classification based on an estimated shape of the object.

15

This way, knowledge or assumptions on the shape, nature and/or size of the object can be used either to filter out further noise or to more quickly determine which pixels represent the object and which pixels do not.

According to a second aspect of the invention, there is provided an image processing method for tracking a moving object within a scene, which comprises performing any one of the above described image processing method for determining a position of the object at consecutive moments in time.

25 This way, the position of a moving object can be accurately tracked over time, in real-time.

According to an embodiment the at least three camera images constitute at least three video frames of at least three corresponding video feeds.

30 This way, the position of the object can be accurately determined over time by using at least three synchronized video cameras.

According to an embodiment the object is a human and the first subspace is between 0 cm and 120 cm, preferably between 0 cm and 100 cm, and more preferably between 0 cm and 75 cm; and wherein the second subspace is between 75 cm and 200 cm, preferably between 95 cm and 185 cm, and more preferably between 115 cm and 170 cm.

35

This way, an extra safety is built into the image processing method which reduces the risk of erroneous determination of position of the human.

It is clear to the skilled person that additional subspaces may be determined within the scene to even more accurately determine the position of the human or track the human. Depending on the actual camera setup and/or the dimensions of the scene and/or particular contextual information regarding the human and/or the scene, the amount, shape and/or dimensions of the subspaces within the scene may be determined or adjusted.

- 10 The skilled person will understand that the hereinabove described technical considerations, functionalities and advantages for method embodiments also apply to the below described corresponding system embodiments, *mutatis mutandis*.

According to a second aspect of the invention there is provided an image processing system for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane. The image processing system comprises:

- an obtaining unit for obtaining at least three contemporaneous camera images of the object within the scene, wherein the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene;
- a first determining unit for determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images;
- a second determining unit for determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D image coordinates; and
- a third determining unit for determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined 3D scene coordinates.

According to a further aspect of the invention, there is provided a computer program comprising computer-executable instructions to perform the method, when the program is run on a computer, according to any one of the steps of any one of the embodiments disclosed above.

According to a further aspect of the invention, there is provided a computer device or other hardware device programmed to perform one or more steps of any one of the embodiments of the method disclosed above.

According to another aspect there is provided a data storage device encoding a program in machine-readable and machine-executable form to perform one or more steps of any one of the embodiments of the method disclosed above.

5 Brief description of the figures

The accompanying drawings are used to illustrate presently preferred non-limiting exemplary embodiments of devices of the present invention. The above and other advantages of the features and objects of the invention will become more apparent and the invention will be better understood from the following detailed description when read in conjunction with the accompanying drawings,
10 in which:

Figure 1 is a flowchart which illustrates an exemplary embodiment of an image processing method according to the invention;

Figure 2 schematically illustrates an object within a scene and an exemplary set-up to perform exemplary embodiment of the image processing method according to the invention;

15 Figure 3 illustrates a step of an exemplary embodiment of the image processing method according to the invention;

Figures 4A and 4B schematically illustrate exemplary scenarios in which an embodiment of the image processing method according to the invention may be applied; and

Figure 5 schematically illustrates an exemplary embodiment of an image processing system
20 according to the invention.

Description of embodiments

Figure 1 illustrates the steps 110, 120, 130 and 140 of an exemplary embodiment of an image processing method 100 for determining a position of an object 250 within a scene 260. The image
25 processing method 100 and corresponding steps 110, 120, 130, 140 thereof will be further elaborated below with respect to Figure 2.

The scene 260 is a predetermined region in space, which is defined between a bottom plane 261 and an upper plane 262 being substantially parallel to the bottom plane 261. In the embodiment of
30 Figure 2, the scene 260 is comprised between a rectangular bottom plane 261 and a rectangular upper plane 262 which is parallel to the bottom plane 261. The scene 260 may be further defined to be bounded by the faces of an imaginary cuboid of which the upper face corresponds with the upper plane 262, the bottom face corresponds with the bottom plane 261 and the side face are defined perpendicularly between the bottom plane 261 and upper plane 262 as illustrated by the
35 dotted lines in Figure 2. It is clear to the skilled person that the application of embodiments of the image processing method 100 is not limited to situations in which the scene 260 is either defined

between two substantially parallel planes 261, 262 or bounded by a rectangular cuboid, but that the scene 260 merely represents a predetermined region in space which can have various shapes, dimensions and/or sizes. Moreover, the skilled person understands that the image processing method 100 according to the invention is applicable in a similar way when the bottom surface 261 is not a plane surface 261 as illustrated in Figure 2, e.g. when the bottom surface 261 is a curved, toothed or serrated surface or a surface with any other shape. A similar reasoning applies to the upper surface 261 *mutatis mutandis*.

Image processing method 100 comprises step 110 of obtaining at least three contemporaneous camera images 211, 212, 213 of the object 250 within the scene 260. The at least three camera images 211, 212, 213 are associated with different predetermined camera locations and orientations with regard to the scene 260. Preferably, the at least three contemporaneous camera images 211, 212 and 213 of the object 250 within the scene 260 are generated by corresponding cameras 201, 202 and 203, respectively. For the sake of simplicity, the object 250 in Figure 2 is illustrated as a cube 250 which is located within the scene 260, and more in particular on the bottom plane 261. However, image processing method 100 can be applied to different objects with any shape and/or size. Moreover, the object 250 does not have to be located on the bottom plane 261 in order for the image processing method 100 to be applied. The object 250 may just as well be located on a different plane, being either a real plane or an imaginary plane, which is positioned somewhere above the bottom surface 261. Furthermore, although only one object 250 is illustrated in Figure 2, the image processing method may be simultaneously applied to a plurality of objects 250 within the scene 260 in order to determine the positions of the respective objects within the scene 260. In Figure 2 three cameras 201, 202, 203 are shown at their corresponding locations which are considered to be known with respect of the scene 260. Along with the location of each camera 201, 202 and 203, also the orientation of each camera 201, 202, and 203 is considered to be known. If these are not known, it is preferred to include a step of calibrating the cameras 201, 202 and 203 in order to obtain or estimate the intrinsic and extrinsic parameters of each camera 201, 202 and 203. Both intrinsic and extrinsic parameters can be determined or estimated in order to create an accurate mapping between the 2D image coordinates x_i, y_i within a camera obtained image 211, 212, 213 and the 3D scene coordinates X_s, Y_s, Z_s within the scene 260. Intrinsic parameters of the cameras 201, 202, 203 may comprise focal length, principal points, distortion coefficients, etc. and extrinsic parameters of the cameras 201, 202, 203 may comprise translation and rotation parameters. The skilled person is aware of a variety of calibration techniques that might be applied in order to perform the calibration. The cameras 201, 202 and 203 are illustrated in Figure 2 as being positioned at different locations outside of the scene 260. However, cameras may also be positioned within the scene or alternatively some cameras might be positioned within the scene and

other cameras may be positioned outside of the scene. Although only drawn schematically, the cameras 201, 202 and 203 may be positioned at different heights, as can be seen from the corresponding camera images 211, 212 and 213, respectively. Image 211 illustrates the object 250 as viewed by camera 201. From the perspective of camera 201, the cube 250 can be seen as an irregular hexagon 211a. Therefrom, it can be understood that the camera 201 is positioned at a height that it is slightly higher than the highest point or plane of the cube 250 and that the camera 201 views the cube at a small angle from above. Image 212 illustrates the object 250 as viewed by camera 202. From the perspective of camera 202, the cube 250 can be seen as another irregular hexagon 212a which is different from the irregular hexagon of image 211. From the hexagon in image 212 it can be understood that the camera position of camera 202 is translated in the X_s direction as compared to the camera position of camera 201, that the camera position of camera 202 is higher as compared to the camera position of camera 201 (i.e. translated in the Z_s direction), and that the camera is rotated as compared to camera 201 and thus that its orientation is different. Image 213 illustrates the object 250 as viewed by camera 203. From the perspective of camera 203, the cube 250 can be seen as a square 213a. Therefrom, it can be understood that the camera 203 is positioned and oriented centrally opposite to one of the planes of the cube 250. From Figure 2 it can be seen that camera 203 views one of the side planes of the square 250 and that the camera is positioned at a lower height as compared to cameras 201 and 202.

In Figure 2, the coordinate axes x_i, y_i as shown in image 211 represent an exemplary origin of the 2D image coordinates. The 2D image coordinates can be used to represent for each image 211, 212 and 213 (only shown for image 211) the position of a pixel within the respective image. Moreover, 2D image coordinates correspond to pixel coordinates within an image can be used to uniquely describe the position of a certain pixel within the respective image. In other words, there is a one to one correspondence between the pixels of an image and the 2D image coordinates. It is clear to the skilled person that the choice of the position of the origin regarding the 2D image coordinates for a certain camera image is an arbitrary choice.

