



(51) International Patent Classification:

G06T 7/13 (2017.01) G06T 17/00 (2006.01)

G06T 7/136 (2017.01) G06T 7/55 (2017.01)

G06T 7/194 (2017.01) H04N 13/10 (2018.01)

G06T 15/00 (2011.01)

(21) International Application Number:

PCT/EP2022/064243

(22) International Filing Date:

25 May 2022 (25.05.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

21177608.3 03 June 2021 (03.06.2021) EP

(71) Applicant: KONINKLIJKE PHILIPS N.V. [NL/NL];
High Tech Campus 52, 5656 AG Eindhoven (NL).

(72) Inventor: VAREKAMP, Christiaan; c/o Philips International B.V. Intellectual Property and Standards, High Tech Campus 52, 5656 AG Eindhoven (NL).

(74) Agent: PHILIPS INTELLECTUAL PROPERTY & STANDARDS; High Tech Campus 52, 5656 AG Eindhoven (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, IQ, IR, IS, IT, JM, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

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

Declarations under Rule 4.17:

— as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

(54) Title: DEPTH SEGMENTATION IN MULTI-VIEW VIDEOS

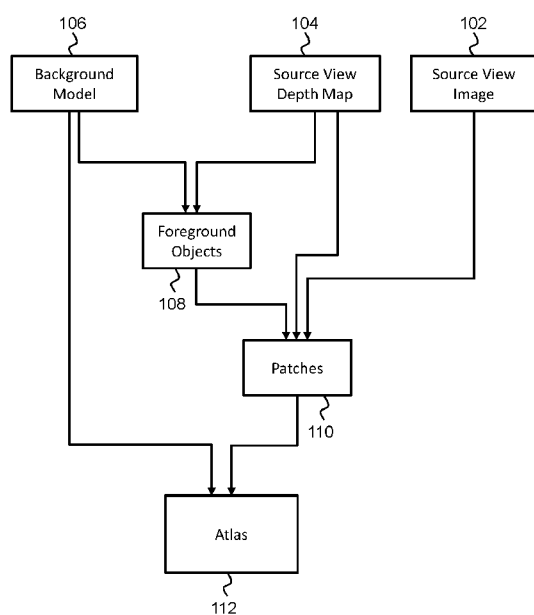


FIG. 1

(57) Abstract: A method of depth segmentation for the generation of a multi-view video data. The method comprises obtaining a plurality of source view images and source view depth maps representative of a 3D scene from a plurality of sensors. Foreground objects in the 3D scene are segmented from the source view images and/or the source view depth maps. One or more patches are then generated for each source view image and source view depth map containing at least one foreground object, wherein each patch corresponds to a foreground object and wherein generating a patch comprises generating a patch texture image, a patch depth map and a patch transparency map based on the source view images and the source view depth maps.

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

DEPTH SEGMENTATION IN MULTI-VIEW VIDEOS

FIELD OF THE INVENTION

The invention relates to the field of multi-view videos. In particular, the invention relates to depth segmentation for the generation of multi-view videos and rendering multi-view videos.

5 BACKGROUND OF THE INVENTION

Existing approaches that render from multi-view image with depth combine the warped textures from multiple source view (capture) cameras using blending. The blending operation can depend on variables such as source and target camera position/orientation (e.g. ray angle differences), depth magnitude, depth variation, de-occlusion, transparency and color. More advanced techniques even use a
10 trained convolutional neural network to align textures in the target viewpoint. There are several formats for storing multi-view images.

Layered Depth Images (LDI) store a set of depth pixels (not just one) along a single line of sight. When the virtual viewpoint moves away from the LDI storage viewpoint then the occluded surfaces become visible.

15 Multi Plane Image (MPI) and Multi Sphere Image (MSI) techniques construct color and transparency for a predefined set of planes or spheres in 3D space. For a new virtual viewpoint, the image is then constructed using back-to-front over-compositing of the layers.

Layered Meshes (LM) can be constructed from MPI and MSI and represent traditional graphics meshes with texture and are hence suitable for atlas construction and transmission using existing
20 video codecs.

While the layered formats (LDI, MPI, MSI, LM) can potentially bring a larger viewing zone due to the explicit occlusion handling, these formats are difficult to produce, especially in real-time, from a multi-camera system.

Loghman Maziar et al. "Segmentation-based view synthesis for multi-view video plus
25 depth", Multimedia Tools and Applications, Kluwer Academy Publishers Boston vol. 74, no. 5, 8 November 2013 discloses a method for image synthesis by segmenting objects from source images and warping the segmented objects individually.

SUMMARY OF THE INVENTION

30 The invention is defined by the claims.

According to examples in accordance with an aspect of the invention, there is provided a method of depth segmentation for the generation of a multi-view video data, the method comprising:

obtaining a plurality of source view images and source view depth maps representative of a 3D scene from a plurality of sensors;

segmenting foreground objects in the 3D scene from the source view images and/or the source view depth maps; and

5 generating one or more patches for each source view image and source view depth map containing at least one foreground object, wherein each patch corresponds to a foreground object and wherein generating a patch comprises generating a patch texture image, a patch depth map and a patch transparency map based on the source view images and the source view depth maps.

10 Typical formats for multi-view video require significant processing power to generate (e.g. layered depth images, multi-plane images etc.) due to the complex analysis that is typically required to estimate a depth value for each pixel. For instance, the inability of finding a robust algorithm to do this has resulted in increased use of data driven approaches based on deep learning. This problem is particularly present in the broadcast of multi-view videos such as live sporting events, as the creation of depth and texture atlas data, for data reduction, must be done in real time for each frame.

15 Thus, the inventor has proposed to “segment” patches from the source views (i.e. images and depth maps) containing the foreground objects. Thus, an atlas would only contain data from patches (and from the background) instead of data for all of the source views. Each patch is based on a section of a source view image which is smaller than the source view image itself and the corresponding depth and transparency data for the section. In other words, each patch functions as a partial source view with
20 texture, transparency and depth data corresponding to a foreground object used to render a scene instead of using source views corresponding to arbitrary parts of the scene. Patches may overlap one another. Different patches may have the same size or they may have different sizes (e.g. depending on their corresponding foreground objects). In some cases, different patches may have identical patch texture images and patch depth maps if, for example, they are based on the same section of a source view image
25 (e.g. when a particular section includes more than one foreground object).

Various methods exist for segmenting the foreground objects from either the source view images or the source view depth maps. For example, segmentation algorithms or object detection algorithms may be used on the source view images to detect/segment the foreground objects. Alternatively, depth differences (above a threshold) in the source view depth maps may be calculated
30 such that the edge of a foreground object can be defined by the large depth differences. For multi-view video, the difference between frames (of either the source view images or source view depth maps) can be used to detect movement of a foreground object and thus detect the foreground object.

Patches can thus be generated for the foreground objects in each source view. Each patch corresponds to a single foreground object. In particular, the extent of the patch may be determined, at
35 least in part, by the extent of the object (or a part of the object). However, a foreground object may have more than one corresponding patch. Alternatively, each foreground object may be associated with a single patch. The patches have a patch texture image which includes the texture/color data for the patch and a

patch depth map which includes the depth data for the patch. The patches also have a patch transparency map (also called a patch alpha map) which includes the transparency values of each pixel in the patch.

The method may further comprise obtaining a plurality of background depth maps of the 3D scene representative of the background of the 3D scene, a background depth map containing depth data of the background from a particular orientation, wherein segmenting foreground objects is based on the difference between a background depth map and a corresponding source view depth map.

When segmenting based on, for example, source view images by themselves, it is likely that the patches generated would bleed across depth boundaries. This could cause problems in the encoding or later rendering. Thus, by using the background depth maps, the foreground object can be robustly segmented and these problems may be avoided.

The background depth map for each source view can be obtained by fitting a pre-defined geometric scene model to a subset of the source views. For example, assuming that the background consists of a horizontal ground surface plane and a vertical background plane, these planes can be initially placed and shifted/rotated with respect to each other and the cameras such that an image based multi-view matching criterion is minimized. After fitting the pre-defined geometric scene to a subset of the source views, a background depth map can be rendered for each source view.