The coordinate axes X_s, Y_s, Z_s as shown on edges of the scene 260 represent an exemplary origin of the 3D scene coordinates. The 3D scene coordinates can be used to represent a location and/or a voxel location within the scene. The 3D scene coordinates may also be used to represent a location outside of the scene, for example to express the corresponding locations of cameras 201, 202 and 203 with respect to the scene.

Although only three cameras are shown in Figure 2 for obtaining three contemporaneous camera images 211, 212, 213, it is clear to the skilled person that according to the image processing

method 100 any higher amount of cameras may be used to obtain more than three contemporaneous camera images, for which steps 120, 130 and 140 can be performed.

Image processing method 100 further comprises step 120 of determining 2D image coordinates of the object 250 for each one of the at least three camera images 211, 212, 213. In view of the above, step 120 comprises determining which pixels in the camera image 211, 212, 213 are pixels that represent the object 250 and determining the 2D image coordinates of the pixels that represent the object 250. In Figure 2, for the sake of simplicity a situation is shown wherein the scene only comprises one object 250 and wherein there is nothing else comprised within the scene. There is not shown any background within the scene, and the images 211, 212, 213 do not contain any noise. However, taking into account more realistic circumstances it may be less straightforward to determine which pixels in the camera image 211, 212, 213 actually correspond to the object, and which pixels do not. Preferably, a background subtraction or other image segmentation is performed for each obtained camera image 211, 212, 213 in order to identify the pixels corresponding to the object 250 and distinguishing within the camera image 211, 212, 213 the pixels corresponding to the object 250 from pixels corresponding to a background within the scene 260. The skilled person is aware of a variety of techniques that can be applied to perform a background subtraction. Alternative to or in addition to background subtraction or image segmentation further step may be taken to filter out the pixels corresponding to the image such as noise filtering and/or shadow filtering. Such filtering steps may be based on one or multiple predetermined threshold values which can be determined for any specific case. The used threshold or thresholds may for example depend on the type of object, the nature of the scene, positioning of the cameras, resolution of the cameras, etc. In situations wherein information on the shape of an object is available or might be estimated, a further step of performing a blob classification may be executed in order to quickly and accurately determine which pixels within the image 211, 212, 213 represent the object, or objects in cases where the location of multiple objects is to be determined. When multiple objects with different shapes are to be identified within a camera image, multiple blob classifications can be performed.

The result of step 120 is that for each of the camera images 211, 212, 213 it is determined on the one hand which pixels represent the object 250 within the respective camera image, and on the other hand for each pixel representing the object 250 within the respective image which 2D image coordinates correspond to the respective pixel. Image processing method 100 further comprises step 130 of determining 3D scene coordinates of the object based on the determined 2D image coordinates.

In an exemplary embodiment of the image processing method 100, step 130 comprises mapping the 2D image coordinates, also referred to as 2D pixel coordinates, to 3D scene coordinates, also referred to as 3D voxel coordinates. In an alternative embodiment, step 130 comprises a mapping in the opposite direction, i.e. mapping 3D scene coordinates to 2D image coordinates in order to
5 determine which 3D scene coordinates correspond to the object 250 and which 3D scene coordinates do not. Both embodiments will be further described below.

In the exemplary embodiment wherein step 130 comprises mapping 2D image coordinates to 3D scene coordinates, for each pixel representing the object 250 in a camera image a set of candidate
10 3D scene coordinates or voxels is determined. This may be done by projecting the respective pixel back to the scene 260, taking into account the intrinsic and extrinsic camera parameters. Such a back projection will result in a “line” of 3D scene coordinates which potentially correspond to the respective pixel. This principle is illustrated in Figure 3 and will be further elaborated below.

Preferably, such determination of which set of 3D scene coordinates correspond to a certain pixel
15 within a camera image, irrespective of the presence of an object within the scene, is performed during a step of calibrating the cameras 201, 202, 203 before the camera images 211, 212, 213 are obtained. The correspondence between the 2D image coordinates and 3D scene coordinates that is obtained this way, may then be stored, e.g. in a look-up table, and be used later on during the image processing method to save time. Since the obtained camera images are 2D images which do
20 not contain depth information, it can be understood that each pixel in a camera image may correspond to multiple 3D scene coordinates, which are located substantially along a hypothetical line and are located at different distances from the camera along that hypothetical line. In other words, one camera image does not allow to accurately determine where an observed object 250 or observed part of an object is located within the scene 260 based on the 2D image coordinates of the
25 respective camera image. However, by determining for each pixel representing the object a set of candidate 3D scene coordinates and by doing this for each of the obtained camera images of the object 250 within the scene 260, these sets of candidate 3D scene coordinates or these pixel projection lines may be combined in order to determine a probability that a certain 3D scene coordinate actually corresponds to the object 250. Pixel projection lines from different camera
30 images, but relating to a same point or area of the object will intersect in a certain voxel or 3D scene coordinate. The meaning of this intersection point is that the particular point is viewed by two different cameras from two different locations, and therefore the probability of the 3D scene coordinates of that point representing the object is raised. From this, it can be understood that if more pixel projection lines, corresponding to a same point or area of the object and originating
35 from different cameras, intersect within a particular 3D voxel or at particular 3D scene coordinates, the likelihood of that particular 3D voxel corresponding to the object rises. Depending on the

specifics of the case and on the amount of cameras used to observe an object within a scene, a probability threshold may be set such that it is determined that an 3D voxel coordinate represents the object, when the amount of intersections in that 3D voxel is higher than the probability threshold.

- 5 By only taking into account pixels that represent the object when determining the sets of candidate voxels, computational resources may be efficiently allocated and step 130 of image processing method 100 may be performed faster as compared to prior art methods which take into account all pixels of all obtained camera images.
- 10 As previously stated, according to an alternative embodiment step 130 comprises a mapping from the 3D scene coordinates to the 2D image coordinates in order to determine 3D voxel coordinates for each voxel representing the object 250 based on the determined 2D pixel coordinates. This embodiment preferably comprises a step of dividing the scene 260 into a voxel grid. In other words, the predefined region in space which is determined by the scene 260 is filled with voxels.
- 15 Preferably the scene is completely and homogeneously divided or filled, such that all voxels have the same dimensions and that the voxels are arranged in a grid such that each 3D scene coordinate corresponds with a voxel. Depending on the size and shape of the voxels, multiple 3D scene coordinates may correspond with one voxel. It is clear to the skilled person that the size and/or shape of the voxels can be determined based on the specifics of the case and on the required
- 20 accuracy. When the voxel grid is established, it can be determined for each voxel within the voxel grid whether that voxel corresponds with the object or not, by mapping its 3D scene coordinates to all corresponding 2D image coordinates of all relevant cameras, i.e. the cameras which have pixels that correspond with the particular voxel. If all corresponding mapped 2D image coordinates correspond with 2D pixel coordinates representing the object, it may be determined that the
- 25 particular voxel represents the object. Moreover, it may be determined that the particular voxel does not represent the object when at least mapped 2D image coordinates of one camera image do not correspond to 2D pixel coordinates representing the image. Alternatively, a first probability threshold may be set, such that it can be determined that a particular voxel represents the object when the amount of cameras for which the mapped 2D image coordinates correspond to 2D pixel
- 30 coordinates representing the object is higher than the first probability threshold. In addition, or alternatively, a second probability threshold may be set, such that it can be determined that a particular voxel does not represent the object when the amount of cameras for which the mapped 2D image coordinates do not correspond to 2D pixel coordinates representing the object is higher than the second probability threshold.
- 35 The steps of the above described approach may be executed for each voxel in the voxel grid subsequently or simultaneously if the computational resources are sufficient. If the steps are

executed subsequently, a preferred order of going through the voxel grid may be determined based on an expected shape or position of the object, the amount or dimensions of the voxels, etc.