The background depth map comprises depth data of the background of the 3D scene. For example, the background depth map may be generated based on the views of a plurality of cameras imaging the 3D scene from different angles. The background depth map may be generated from a different set cameras than the ones used to obtain the source views. For example, if the 3D scene is a soccer field, cameras on the side of the soccer field may be used to image the foreground objects (i.e. the players and the ball) and cameras viewing the soccer field from the top (e.g. top down cameras) may be used to generate the background depth map.

The foreground objects can be segmented by thresholding the difference between the source view depth maps and the background depth map for each source view. After this global threshold, a second, local, thresholding may be applied to separate connected foreground objects based on a relative depth step.

A trained human person detection algorithm can be used to detect foreground objects. A ball detector can be used to detect the ball in a sports game. Motion estimation or temporal frame differencing can be used to further improve foreground object detection.

The method may further comprise obtaining a background model comprising the background depth map and background texture data.

The method may further comprise generating an atlas based on the patch texture images, the patch depth maps, the patch transparency maps and the background model. For example, the atlas may contain the patch texture images, the patch depth maps, the patch transparency maps and the background model.

An atlas is, in essence, a data matrix which contains various images and/or maps (e.g. texture, depth and transparency data). In order to find an image or map in the atlas, the “coordinates” of each image are specified (i.e. the column and row values for the matrix). Thus, the atlas contains data from multiple source views.

5 Typically, the patch data is all separately contained in the atlas. However, it is also possible to define, for example, the patch transparency maps in binary (i.e. transparency values of zero or one) and coded via a reserved value in the depth map.

10 Detecting foreground objects may comprise subtracting the respective background depth maps from the source view depth maps, to produce difference images, and thresholding the difference images, wherein thresholding comprises comparing the pixel values of the difference images with a threshold value, to produce threshold maps, thereby to differentiate between the background and the foreground objects. Pixels in the source view depth maps corresponding to depth steps are identified, wherein depth steps are defined by the differences between neighboring depth values in a source view depth map being larger than a depth threshold value. All the depth values corresponding to the depth steps
15 are adjusted in the threshold maps thereby to differentiate the foreground objects from each other and bounding boxes are generated for the foreground objects based on the adjusted threshold maps.

Thresholding the difference image may result in a binary map where the pixel value “1” means foreground and “0” means background. To identify foreground objects, connected components are identified via a 4-connected or 8-connected component labelling algorithm. Doing this immediately after
20 the initial thresholding operation would result in multiple foreground objects being falsely identified as a single object. To avoid this, the spatial derivative of the original source view depth map, for example, is analyzed. When a depth step exceeds a “depth threshold”, then the binary map is set to “0” (i.e. background) on the further side of the step. When the resulting binary map is input to the connected component labelling algorithm, then the foreground objects can receive different labels.

25 The size and position of a patch may be based on the size and position of a bounding box.

The depth values of the background depth map are subtracted from the source view depth maps in order to make all values zero (or close to zero) apart from the foreground objects present in the source view depth maps. The subtracted maps are then thresholded based on a “threshold value” in order to, for example, set all of the depth values corresponding to the background to zero (or black) and all of
30 the depth values corresponding to the foreground objects to one (or white).

Depth steps are also identified in the source view depth maps. The depth steps correspond to large changes in depth for adjacent/neighboring pixels which indicate the edges of the foreground objects. The depth steps can be identified by the difference between neighboring depth values being larger than a depth threshold value (e.g. larger than 0.1 in a normalized depth map).

35 The depth values of the threshold maps can then be adjusted at the depth steps to be, for example, zero (or black) in order to highlight and distinguish the edges of each foreground object. A

bounding box is generated for each foreground object based on the adjusted threshold maps (e.g. segmenting the foreground objects in the adjusted threshold maps).

The size and position of a patch could be the size and position of the bounding box. Alternatively, multiple patches could be generated for a bounding box. For example, the number of patches per bounding box may depend on the size of the bounding box, the type of foreground object, the position of the foreground object etc.

Detecting foreground objects may further comprise extending the bounding boxes thereby to include regions of the foreground objects in the subtracted maps below the threshold value.

For example, extending the bounding boxes may be based on the difference between the source view depth map and the background depth map being less than the threshold value for regions which include a foreground object, wherein the bounding boxes are extended such that each foreground object is enclosed by a bounding box.

In some instances, the foreground object may have parts which have a depth value similar to the depth value of the background. Thus, during the thresholding, the foreground object will look smaller and the bounding box may not fully enclose the foreground object in the source views.

For example, the feet of a soccer player have a similar depth to the soccer field they are standing on. In these cases, the bounding boxes are extended (for example, extended downwards) such that the bounding box(es) corresponding to the foreground object fully enclose the foreground object.

Generating a patch texture image and a patch transparency map may be based on alpha matting the source view images. Alpha matting is based on extracting the foreground from an image. Thus, the texture and transparency (alpha values) of each pixel of a patch can be estimated using alpha matting.

The method may further comprise adapting pixel depth values of the patch depth maps such that all of the pixel depth values of a patch depth map consist of values equal to or lower than the depth values of the corresponding foreground object.

For the sake of consistency and clarity, any depth maps defined in this application will be constructed such that a maximum value (e.g. 255) represents the closest distance to a viewpoint (i.e. smallest depth value) and a minimum value (e.g. 0) represent the furthest distance (i.e. highest depth value). Any mention of “lower” or “higher” in this application with respect to the value of the pixels in a depth map should be interpreted with respect to the aforementioned definition. However, it must be noted that any other format of representing depth maps could also be used and will be known to the person skilled in the art. For example, a “0” pixel may represent the closest distance and a “1” value may represent the furthest.

Some patch depth maps may contain depth data from other foreground objects occluding the corresponding foreground object. The “unwanted” (or left-over) depth data may cause artefacts when rendering the foreground objects. Thus, it may be beneficial to adapt the pixel depth values of the patch depth map (i.e. change the pixel values) such that all of the pixel depth values are equal to or higher than

the depth values of the target foreground object (i.e. the foreground object corresponding to the patch depth map in question).

Additionally, the method may further comprise identifying a plurality of patches originating from different source views corresponding to a first foreground object based on identifying the patch depth maps of the identified patches within an object depth range and correcting the identified patch depth maps such that they correspond to an object location in the 3D scene.

For example, the patch depth maps can be corrected (e.g. filtered) by projecting the centroid locations of the patch depth maps of all views to a common world coordinate system. Patches from different source views that map to similar world space coordinates (i.e. within a given inter-object distance) likely originate from one and the same physical foreground object. Patch depth maps can hence be corrected (i.e. made to have more similar world space coordinate). After correcting, a back-projection to the source view results in a filtered depth map per patch.

The method may further comprise pruning the patches based on measuring the consistency between patches in multiple source views. For example, the method may comprise filtering a particular patch to possibly remove it in case there are not enough corresponding patches in other source views (indicating that the patch is an isolated error likely as a result of estimation noise).

This may help in identifying falsely detected foreground patches. For example, after projecting a patch to a common world coordinate system the number of patches from other source views closer than a minimum world-space (Euclidean) distance of the patch may be calculated. If this number is lower than a patch number threshold (e.g. a given fraction of the number of source views) then the patch is discarded. For example, if a “foreground object” is only identified in less than three of eight source views, the patches for that particular foreground object are discarded. If the patch is discarded it will not be used in the atlas.

Generating one or more patches may comprise identifying a sub-region in a source view depth map, determining a number of depth surfaces of different depths present in the sub-region and generating a patch for each depth surface in the sub-region, wherein each patch comprises a different patch transparency map.

Alternatively or additionally, generating one or more patches may comprise identifying a sub-region in a source view image.

The invention also provides a system comprising:

one or more processors comprising computer program code which, when executed on a computing device, cause the computing system to perform the aforementioned method; and a plurality of sensors configured to obtain the source view images and source view depth maps.

The invention also provides a method for rendering multi-view videos, the method comprising:

receiving an atlas with a plurality of patches and a background model of a 3D scene, wherein each patch corresponds to a foreground object and wherein each patch comprises a patch texture image, a patch depth map and a patch transparency map derived from source view images and source view depth maps;

- 5 receiving a virtual viewpoint within the 3D scene;
sorting the patches based on the difference between the position of the virtual viewpoint and the position of the foreground objects corresponding to each patch; and
rendering the background model and the sorted patches.