In Figure 3, the above described principles are illustrated in view of an elementary example. Figure 3 illustrates eight voxels from a scene 360, from which voxels voxel 350 is a voxel representing the object. Also, four pixels from a first camera image 311 and four pixels from a second camera 313 image are illustrated, from which shaded pixels 311a and 313a are pixels representing the object. For illustrative purposes, and for the sake of simplicity, only eight voxels of the scene 360 are shown with one of those voxels representing the object, likewise only two camera images 311, 313 are shown, each comprising four pixels, more in particular one pixel 311a, 313a representing the object, and three pixels not representing the object. No cameras are shown in Figure 3, although it is clear to the skilled person that the camera images 311, 313 may have been taken with cameras of which the position and orientation with respect to the scene 360 is similar, mutatis mutandis, to the position and orientation of cameras 201 and 203 with respect to the scene 260 as illustrated in Figure 2. As described above, each pixel in a camera image 311, 313 may correspond to multiple voxels within the scene, which are located substantially along a hypothetical line 311b and 313b respectively, and are located at different distances from the camera along that hypothetical line. One camera image does not allow to accurately determine where an observed object 250 or observed part of an object 350 is located within the scene 360 based on the 2D image coordinates of the respective camera image 311, 313. However, by determining for each pixel representing the object 311a and 313a a set of candidate 3D scene coordinates or voxels, and by doing this for each of the obtained camera images 311, 313 of the object 350 within the scene 360, the resulting sets of candidate 3D scene coordinates or these pixel projection lines may be combined in order to determine a probability that a certain 3D scene coordinate or voxel actually corresponds to the object 350. Lines 311b and 313b illustrate pixel projection lines from different camera images 311 and 313, which relate to a same point or area of the object 350. For each of the pixels 311a and 313a two candidate voxels are determined, which are indicated by means of a similar shading as corresponding to the shading of the respective pixels. The pixel lines intersect voxel 350, which is indicated by the double shading, corresponding with both pixel 311a and pixel 313a. The meaning of this intersection point or intersection voxel is that it is viewed by two different cameras from two different locations, and therefore the probability of the voxel or 3D scene coordinates of that point representing the object is raised.

Alternatively or in addition, the scene 360 in Figure 3 may be viewed as being divided into a voxel grid comprising eight voxels. In other words, the predefined region in space which is determined by the scene 260 is filled with voxels. For each of the eight voxels it can be determined for each

voxel within the voxel grid whether that voxel corresponds with the object or not. This can be done by mapping the 3D scene coordinates of such a voxel to all corresponding 2D image coordinates of all relevant cameras, in this case to the 2D image coordinates represented by the images 311 and 313. This mapping then occurs in the opposite direction of the arrows indicated on the projection lines 311b and 313b. Voxel 350 is the only voxel for which each of the mapped 2D image coordinates 311a and 313a correspond with 2D pixel coordinates representing the object, therefore it can be determined that voxel 350 represents the object.

The result of step 130 of Figure 1 is that the 3D scene coordinates or 3D voxel coordinates which represent the object are determined. In other words, the object 250 may be represented by a plurality of voxels, i.e. a voxel cloud, within the scene 260 of Figure 2. Such voxel clouds, however, may present a significant burden on computational resources and/or storage resources for purposes such as determining a location or position of a voxel cloud, tracking a voxel cloud or storing a voxel cloud. To lower this burden, image processing method 100 further comprises step 140 of determining 2D scene coordinates of the object 250 which are representative for the position of the object 250 within the scene, based on the determined 3D scene coordinates. Because the 3D scene coordinates of the object are available in the form of voxel clouds, the 3D scene coordinates may be transformed to generate 2D scene coordinates which represent any other view of the object in 2D. For example, when the position and orientation with respect to the scene of a camera is known, the generated camera image of that camera may be reconstructed based on the 3D scene coordinates by performing a coordinate transfer. Likewise, also so-called virtual camera images may be reconstructed, i.e. camera images that would have been taken from a different location within or outside the scene. This way, 2D scene coordinates of the object which are representative for the position of the object within the scene can be determined by performing a mapping, projection or transformation to another coordinate system. For example, if an object is expected or estimated to be located somewhere on the bottom surface of the scene, e.g. the cube 250 in scene 260, the determined 3D scene coordinates can be projected onto the bottom surface, in this case more particular the bottom plane, of the scene 260. In other words, a coordinate transfer is performed from the 3D scene coordinates X_s, Y_s, Z_s to 2D scene coordinates X_s, Y_s by projecting the 3D scene coordinates along the direction of the Z_s axis onto the bottom plane. Depending on a particular shape of the object and/or expected location of the object and/or a particular dimensions and/or shape of the scene and/or bottom plane it might be preferred to project along a different axis and/or onto a different plane. The 2D scene coordinates which are obtained this way take in less storage space, require less computational resources to process and can be tracked faster as compared to 3D volumes.

Preferably, the projected 3D scene coordinates may even be further simplified by determining a centre of the projected determined 3D scene coordinates. This way eventual required tracking resources may further be limited by narrowing down the size of the information or coordinates being representative for the position of the object. Determining a center of the projected

5 determined 3D scene coordinates may comprise determining a smaller surface within boundaries of the projected determined 3D scene coordinates which is located substantially in the middle of the projected determined 3D scene coordinates. Alternatively, determining a center of the projected determined 3D scene coordinates may comprise determining a 1D point within boundaries of the projected determined 3D scene coordinates, which 1D point is located substantially in the middle
10 of the projected determined 3D scene coordinates.

Image processing method 100 may further comprise a step of determining within the scene at least a first subspace and a second subspace. For example the scene 260 in the shape of a rectangular cuboid may be divided in a first subspace and a second subspace with equal volume, both in the
15 shape of a rectangular cuboid and being half of the rectangular cuboid 260. The first subspace is then comprised between a first bottom subplane, which in the given example coincides with the bottom plane 261, and a first upper subplane, which in the given example would be a plane parallel to planes 261 and 262 at an equal distance of both bottom plane 261 and upper plane 262. The second subspace may then be comprised between a second bottom subplane, which in the given
20 example coincides with the earlier described first upper subplane, and a second upper subplane, which in the given example coincides with the upper plane 262. After the scene has been divided this way, it can be determined for each of the first subspace and second subspace, whether determined 3D scene coordinates are present in the respective subspace. In other words, for each subspace it may be checked whether the determined 3D scene coordinates or voxels representing
25 the object are present in the respective subspace. This might be beneficial if the object is expected to have a certain height and/or orientation, such that it should be present in both subspaces. If in such a case, voxels allegedly representing the object are present in only one subspace, it may be determined that those voxels actually do not represent the object, and consequently 2D scene coordinates are not determined for those particular voxels. Depending on the expected or estimated
30 nature, size and/or shape of the object, more specific decisions can be made in determining the at least first and second subspace. In a similar way, also more than two subspaces may be determined. It is clear to the skilled person that when two or more subspaces are determined within the scene, the collection of determined subspaces may correspond to the entire scene or alternatively may only correspond to a part of the scene. It is further clear to the skilled person, that although in the
35 above example the at least first and second subspace are of equal volume, the at least first and second subspace may alternatively have mutually different volumes. In other words, the image

processing method 100 can be adapted or fine-tuned in respect of a variety of objects of which the position is to be determined or which objects are to be tracked. In exemplary embodiments an object may have different shapes at different moments in time, e.g. when the object is a human, the human may be, depending on the timing, standing up, lying down, running, sitting down, walking, jumping, etc. To take this into account the amount, shape and dimensions of determined subspaces may vary in time. In order to adequately determine the most efficient amount and size of subspaces a machine learning method, more in particular a deep learning method may be used to learn the nature of an object and to predict the particular shape and/or dimensions of the object at a particular moment in time. The determined amount and/or sizes of subspaces may also be different for each object when multiple objects are present within the scene.

To illustrate the versatility of embodiments of the image processing method according to the invention, Figures 4A and 4B illustrate a scenario wherein the scene 460 comprises a region in space which is defined by a football pitch as bottom surface 461. No upper surface of the scene is illustrated in Figures 4A and 4B, but an upper surface may for example be chosen to be a plane parallel to the football pitch at a distance of three meters or any other desirable height. In the scenario of Figures 4A and 4B, multiple moving objects 450, 451, 452, 453, 454 are located within the scene. More in particular the moving objects comprise multiple humans 450, 451, 452, 453, e.g. football players, and a ball 454. Below will be described how embodiments of the image processing method 100 may be used to determine a position for each football player 450, 451, 452, 453 on the football pitch 461 and/or to track the football players 450, 451, 452, 453 by determining their position at consecutive moments in time.

In the scenario of Figures 4A and 4B, fourteen cameras 401-414 are placed at fixed predetermined locations around the scene 460 which comprises the football pitch 461 and the space above the football pitch 461, e.g. up to three meters high. By defining the scene this high, football players 450-453 will stay completely within the scene when either jumping up or gesticulating with their arms up. In the embodiment of Figure 4A, the cameras 401-414 are arranged differently as compared to the embodiment of Figure 4B. It is however clear to the skilled person that other arrangements may be applied. During a football game typically twenty-two football players and at least one referee are simultaneously on, or close to, the football pitch 461, however for illustrative reasons only four football players 450, 451, 452, 453 are shown. In order to accurately determine positions of the twenty-two football players in a scene with the dimensions of a football pitch 461, and to accurately track the positions of these players over time, more than three cameras 401-414 are preferably used to view the scene 460. In the embodiments of Figures 4A and 4B fourteen cameras 401-414 are used, however depending on the intrinsic parameters, the resolution of the cameras, and/or the required accuracy more or less cameras may be used. It is further noted that in

the examples of Figures 4A and 4B, any set of three or more cameras of the fourteen cameras 401-414 may be used to perform step 100 of the allocation method of Figure 1. Preferably, the cameras 401-414 are set up in such a way that for every area of the football pitch 461 at least three cameras are able to view that area. In other words, preferably every area on the football pitch 361 falls within the field of view of at least three cameras, such that the method 100 as illustrated in Figure 1 can be performed for that particular area.

For reasons of accuracy, the cameras 401-414 may be set up in such a way that areas of the football pitch which might get crowded, i.e. a lot of football players flock together on a relatively small area, fall within the field of view of more than three cameras. For example, the penalty area at the left side in Figure 4A may fall within the field of view of five cameras 401, 411, 412, 413 and 414, and the penalty area at the left side in Figure 4B may fall within the field of view of five cameras 401, 402, 412, 413 and 414, such that during corner kick situations or free kicks at the edge of the penalty area when a lot of players are present within the penalty area, the position of each individual football player can be accurately determined.

According to an embodiment, in order to determine which pixels within the obtained camera images are pixels representing the football players 450, 451, 452 and 453, preprocessing steps may be performed on each of the obtained camera images. Typically the cameras 401-414 are synchronized and deliver a RGB video feed, wherein each frame can be seen to constitute a camera image. The resulting RGB camera images may not only comprise pixels representing the players 450, 451, 452 and 453 but may also comprise pixels representing part of the football pitch 461, parts of advertising boards (not shown) which are typically positioned around the football pitch 461, parts of the stadium (not shown) or even spectators (not shown) who are watching the football game. To filter out the pixels which represent the football players 450, 451, 452, 453 a background segmentation may be performed for each of the obtained camera images or video frames. This results in a black and white (BW) images wherein the football players are represented in white, whereas parts pertaining to the background are represented in black. In an embodiment, the background subtraction may be based on images of the scene, taken by the respective cameras 401-414 at an earlier moment in time. After performing the background subtraction a noise filtering may be performed. The noise filtering step may be based on a suspected or estimated size of the noise, wherein white pixels which are isolated within the camera image, or groups of white pixels with a size smaller than a predetermined threshold size, are considered to be noise and are filtered out of the camera image. In addition to, or alternative to, the noise filtering step a blob classification is carried out. Blob classification comprises comparing the shape or shapes of groups of pixels which represent the object, with predetermined shapes which correspond with the object as viewed from the particular camera providing the camera images. This way it can be determined

whether a group of pixels actually corresponds with an object, in this case a football player, corresponds with two or more football players positioned near to each other, or with the ball 454.

When the 2D pixel coordinates of the pixels representing a football player have been determined, based thereon the 3D scene coordinates or voxel coordinates may be determined. According to an embodiment the scene 460, which is the football pitch 461 and a volume of space above it, is divided in voxels, to form a voxel grid. In other words, the scene 460 is, preferably completely, filled with voxels, which may be represented as cubes. The cubes may have various dimensions, but in the scenario of Figures 4A and 4B, wherein the scene 460 is defined by means of the football pitch 461 as its bottom surface, cubes with a length of 5cm, width of 5cm and height of 5cm (5x5x5) are preferred. However, it is clear to the skilled person that cubes or any other shape of voxels may be used which have other dimensions, depending on the particular application of the image processing method according to the invention. For each of the 5x5x5 cubes within the scene it can then be determined whether that particular cube corresponds with a football player or not, by mapping the corresponding 3D voxel coordinates to the 2D pixel coordinates of the camera images obtained by the camera of which the field of view comprises the region in space which corresponds with that particular cube. If the particular cube is located close to the illustrated origin of the 3D scene coordinates, i.e. close to the corner area of the football pitch at the low left side of Figures 4A and 4B, the corresponding area may fall within the field of view of cameras 401, 413 and 414 for Figure 4A and within the field of view of cameras 401, 402 and 414 for Figure 4B. The 3D scene coordinates of that particular cube are then mapped to the 2D image coordinates of the camera images taken by cameras 401, 413 and 414 and it is checked whether the mapped 3D scene coordinates correspond with 2D pixel coordinates representing a football player or not. In a preferred embodiment, when the mapped 3D scene coordinates do not correspond with 2D pixel coordinates representing a football player in a camera image obtained by one of the cameras 401, 413 and 414, it is determined that the particular cube and its corresponding 3D scene coordinates or voxel coordinates do not represent a football player. On the other hand, if the mapped 3D scene coordinates correspond with 2D pixel coordinates representing a football player in a camera image obtained by each one of the cameras 401, 413, 414, it is determined that the particular cube and its corresponding 3D coordinates or voxel coordinates represent a football player. This process may be repeated for every voxel or cube of the voxel grid in order to create one or more voxel clouds within the 3D scene coordinate system which represent football players, and/or a ball and/or a referee.

Based on the voxel coordinates representing the football players, typically each resulting voxel cloud represents a football player, the corresponding positions of the football players 450, 451,

452, 453 can be determined. In a preferred embodiment, the positions can be determined by projecting the 3D scene coordinates to 2D scene coordinates, i.e. by projecting the voxel clouds representing the football players to the bottom surface 461. This way each voxel cloud representing a football player is transformed to a 2D area on the football pitch 461, which area is representative for the position of the respective football player. In order to avoid projecting voxels which do not represent a football player, an amount of voxels that is actually projected along a direction perpendicular to the bottom surface 461 can be taken into account and the amount of projected voxels may be compared with a threshold projection value. The threshold projection value can be predetermined based on the dimensions of the voxels and an estimate of the dimensions of a football player. For example, given the 5x5x5 voxels as described earlier, and an average estimated height of 175cm of a football player, the projections threshold may be set to 30, meaning that at least 30 voxels, representing a height of 150cm must at least be present along a projection direction in order for the voxels to result in a valid projected 2D scene coordinate. It is however clear to the skilled person, that depending on the applications of the image processing method according to the invention other projection thresholds can be set.

According to a further preferred embodiment, two or more partial projection thresholds relating to two or more subspace within the scene 460 can be applied. For example, the scene 460 may be divided into a first subspace and a second subspace, the first subspace being separated from the second subspace by a subplane which is substantially parallel to the bottom surface 361 at a height of 75cm above the bottom surface 461. The second subspace may be further delimited by an upper subplane which is substantially parallel to the bottom surface 461 at a height of 150cm above the bottom surface 461. For the first subspace a first projection threshold may be set and for the second subspace a second projection threshold may be set as described above. It may then be determined that voxels are validly projected along a direction perpendicular to the bottom surface 461, when along that direction the amount of voxels within the first subspace which are being projected is higher than the first projection threshold and the amount of voxels within the second subspace which are being projected is higher than the second projection threshold. This way, the erroneous projection and an erroneous position determination of voxels which do not represent the football players is avoided. To even further improve accuracy, additional subspaces and corresponding additional projection thresholds may be used. Moreover, deep learning methods may be used to estimate the shape of a player which may vary depending on whether the player is running, jumping or sliding. Based on the estimated shape of the player two or more subspaces may be determined within the scene.

Preferably, the obtained projected 2D area may even be further simplified by determining a centre of the projected 2D area. Determining a center of the projected 2D area may comprise determining a smaller surface within boundaries of the projected 2D area which is located substantially in the

middle of the projected 2D area. Alternatively, determining a center of the projected determined 3D scene coordinates may comprise determining a point, i.e. an area of 5x5 in the scenario where 5x5x5 voxels are used, within boundaries of the projected 2D area, which point is located substantially in the middle of the projected 2D area.

5 The above described steps of projecting the determined 3D scene coordinates of the football players to 2D scene coordinates and determining based thereon the position of the football players allows for an accurate and quick determination of the position of the football players.

In comparison to prior art methods which use an overhead camera, i.e. a camera which is positioned directly above the scene, to determine positions of players, the described method is more accurate and less susceptible to noise since data corresponding to the Z_s direction is taken into account according to the described method, whereas this information is not available to the overhead camera. Moreover, the described method is more flexible as compared to the use of an overhead camera since an overhead camera cannot be set up at every scene, e.g. a lot of football pitches do not offer the infrastructure to position a camera directly above the football pitch, whereas cameras
15 can be easily set-up around the football pitch.