10 The rendering method may further comprise grouping the patches based on the position of the corresponding foreground objects relative to the virtual viewpoint.

Rendering the background model and the sorted patches may comprise rendering the background model, warping and/or blending a first patch group, compositing the warped and/or blended first patch group onto the rendered background model, warping and/or blending a second patch group, wherein the position of the foreground objects corresponding to the second patch group relative to the virtual viewpoint is closer to the virtual viewpoint than the position of the foreground objects
15 corresponding to the first patch group and compositing the warped and/or blended second patch group onto the warped and/or blended first patch group.

The method may further comprise receiving metadata comprising the position and geometry of each patch in the atlas and the position and geometry of each patch in a source view image
20 and/or a source view depth map, wherein rendering the patches is based on both the positions and geometries.

The invention also provides a computer program product comprising computer program code which, when executed on a computing device having a processing system, cause the processing system to perform all of the steps of the method of depth segmentation for the generation of a multi-view
25 video data and/or the method for rendering multi-view videos and a processor configured to execute the computer program code.

These and other aspects of the invention will be apparent from and elucidated with reference to the embodiment(s) described hereinafter.

30 BRIEF DESCRIPTION OF THE DRAWINGS

For a better understanding of the invention, and to show more clearly how it may be carried into effect, reference will now be made, by way of example only, to the accompanying drawings, in which:

- 35 Fig. 1 shows a method of depth segmentation for multi-view videos;
Fig. 2 shows a source view image and a source view depth map;
Fig. 3 shows a patch corresponding to the triangular object in Fig. 2;
Fig. 4 illustrates a first process for generating patches;

Fig. 5 shows a source view depth map with three foreground objects;
Fig. 6 shows four patch transparency maps generated for a single region;
Fig. 7 illustrates the step of warping and blending a patch group from the atlas; and
Fig. 8 shows a 3D scene with two source views and two target views.

5

DETAILED DESCRIPTION OF THE EMBODIMENTS

The invention will be described with reference to the Figures.

It should be understood that the detailed description and specific examples, while indicating exemplary embodiments of the apparatus, systems and methods, are intended for purposes of illustration only and are not intended to limit the scope of the invention. These and other features, aspects, and advantages of the apparatus, systems and methods of the present invention will become better understood from the following description, appended claims, and accompanying drawings. It should be understood that the Figures are merely schematic and are not drawn to scale. It should also be understood that the same reference numerals are used throughout the Figures to indicate the same or similar parts.

15 The invention provides a method of depth segmentation for the generation of a multi-view video data. The method comprises obtaining a plurality of source view images and source view depth maps representative of a 3D scene from a plurality of sensors. Foreground objects in the 3D scene are segmented from the source view images (102) and/or the source view depth maps (104). One or more patches are then generated for each source view image and source view depth map containing at least one foreground object, wherein each patch corresponds to a foreground object and wherein generating a patch comprises generating a patch texture image, a patch depth map and a patch transparency map based on the source view images and the source view depth maps.

Fig. 1 shows a method of depth segmentation for multi-view videos. The method is based on creating an atlas 112 containing all the data necessary for rendering a multi-view video. Multiple source views are typically necessary to render a multi-view image frame. Each source view will typically have a source view image 102 and a source view depth map 104. A depth map 104 may be derived using multi-view image based matching or alternatively by adding one or more depth sensors (e.g. laser depth sensors) to the multi-camera setup (used to obtain the source view data) and then warping the measured depth in each source view after which a filtering/hole filling can be used to make the depth maps complete for each source view.

The inventor proposes to segment all source view depth maps 104 based on the difference with a globally determined background model 106. The background model 106 is used to generate a background depth map for each source view. Foreground objects 108 are then further segmented based on the relative depth differences between pixels of the background depth map and the source view depth maps 104. Instead of producing a single layered representation, segmented patches 110 of the foreground objects 108 are kept for all source views and packed together in an atlas 112.

35

A client device can sort the patches 110 along the z-axis of a new virtual viewpoint. The view synthesis algorithm can then visit patches 110 in this order and alternate between blending patches 110 from different source views when these patches 110 have similar depth and compositing the blended view over the previous composited output.

Fig. 2 shows a source view image 102 and a source view depth map 104. Fig. 2 (a) shows the source view image 102 and Fig. 2 (b) shows the source view depth map 104. The source view image 102 and the source view depth map 104 are of a 3D scene containing a background 206 and two foreground objects (i.e. the rectangle 202 and the triangle 204). As can be seen in Fig. 2 (b), the depth of the background 206 varies and the depth of the two objects 202 and 204 remains constant. The lower section of the rectangular object 202 has a similar depth to the closest section of the background 206.

Fig. 3 shows a patch 110 (as indicated in Fig. 1) corresponding to the triangular object 204 (as indicated in Fig. 1). For each patch 110, a patch texture image 302, a patch depth map 306 and a patch transparency map 304 (e.g. an alpha channel) is generated. Fig. 3 (a) shows the patch texture image 302 for the triangular object 204. Fig. 3 (b) shows the patch transparency map 304 corresponding to the triangular object 204 of Fig. 2, where the black sections of the patch transparency map 304 show fully transparent areas and the white sections show non-transparent areas. Fig. 3 (c) shows the patch depth map 306 of the triangular object 204.

The patch texture image 302 and the patch depth map 306 can be generated by directly copying the data from the source view image 102 (Fig. 2) and the source view depth map 104 (Fig. 2) at the section corresponding to the patch 110. The transparency values can be set to one where the triangular object 204 is present and zero where it is absent.

Alternatively, a more accurate algorithm can be used to estimate, for each patch 110, the foreground color and alpha (transparency) using so called alpha matting. In that case, the color of a pixel i may be a linear combination of the local foreground color F and the background color B based on the transparency value α of the pixel:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i$$

A trimap can be constructed based on the per pixel object component label map inside each patch 110. The trimap may consist of the classes ‘definitely foreground’ ($\alpha = 1$), ‘definitely background’ ($\alpha = 0$) and ‘uncertain’ (α needs to be estimated). The alpha matting algorithm then estimates both α_i and F_i for the pixels that are uncertain.

If the patch depth map 306 is used for depth image based rendering, the triangle object 204 could get covered behind the ‘left-over’ of the rectangle (e.g. the part of the rectangle at the bottom left of Fig. 3 (c)). This is because the pixels in this left-over region are closer to the camera than the pixels for the triangle object 204. This has to do with the view synthesis logic which renders the patch 110 using a triangular mesh where each vertex depends on the patch depth map 306.

To solve this problem, it may be advantageous to process the patch depth map 306 such that the 'left-over' regions of other patches 110 are changed to a depth value at least further than the local foreground of the corresponding foreground object of the patch 110. The amended patch depth map 308 in Fig. 3 (c) illustrates a simple method that just sets all pixels outside the triangle object 204 to a depth value that is equal to a minimum taken over all valid object pixels. In this example, the entire amended patch depth map 308 now receives the constant depth that the triangle object 204 has. Alternatively, all pixels not corresponding to the triangle object 204 may be set to a depth value lower than any depth value corresponding to the triangle object 204.

The approach of modifying the depth pixels inside the rectangle when outside the object region, in the present example, results from the design of the client rendering system. The client rendering application will typically warp each rectangle as a single regular mesh whole (ignoring which pixels correspond to the background). Pixels outside the object region that are closer to the camera would result in the mesh folding back over the object itself.

Alternatively, a so called geometry shader could be used to select/cut triangles during rendering such that only the pixels with the object label are warped. However, this would be more complex in terms of implementation of the real-time renderer.

In case the depth of the triangle object 204 does not vary or varies only a little, then the patch depth map 306 does not need to be stored as a map and a scalar depth value can be indicated for the entire patch 110.