By carrying out the described method for subsequent video frames of the obtained video feed, at subsequent moments in time, the position of the football players can be accurately tracked over time.

20

Figure 5 schematically illustrates an image processing system 500 for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane. The image processing system 500 comprises an obtaining unit 510 for obtaining at least three

25 contemporaneous camera images of the object within the scene, wherein the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene. The obtaining unit 510 may be comprised within one or more cameras which capture the at least three contemporaneous camera images. Alternatively or in addition, the obtaining unit 510 may be located externally to one or more of those cameras. The system 500 further comprises
30 a first determining unit 520 for determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images, a second determining unit 530 for determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D image coordinates, and a third determining unit 540 for determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined
35 3D scene coordinates.

A person of skill in the art would readily recognize that steps of various above-described methods can be performed by programmed computers. Herein, some embodiments are also intended to cover program storage devices, e.g., digital data storage media, which are machine or computer readable and encode machine-executable or computer-executable programs of instructions, wherein said instructions perform some or all of the steps of said above-described methods. The program storage devices may be, e.g., digital memories, magnetic storage media such as a magnetic disks and magnetic tapes, hard drives, or optically readable digital data storage media. The embodiments are also intended to cover computers programmed to perform said steps of the above-described methods.

The functions of the various elements shown in the Figures, including any functional blocks labelled as “units”, “processors” or “modules”, may be provided through the use of dedicated hardware as well as hardware capable of executing software in association with appropriate software. When provided by a processor, the functions may be provided by a single dedicated processor, by a single shared processor, or by a plurality of individual processors, some of which may be shared. Moreover, explicit use of the term “processor” or “controller” should not be construed to refer exclusively to hardware capable of executing software, and may implicitly include, without limitation, digital signal processor (DSP) hardware, network processor, application specific integrated circuit (ASIC), field programmable gate array (FPGA), read only memory (ROM) for storing software, random access memory (RAM), and non volatile storage. Other hardware, conventional and/or custom, may also be included. Similarly, any switches shown in the Figures are conceptual only. Their function may be carried out through the operation of program logic, through dedicated logic, through the interaction of program control and dedicated logic, or even manually, the particular technique being selectable by the implementer as more specifically understood from the context.

It should be appreciated by those skilled in the art that any block diagrams herein represent conceptual views of illustrative circuitry embodying the principles of the invention. Similarly, it will be appreciated that any flow charts, flow diagrams, state transition diagrams, pseudo code, and the like represent various processes which may be substantially represented in computer readable medium and so executed by a computer or processor, whether or not such computer or processor is explicitly shown.

Whilst the principles of the invention have been set out above in connection with specific embodiments, it is to be understood that this description is merely made by way of example and not as a limitation of the scope of protection which is determined by the appended claims.

Claims

1. Image processing method for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane, the method comprising:

- obtaining at least three contemporaneous camera images of the object within the scene, wherein the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene;

- determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images;

- determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D image coordinates;

- determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined 3D scene coordinates.

2. Image processing method according to claim 1, wherein

- determining 2D image coordinates of the object comprises determining 2D pixel coordinates for each pixel representing the object; and

- determining 3D scene coordinates of the object comprises determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates.

3. Image processing method according to claim 2, wherein determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates comprises mapping the 2D pixel coordinates to 3D voxel coordinates.

4. Image processing method according to claim 3, wherein mapping the 2D pixel coordinates to 3D voxel coordinates comprises:

- determining for each pixel representing the object in a first camera image a set of candidate voxels within the scene which correspond with said respective pixel;

- selecting from said set of candidate voxels a voxel representing the object, based on sets of candidate voxels corresponding to pixels representing the object in at least one camera image other than said first camera image.

5. Image processing method according to claim 4, wherein determining the set of candidate voxels for said respective pixel comprises determining a pixel projection line from said respective

pixel to the scene and determining that a voxel is a candidate voxel when said voxel is situated substantially along said pixel projection line.

6. Image processing method according to claim 2, wherein determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates comprises dividing the scene into a voxel grid comprising a plurality of voxels, and determining for each voxel of the plurality of voxels whether said voxel represents the object based on the determined 2D pixel coordinates.

7. Image processing method according to claim 6, wherein determining for each voxel of the plurality of voxels whether said voxel represents the object based on the determined 2D pixel coordinates comprises mapping corresponding 3D voxel coordinates of said voxel to mapped 2D image coordinates for each one of the at least three camera images, and determining whether said mapped 2D image coordinates correspond with 2D pixel coordinates representing the object.

8. Image processing method according to any one of the preceding claims, further comprising dividing the scene in a plurality of subspaces, wherein each subspace of the plurality of subspaces is comprised between a bottom subplane and an upper subplane, which subplanes are substantially parallel with the bottom plane and the upper plane, and wherein determining 3D voxel coordinates for each voxel representing the object based on the determined 2D pixel coordinates is executed consecutively for each subspace of the plurality of subspaces or in parallel for each subspace of the plurality of subspace.

9. Image processing method according to any one of the preceding claims, wherein determining 2D scene coordinates of the object which are representative for the position of the object within the scene comprises projecting the determined 3D scene coordinates onto the bottom plane.

10. Image processing method according to claim 9, wherein determining 2D scene coordinates of the object which are representative for the position of the object within the scene comprises determining a center of the projected determined 3D scene coordinates.

11. Image processing method according to any one of the preceding claims, further comprising:

- dividing the scene or a part thereof in at least a first subspace and a second subspace, wherein the first subspace is comprised between a first bottom subplane and a first upper subplane,

and wherein the second subspace is comprised between a second bottom subplane and a second upper subplane, which subplanes are substantially parallel with the bottom plane and the upper plane;

- determining for each of the at least first subspace and second subspace, whether
5 determined 3D scene coordinates are present in the respective subspace; and
- determining 2D scene coordinates of the object which are representative for the position of the object within the scene if determined 3D scene coordinates are present in each subspace of the at least first subspace and second subspace.

10 12. Image processing method according to claim 11, wherein the first bottom subplane coincides with the bottom plane and wherein the first upper subplane coincides with the second bottom subplane.

15 13. Image processing method according to any one of the preceding claims, wherein each predetermined camera location and orientation corresponds with a predetermined position and orientation with respect to a scene based coordinate system, respectively.

20 14. Image processing method according to any one of the preceding claims, wherein obtaining each one of the at least three contemporaneous camera images of the object within the scene is performed by at least three corresponding cameras.

25 15. Image processing method according to claim 10, further comprising a step of calibrating the at least three cameras, prior to obtaining the at least three contemporaneous camera images of the object within the scene.

16. Image processing method according to claim 15, wherein calibrating the at least three cameras comprises minimizing a reprojection error of the cameras.

30 17. Image processing method according to any one of the preceding claims, wherein determining 2D image coordinates of the object for each one of the at least three camera images comprises determining which pixels represent the object in each one of the at least three camera images.

35 18. Image processing method according to claim 17, wherein determining which pixels represent the object in each one of the at least three camera images comprises performing a background subtraction for each one of the at least three camera images.

19. Image processing method according to claim 17 or 18, wherein determining which pixels represent the object in each one of the at least three camera images comprises at least one of noise filtering and shadow filtering based on a predetermined threshold.

5

20. Image processing method according to claim 17, 18 or 19, wherein determining which pixels represent the object in each one of the at least three camera images comprises a blob classification based on an estimated shape of the object.

10

21. Image processing method for tracking a moving object within a scene, comprising:

- performing the method according to any one of the preceding claims 1-20 at consecutive moments in time.

15

22. Image processing method according to claim 21, wherein the at least three camera images constitute at least three video frames of at least three corresponding video feeds.

20

23. Image processing method according to any one of the preceding claims 11 to 22, wherein the object is a human and wherein the first subspace is between 0 cm and 120 cm, preferably between 0 cm and 100 cm, and more preferably between 0 cm and 75 cm; and wherein the second subspace is between 75 cm and 200 cm, preferably between 95 cm and 185 cm, and more preferably between 115 cm and 170 cm.

25

24. A computer program product comprising computer-executable instructions for performing the method of any one of the claims 1-23, when the program is run on a computer.

30

25. Image processing system for determining a position of an object within a scene, wherein the scene is a predetermined region in space, said region being defined between a bottom plane and an upper plane substantially parallel to the bottom plane, the system comprising:

- an obtaining unit for obtaining at least three contemporaneous camera images of the object within the scene, wherein the at least three camera images are associated with different predetermined camera locations and orientations with regard to the scene;

- a first determining unit for determining two-dimensional, 2D, image coordinates of the object for each one of the at least three camera images;

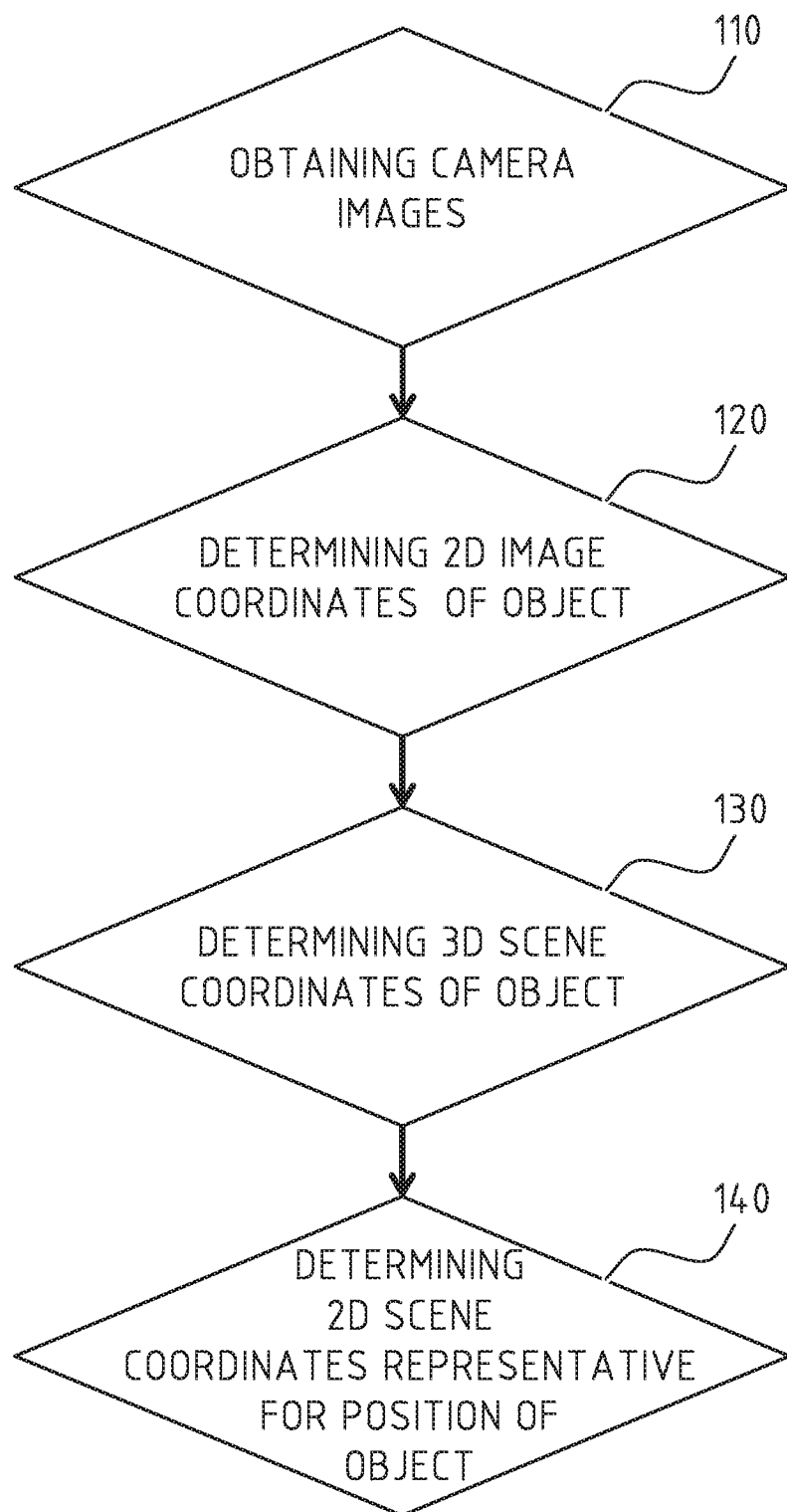
35

- a second determining unit for determining three-dimensional, 3D, scene coordinates of the object based on the determined 2D image coordinates; and

- a third determining unit for determining 2D scene coordinates of the object which are representative for the position of the object within the scene, based on the determined 3D scene coordinates.

- 5 26. Image processing system according to claim 25, configured for performing the method of any of the claims 1-23.