Fig. 4 illustrates a first process for generating patches 110 (Fig. 1). Each source view is assumed to have an associated background depth map 106 (Fig. 1). For a sports stadium, this may be a depth map rendering of the ground surface model combined with a geometric model of the stadium. Fig. 4 (a) shows a source view depth map 104 with two foreground objects 202 and 204 (e.g. sports players). The foreground objects 202 and 204 can be detected by subtracting the depth of the background model 106 from the estimated source view depth map 104 followed by thresholding the result. Fig. 4 (b) shows the thresholded map 402. As can be seen from Fig. 4 (b), foreground objects 202 and 204 that occlude each other are still attached together in the (binary) thresholded map 402.

To separate the attached objects, depth step edge pixels are detected and set to, for example, zero. Fig. 4 (c) shows a thresholded map 404 with the depth step pixels set to zero. Note that a binary mask of the local background pixels is set bounding the depth step to zero and not the pixels corresponding to the local foreground. This is done to avoid losing foreground pixels. Consecutive connected component labelling then separates the foreground objects 202 and 204 from the background 206 and from each other into segments. Fig. 4 (d) shows an image 406 with the segments labelling each object (background 206 labelled as 0, triangular object 204 labelled as 1 and rectangular object 202 labelled as 2).

Bounding boxes 408 can then be detected for each segment. Fig. 4 (e) shows the source view depth map 104 with the bounding boxes 408 for the foreground objects 202 and 204. Each bounding box

408 can then further be extended vertically downwards to include the rectangular object 202 section where it touches (or is near) the depth of the ground surface of the background 206. Fig. 4 (f) shows the source view depth map 104 with the extended bounding boxes 410. The patches 110 can then be determined based on the bounding boxes 408 and/or the extended bounding boxes 410. Extending the bounding boxes 408 allows the bounding boxes 408 to further enclose any parts of the foreground objects 202 and 204 which were cut off at the thresholding step (e.g. due to parts of the foreground objects 202 and 204 being close to the background 206).

Fig. 5 shows a source view depth map 104 with three foreground objects 202, 204 and 502. A rectangular grid 504 is shown defining a sub-region 506 of the source view depth map 104 which contains all three foreground objects.

An alternative approach for the step of generating a patch 110 (Fig. 1) uses sub-regions 506 such as the sub-region 506 shown in Fig. 5. Depending on the number of depth surfaces present in the sub-region 506, one or multiple patches 110 are constructed corresponding to the sub-region 506. Fig. 5 shows an example sub-region 506 where four patches are needed to model all depth surfaces in the sub-region 506 (i.e. background 206 and three foreground objects 202, 204 and 502).

Having multiple patches per sub-region 506 allows the multiple patches to share the same patch texture image 302 and, potentially, the same patch depth map 304, thus reducing the overall amount of data that needs to be broadcast. Additionally, the spatial relationship between patches could also be defined by a grid (of sub-regions 506) instead of having to define the position of each patch.

The background depth map, as generated from the background model 106 (Fig. 1), can be used to segment the foreground objects 108 but is not essential. In the example of Fig. 5, instead of a background model, a classification, clustering and/or binning of pixels based on the depth values of the pixels in the source view depth maps 104 can be used to identify the different segments and hence produce the multiple patches per sub-region 506.

The number and position of the sub-regions 506 may depend on, for example, an object detection algorithm (or similar) detecting foreground objects in the source view images 102 (Fig. 1) or the source view depth maps 104. Alternatively, the number and position of sub-regions 506 may be fixed based on known positions of the foreground objects 202, 204 and 502.

Fig. 6 shows four patch transparency maps generated for a single sub-region 506 (Fig. 5). The four patch transparency maps 602, 604, 606 and 608 are all for the sub-region 506 shown in Fig. 5 and each correspond to a different patch 110 (Fig. 1). In this example, four patches 110 (and thus four patch transparency maps) are generated to a single sub-region 506 due to there being four depth surfaces (background and three objects) present in the sub-region 506.

Fig. 6 (a) shows a first patch transparency map 602 for the thin object 502 (Fig. 5). Fig. 6 (b) shows a second patch transparency map 604 for the rectangular object 202 (Fig. 5). Fig. 6 (c) shows a third patch transparency map 606 for the triangular object 204 (Fig. 5). Fig. 6 (d) shows a fourth patch transparency map 608 for the background 206 (Fig. 5).

Thus, multiple patches 110 can be generated when a region of a source view depth map 104 contains multiple depth surfaces. Each surface results in a separate patch 110. Only the transparency (alpha) maps are shown in Fig. 6 as a patch texture image 302 (Fig. 3) only needs to be stored once since it is the same for all four patches 110. Additionally, for some patches 110, a scalar depth map will suffice while for other a (differential) patch depth map may be needed.

A single background model 106 (video sprite texture and depth) can be constructed from the source views by accounting for the fact that it is known where the foreground patches 110 were removed and using this to fill in any gaps. For example, when multiple cameras image a hockey game, a single background sprite image can be generated that only contains the ground and the audience but not the players. This single background sprite can be modelled via a perspective projection with a wider field of view than the source views. The background sprite and depth can be packed together with the source view patches 110 into a single atlas 112 (Fig. 1).

View Synthesis

Patch 110 view synthesis starts after decoding the atlas 112 data that contains the patch texture images 302, the patch transparency map 304 (Fig. 3) and the patch depth maps 306 (Fig. 3) of the same foreground object 108 but for multiple source views. For example, a foreground object 108 may be visible in only five of in total eight source views. The background model 106 is rendered before all other patches 110 since it is known that it is always the furthest object in the 3D scene.

Given a target view matrix (defining the position and orientation from which the 3D scene is being viewed), the patches 110 are then first sorted in decreasing order based on the distance (z-axis) from the virtual viewpoint. The sorted patches then form patch groups where the z-coordinate variation within a group is typically smaller than the z-coordinate variation between patch groups. Note that patches 110 from multiple source views will end up in the same group depending on the virtual viewpoint.

View synthesis then alternates between warping and/or blending a patch group and compositing the blended result onto the previous compositing result.

Fig. 7 illustrates the step of warping and blending a patch group from the atlas 112. After fetching a patch 110 as shown in Fig. 1 (i.e. the patch texture image 302, the patch transparency map 304 (Fig. 3) and the patch depth map 306) in the patch group from the atlas 112, the patch 110 is warped to its associated source view buffer using the target view matrix. The result is hence directly represented into the target view image coordinates.

For illustrative purposes, only the patch texture images 302 and the patch depth maps 306 are shown in the atlas 112. The patch transparency maps 304 may also be included in the atlas 112 or may be embedded in, for example, the patch depth maps 306.

Each patch 110 is warped to its associated source view buffer 702 and all (or some) of the source view buffers 702 are used to composite the foreground object(s) 108 corresponding to the patch

group onto either the background model 106 (Fig. 1) if no other patch groups composited or onto the previously composited image 704 (if patch group already composited).

The number of source view buffers 702 used to composite the foreground object(s) 108 (Fig. 1) may be fixed in order to maintain a constant memory use during compositing. For example, only eight of the source view images (based on proximity of source view to target view) may be selected to perform compositing.

Fig. 8 shows a 3D scene with two source views 800 and two target views 808 and 810. Fig. 8 (a) shows a first target viewpoint 808 and Fig. 8 (b) shows a second viewpoint 810. Fig. 8 illustrates the need for sorting patches 110 (Fig. 1) at the client side based on the coordinates of the target views 808 or 810. Target views 808 and 810 may also be called virtual viewpoints or virtual views.

For the first target view 808 shown in Fig. 8 (a), objects 804 and 806 may end up in the same patch group which is warped and blended first. Object 802 is closest to the coordinates of the first target view 808 and is hence warped, blended and composited last.

However, this is different for the second target view 810 shown in Fig. 8 (b). For the second target view 810, objects 802 and 804 are warped, blended and composited first and object 806 is warped blended and composited last as object 806 is closer to the second target view 810 than objects 802 and 804.

Metadata

Metadata may also be stored for each patch 110. For example, for each patch 110, a source view identifier, a source view position and size $(u_{0,v}, v_{0,v}, w_v, h_v)$ and an atlas 112 (Fig. 1) position and size $(u_{0,a}, v_{0,a}, w_a, h_a)$ may be stored.