1/6

FIG. 1

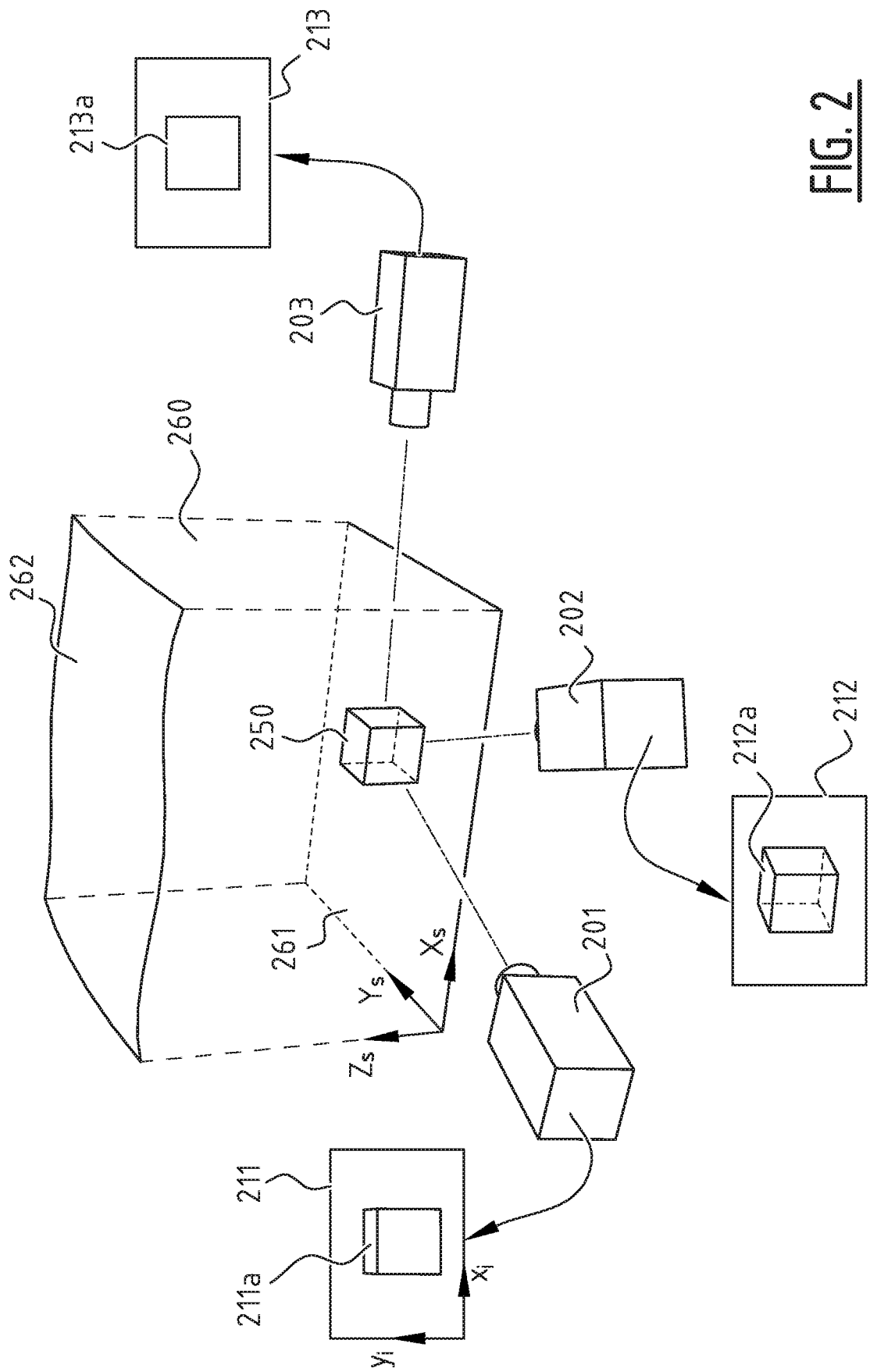


FIG. 2

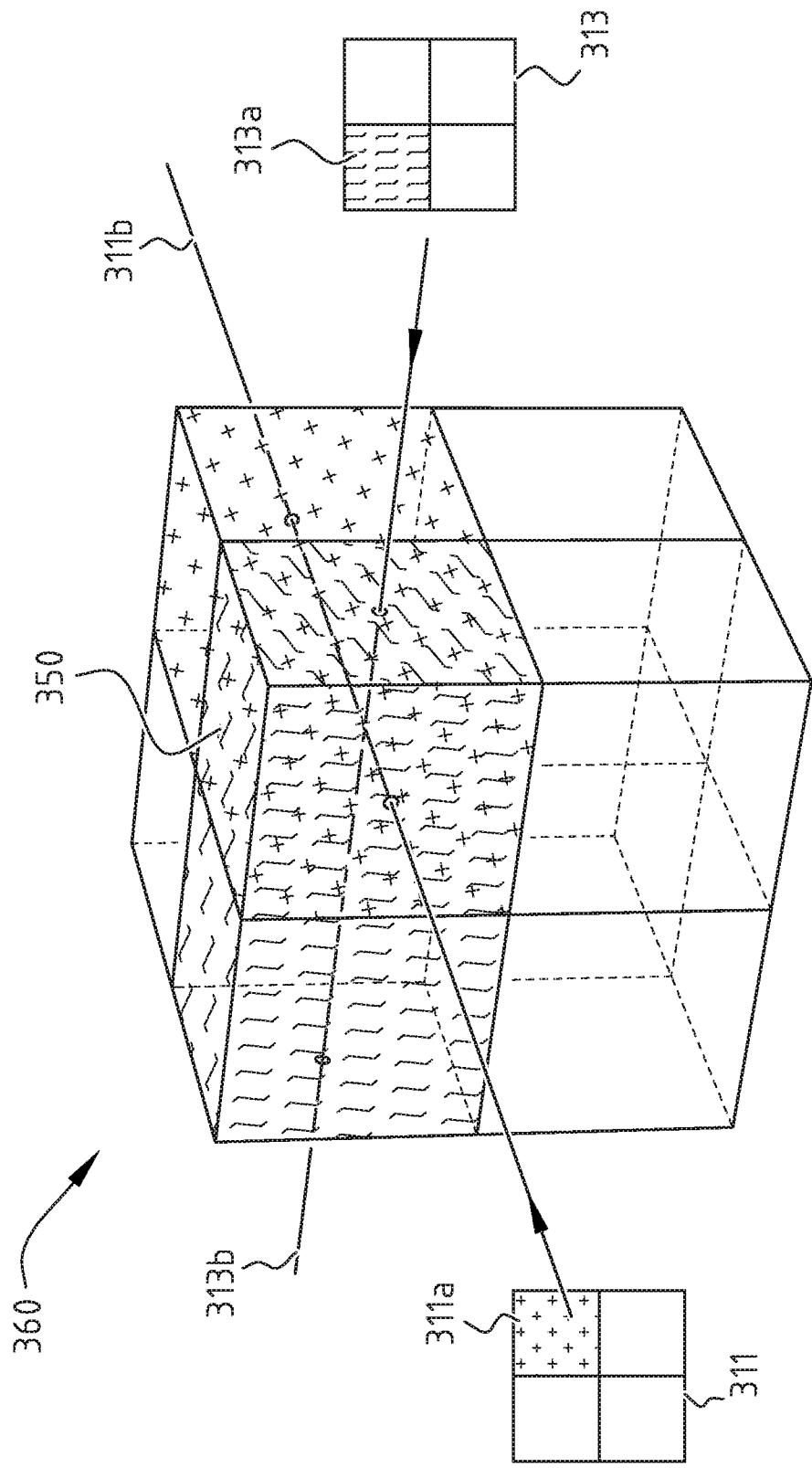


FIG. 3

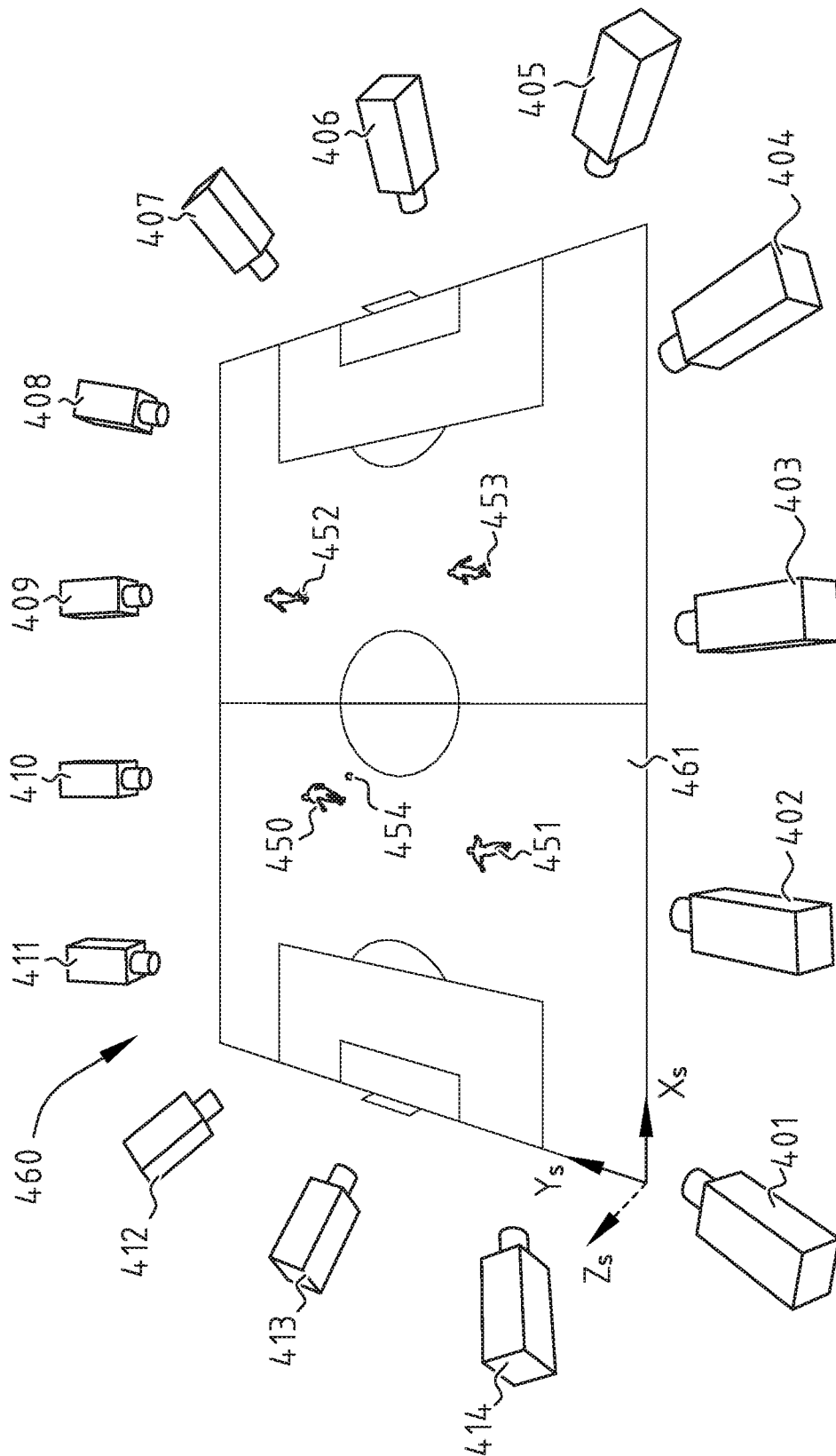


FIG. 4A

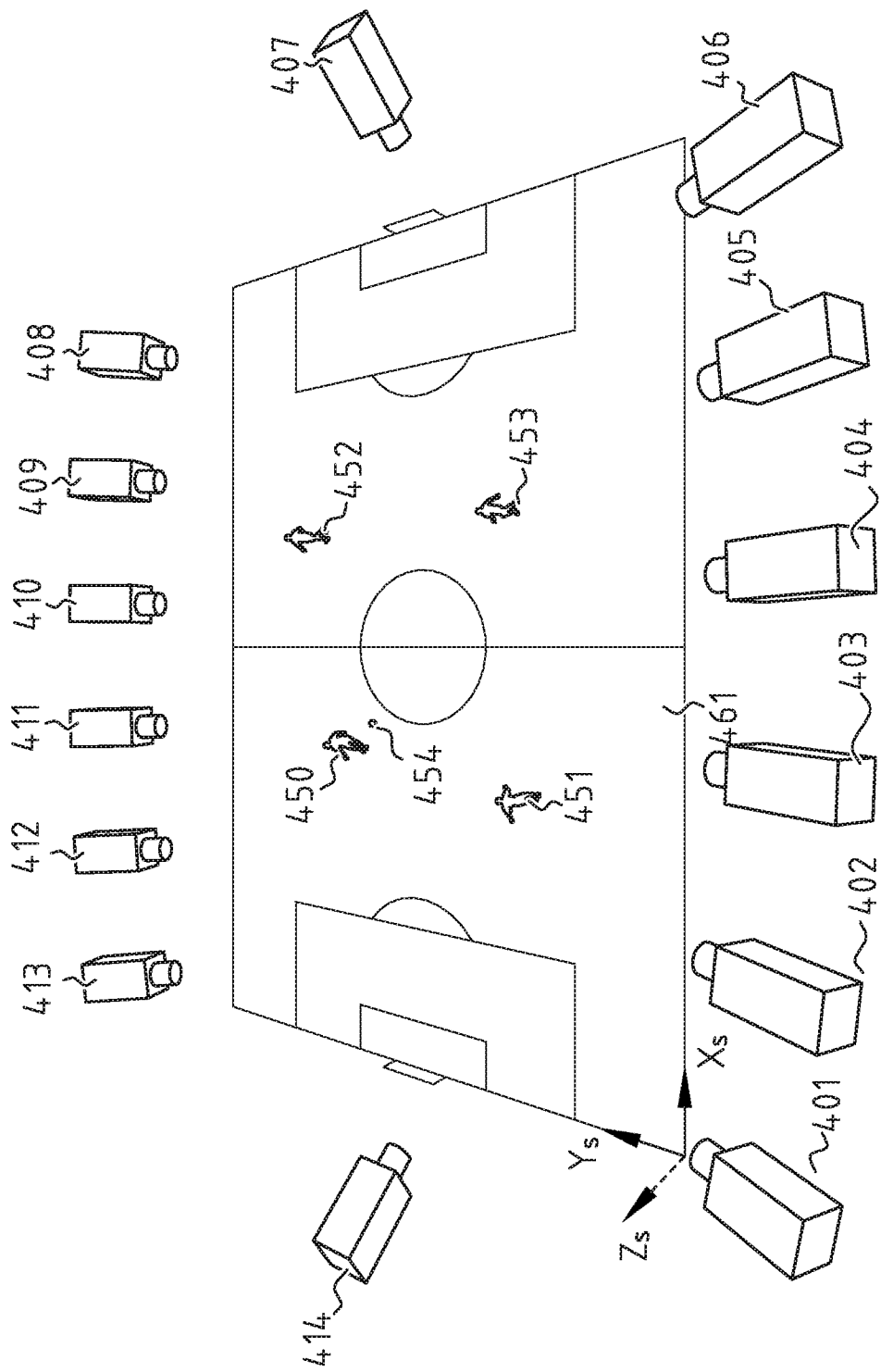
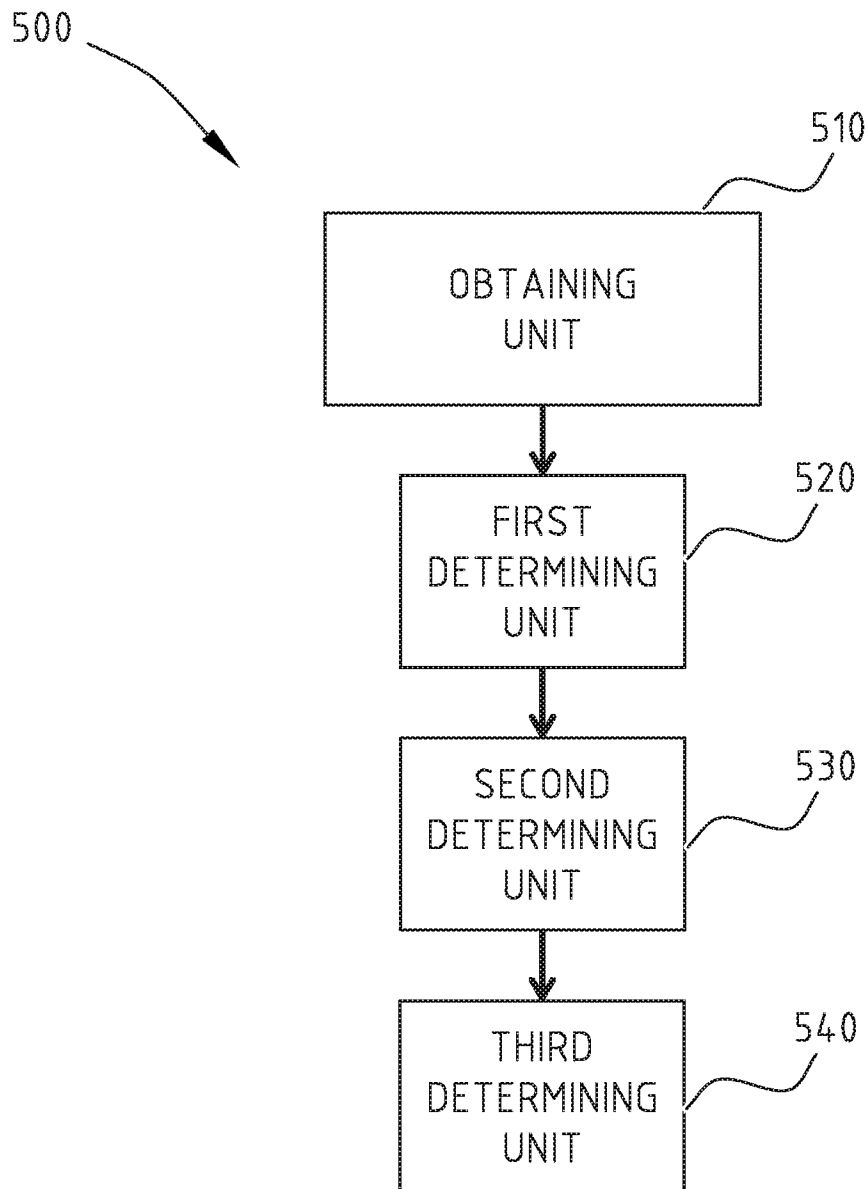


FIG. 4B

6/6

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/NL2018/050349

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06T7/70 G06T7/564
ADD.

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)
G06T

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	ANDERSEN M ET AL: "Three-dimensional adaptive sensing of people in a multi-camera setup", 2006 14TH EUROPEAN SIGNAL PROCESSING CONFERENCE, IEEE, 23 August 2010 (2010-08-23), pages 964-968, XP032770492, ISSN: 2219-5491 [retrieved on 2015-04-27]	1,2, 6-10, 13-22, 24-26
A	sections 2.1, 2.1.2-2.1.2; figure 1 section 2.3 section 2.3.1, last paragraph; figure 5 section 2.3.2, paragraph 1 section 3, first and last paragraphs ----- -/--	11,12,23



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

17 July 2018

Date of mailing of the international search report

25/07/2018

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

Kollreider, Klaus

INTERNATIONAL SEARCH REPORT

International application No

PCT/NL2018/050349