Let $(u_{0,a}, v_{0,a})$ represent the lower left corner of a rectangular patch 110 in the atlas coordinates. Thus, it is only necessary to sample atlas coordinates that lie inside the rectangle that is being warped for a given patch size. Given that the normalized (u, v) coordinates lie in the domain $[0,1]$, the normalized atlas coordinates (u_a, v_a) of a point (u, v) of the rectangle can be calculated as:

$$u_a = u_{0,a} + uw_a$$

$$v_a = v_{0,a} + vh_a$$

The atlas coordinates (u_a, v_a) are used to access the depth value inside the patch depth maps 306 (Fig 3) during a vertex shader stage and interpolate the color and transparency by passing the atlas coordinates to a fragment shader.

However, to warp the patch 110 to an output view it may be necessary to know the rectangle source view image 102 (Fig. 1) coordinates. Given that the normalized (u, v) coordinates lie in the domain $[0,1]$, the normalized atlas coordinates (u_v, v_v) of a point (u, v) of the rectangle can be calculated as:

$$u_v = u_{0,v} + uw_v$$

$$v_v = v_{0,v} + vh_v$$

5 Note that because normalized coordinates are used, the source view images 102 and/or depth maps can have less or more pixels than stored in the atlas 112. With a fixed pixel budget in the atlas 112, patches 110 can be scaled to always fit.

 The skilled person would be readily capable of developing a processor for carrying out any herein described method. Thus, each step of a flow chart may represent a different action performed
10 by a processor, and may be performed by a respective module of the processing processor.

 As discussed above, the system makes use of a processor to perform the data processing. The processor can be implemented in numerous ways, with software and/or hardware, to perform the various functions required. The processor typically employs one or more microprocessors that may be programmed using software (e.g., microcode) to perform the required functions. The processor may be
15 implemented as a combination of dedicated hardware to perform some functions and one or more programmed microprocessors and associated circuitry to perform other functions.

 Examples of circuitry that may be employed in various embodiments of the present disclosure include, but are not limited to, conventional microprocessors, application specific integrated circuits (ASICs), and field-programmable gate arrays (FPGAs).

20 In various implementations, the processor may be associated with one or more storage media such as volatile and non-volatile computer memory such as RAM, PROM, EPROM, and EEPROM. The storage media may be encoded with one or more programs that, when executed on one or more processors and/or controllers, perform the required functions. Various storage media may be fixed within a processor or controller or may be transportable, such that the one or more programs stored
25 thereon can be loaded into a processor.

 Variations to the disclosed embodiments can be understood and effected by those skilled in the art in practicing the claimed invention, from a study of the drawings, the disclosure and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality.

30 A single processor or other unit may fulfill the functions of several items recited in the claims. A computer program may be stored/distributed on a suitable medium, such as an optical storage medium or a solid-state medium supplied together with or as part of other hardware, but may also be distributed in other forms, such as via the Internet or other wired or wireless telecommunication systems.

35 If the term "adapted to" is used in the claims or description, it is noted the term "adapted to" is intended to be equivalent to the term "configured to".

 Any reference signs in the claims should not be construed as limiting the scope.

CLAIMS:

Claim 1. A method of depth segmentation for the generation of a multi-view video data, the method comprising:

obtaining a plurality of source view images (102) and source view depth maps (104) representative of a 3D scene from a plurality of sensors;

5 segmenting foreground objects (108) in the 3D scene from the source view images (102) and/or the source view depth maps (104); and

generating one or more patches (110) for each source view image (102) and source view depth map (104) containing at least one foreground object (108), wherein:

each patch (110) corresponds to a foreground object (108),

10 generating a patch (110) comprises generating a patch texture image (302), a patch depth map (306) and a patch transparency map (304) based on the source view images (102) and the source view depth maps (104), and

each patch (110) is based on a section of the corresponding source view image which is smaller than the source view image;

15 obtaining a background model (106) comprising at least one background depth map and background texture data; and

generating an atlas (112) based on the patch texture images (302), the patch depth maps (306), the patch transparency maps (304) and the background model (106).

20 Claim 2. The method of claim 1, further comprising obtaining a plurality of background depth maps of the 3D scene representative of the background (206) of the 3D scene, each background depth map containing depth data of the background (206) from a particular orientation, wherein segmenting foreground objects (108) is based on a difference between a background depth map and the corresponding source view depth map (104).

25

Claim 3. The method of claims 1 or 2, wherein detecting foreground objects (108) comprises:

subtracting a background depth map from each corresponding source view depth map (104), to produce a respective difference image;

30 thresholding the difference images, wherein thresholding comprises comparing the pixel values of the difference images with a threshold value to produce threshold maps which differentiate between the background and the foreground objects (108);

identifying pixels in the source view depth maps (104) corresponding to depth steps, wherein depth steps are defined by the differences between neighboring depth values in a source view

depth map (104) being larger than a depth threshold value;

marking the depth steps in the threshold maps (402) as background, thereby to differentiate the foreground objects (108) from each other; and

generating bounding boxes (408) around the foreground objects (108) based on the adjusted threshold maps (404).

Claim 4. The method of claim 3, wherein detecting foreground objects (108) further comprises enlarging the bounding boxes (408) thereby to include regions of the foreground objects (108) in the subtracted maps (402) below the threshold value.

Claim 5. The method of any one of claims 1 to 4, further comprising adapting pixel depth values of a patch depth map (306) such that all of the pixel depth values of the patch depth map (306) consist of values equal to or lower than the depth values of the corresponding foreground object (108).

Claim 6. The method of any one of claims 1 to 5, further comprising:
identifying a plurality of patches from the patches (110) corresponding to a first foreground object based on identifying the patch depth maps (306) of the patches (110) within an object depth range; and
correcting the identified patch depth maps such that they correspond to an object location in the 3D scene.

Claim 7. The method of any one of claims 1 to 6, further comprising pruning the patches based on a measure of consistency between patches in multiple source views.

Claim 8. The method of any one of claims 1 to 7, wherein generating one or more patches (110) comprises:

identifying a sub-region (506) in a source view depth map (104);
determining a number of depth surfaces of different depths present in the sub-region (506); and

generating a patch (110) for each depth surface in the sub-region (506), wherein each patch (110) comprises a different patch transparency map (304).

Claim 9. A system comprising:

one or more processors comprising computer program code which, when executed on a computing device, cause the computing device to perform all of the steps of the method according to any of claims 1 to 8; and

a plurality of sensors configured to obtain the source view images (102) and source view depth maps (104).

Claim 10. A method for rendering multi-view videos, the method comprising:

5 receiving an atlas (112) with a plurality of patches (110) and a background model (106) of a 3D scene, wherein:

each patch (110) corresponds to a foreground object (108),

each patch (110) comprises a patch texture image (302), a patch depth map (306) and a patch transparency map (304) derived from source view images (102) and source view depth maps (104), and

10 each patch (110) is based on a section of the corresponding source view image which is smaller than the source view image;

receiving a virtual viewpoint (808, 810) within the 3D scene;

15 sorting the patches (110) based on the difference between the position of the virtual viewpoint (808, 810) and the position of the foreground objects (108) corresponding to each patch (110); and

rendering the background model (106) and the sorted patches.

Claim 11. The method of claim 10, further comprising grouping the patches (110) based on the position of the corresponding foreground objects (108) relative to the virtual viewpoint (808, 810).

Claim 12. The method of claim 11, wherein rendering the background model (106) and the sorted patches comprises:

rendering the background model (106);

25 warping and/or blending a first patch group;

compositing the warped and/or blended first patch group onto the rendered background model (106);

30 warping and/or blending a second patch group, wherein the position of the foreground objects (108) corresponding to the second patch group relative to the virtual viewpoint (808, 810) is closer to the virtual viewpoint than the position of the foreground objects (108) corresponding to the first patch group; and

compositing the warped and/or blended second patch group onto the warped and/or blended first patch group.

35 Claim 13. A computer program product comprising computer program code which, when executed on a computing device having a processing system, cause the processing system to perform all of the

steps of the method according to any of claims 1 to 8 and/or the method according to any one of claims 10 to 12.

Claim 14. A processor configured to execute the computer program code of claim 13.