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ZABULIS X ET AL: "A Platform for Monitoring Aspects of Human Presence in Real-Time", 1 December 2010 (2010-12-01), NETWORK AND PARALLEL COMPUTING; [LECTURE NOTES IN COMPUTER SCIENCE; LECT.NOTES COMPUTER], SPRINGER INTERNATIONAL PUBLISHING, CHAM, PAGE(S) 584 - 595, XP047441621, ISSN: 0302-9743 ISBN: 978-3-642-01969-2	1,2, 6-10, 13-18, 20-22, 24-26
A	section 3.1 sections 4.1-4.2; figures 1-2	11,12,23
X	----- VAN HAMME DAVID ET AL: "Parameter-unaware autocalibration for occupancy Mapping", 2013 SEVENTH INTERNATIONAL CONFERENCE ON DISTRIBUTED SMART CAMERAS (ICDSC), IEEE, 29 October 2013 (2013-10-29), pages 1-7, XP032580815, DOI: 10.1109/ICDSC.2013.6778205	1-5,13, 14,18, 20,22, 24-26
A	section IV.B, paragraph 3; figure 8	11,12,23
X	----- ARSIC DEJAN ET AL: "Real Time Person Tracking and Behavior Interpretation in Multi Camera Scenarios Applying Homography and Coupled HMMs", 7 September 2010 (2010-09-07), NETWORK AND PARALLEL COMPUTING; [LECTURE NOTES IN COMPUTER SCIENCE; LECT.NOTES COMPUTER], SPRINGER INTERNATIONAL PUBLISHING, CHAM, PAGE(S) 1 - 18, XP047373356, ISSN: 0302-9743 ISBN: 978-3-642-01969-2	1-5,9, 13,14, 17,18, 20-22, 24-26
A	section 1, paragraph 2, last sentence section 2.3; figure 6 section 3.1, paragraph 1 -----	11,12,23