1/6

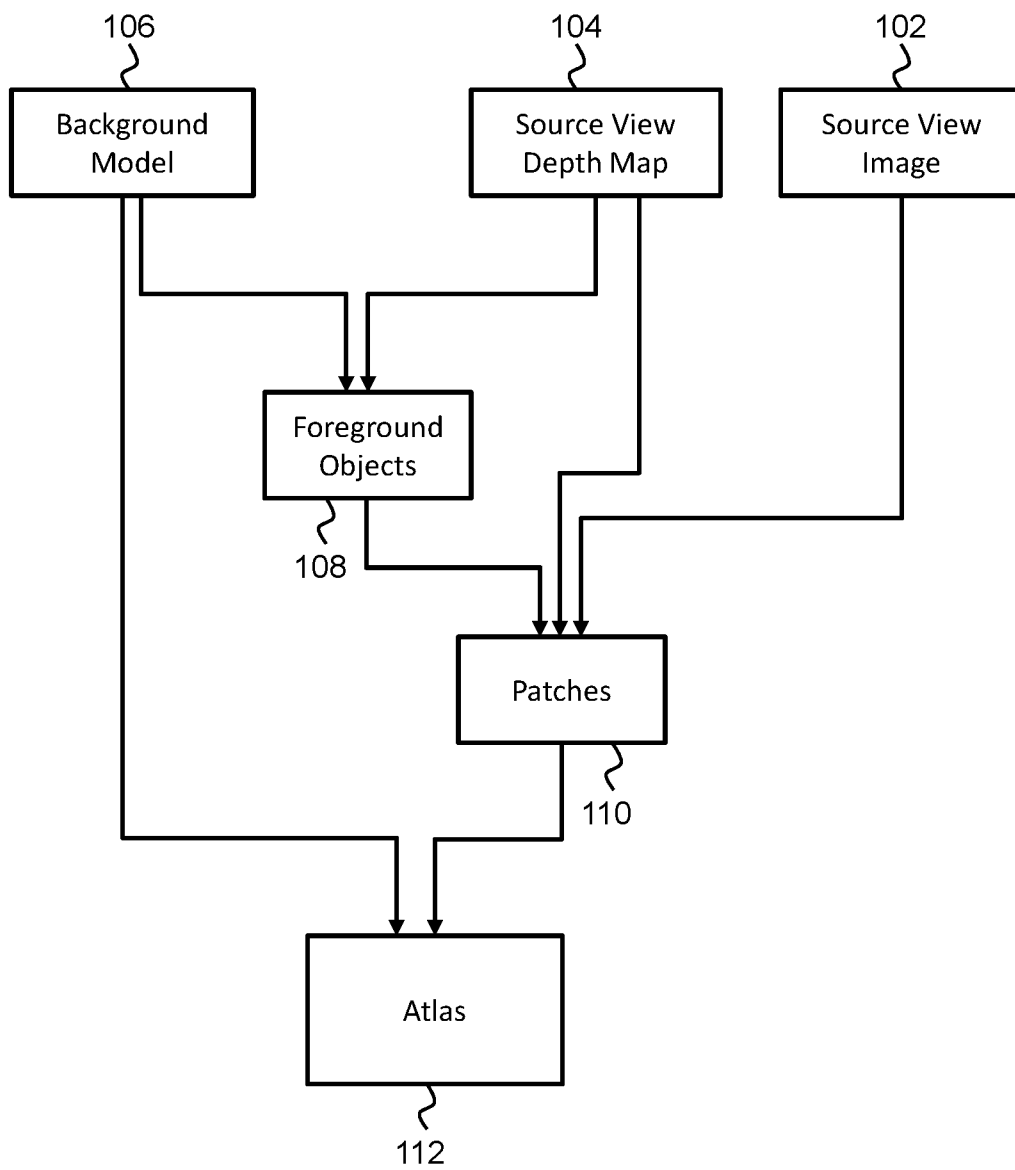


FIG. 1

2/6

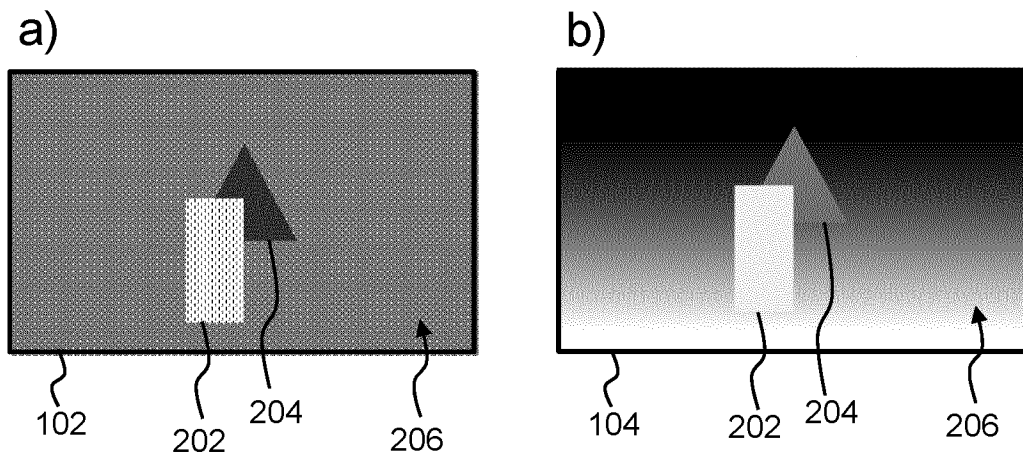


FIG. 2

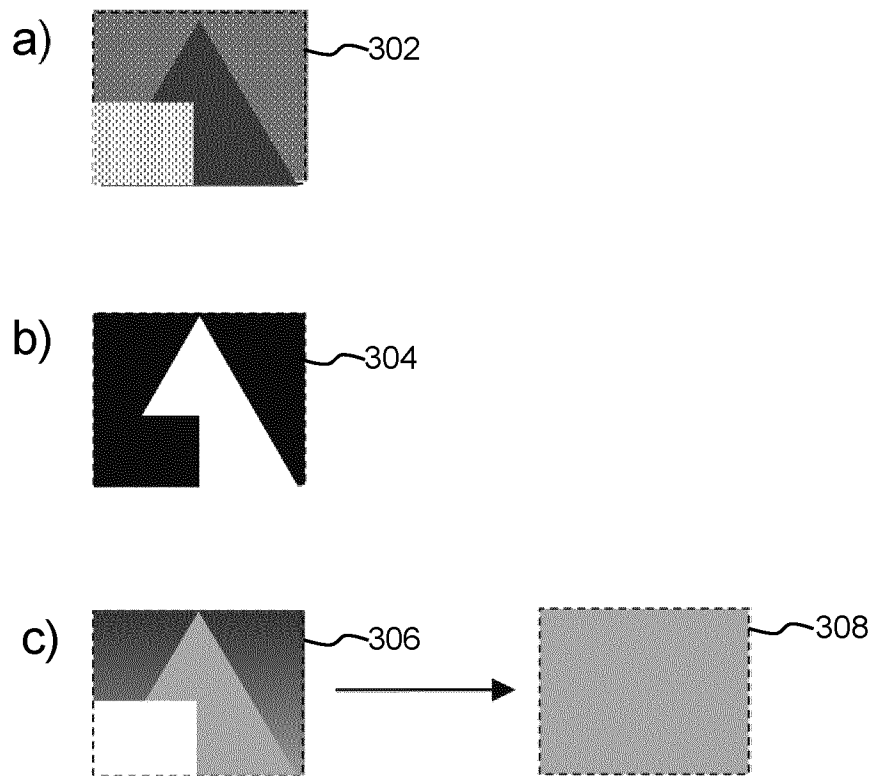


FIG. 3

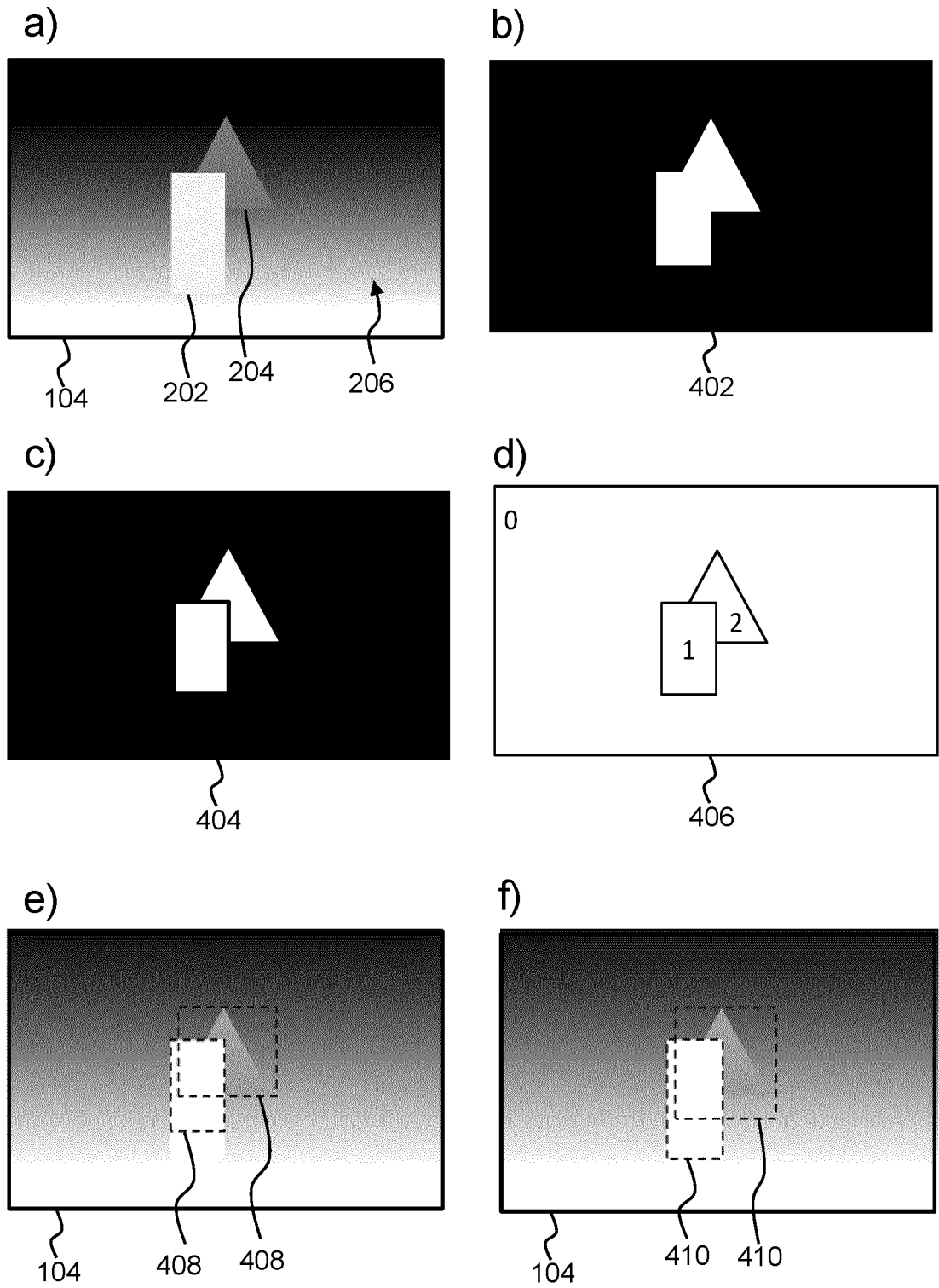


FIG. 4

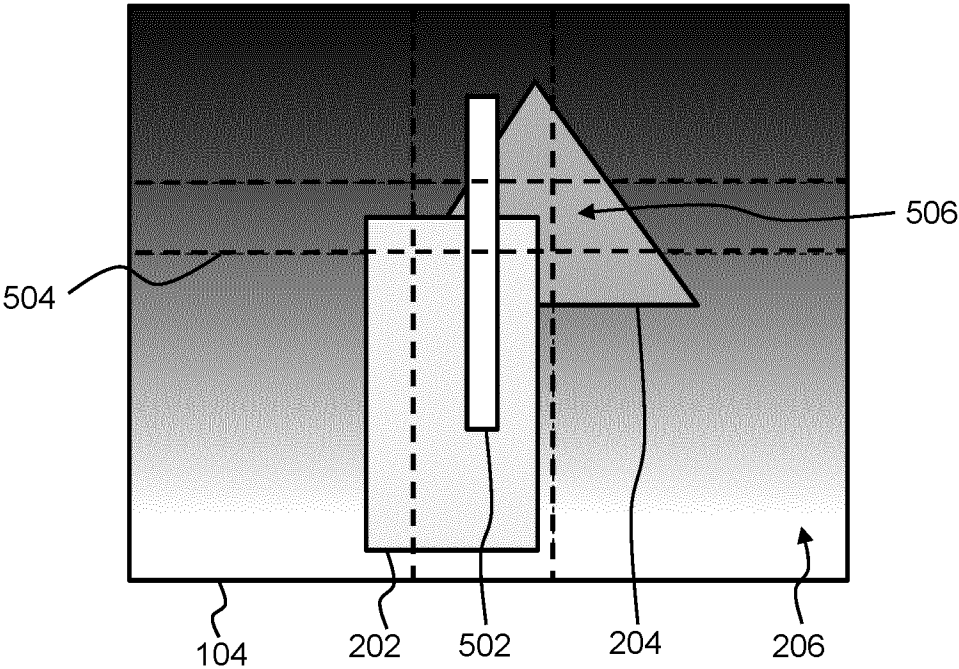


FIG. 5

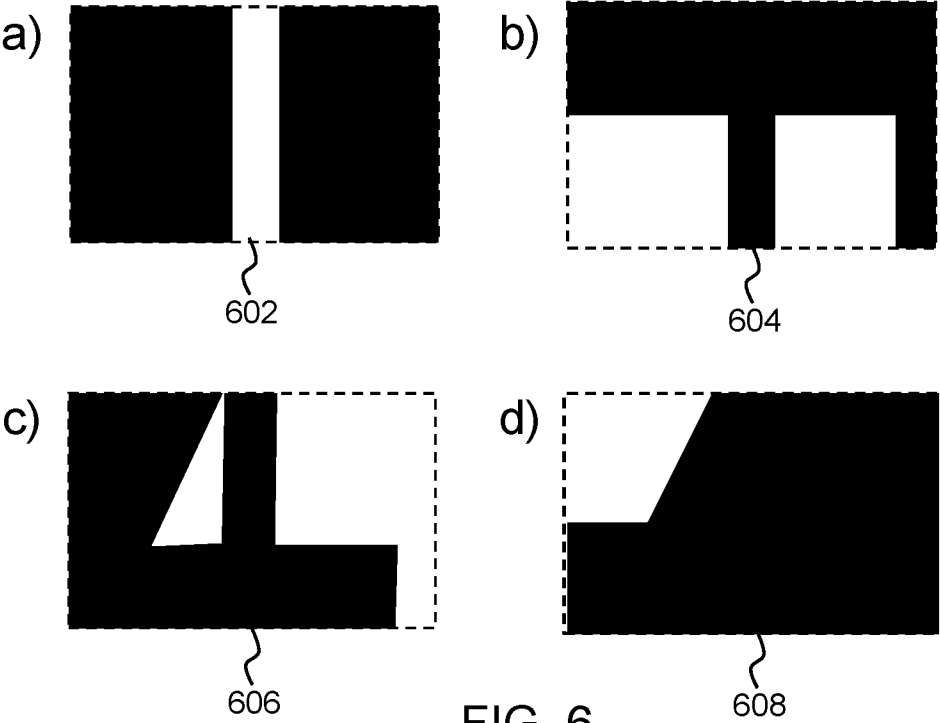


FIG. 6

5/6

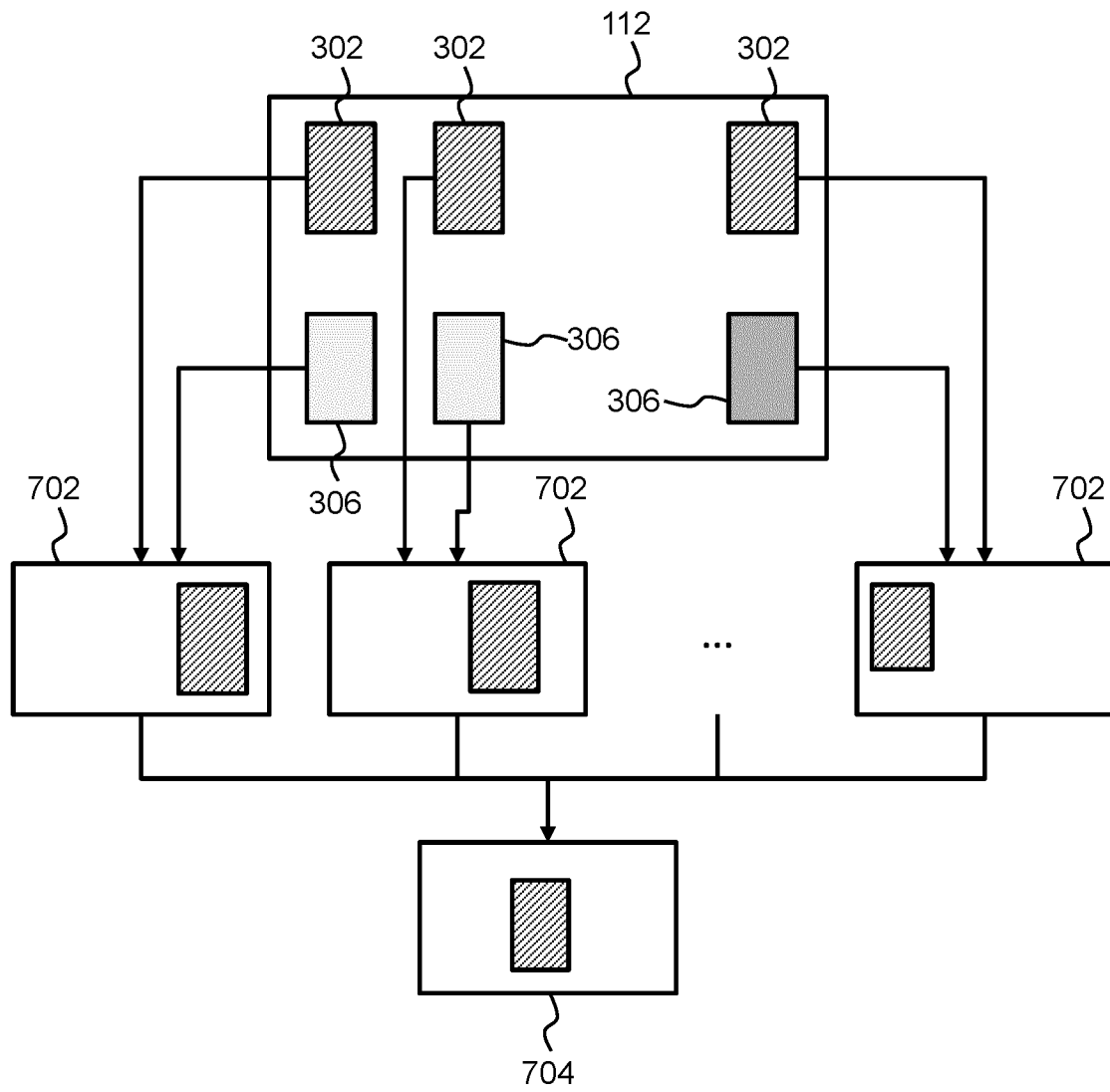
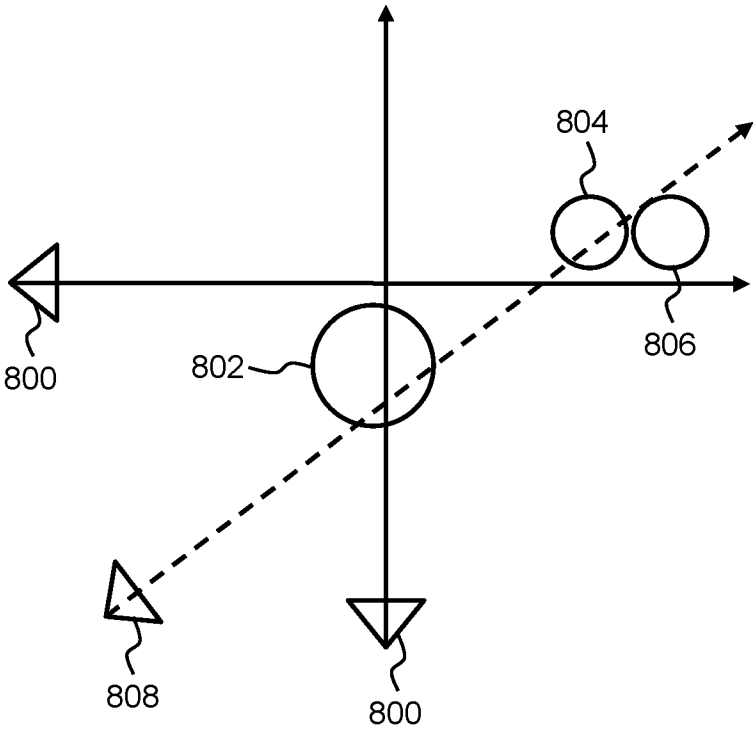


FIG. 7

6/6

a)



b)

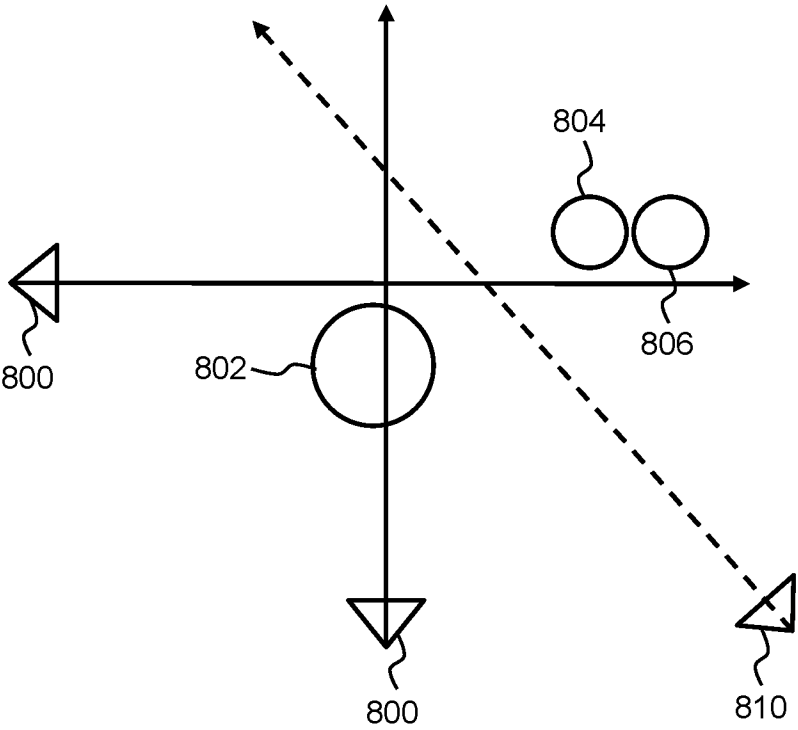


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/064243

A. CLASSIFICATION OF SUBJECT MATTER		
INV.	G06T7/13	G06T7/136
	G06T7/55	H04N13/10
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 H04N		
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	WO 2021/067501 A1 (INTEL CORP [US]) 8 April 2021 (2021-04-08)	1, 2, 5-14
A	[0029-30, 42, 46, 66, 71-73, 79-80] -----	3, 4
A	CN 109 147 025 A (UNIV BEIHANG) 4 January 2019 (2019-01-04) the whole document -----	1-14
A	DIAZ RODRIGO ET AL: "Region Dependent Mesh Refinement for Volumetric Video Workflows", 2019 INTERNATIONAL CONFERENCE ON 3D IMMERSION (IC3D), IEEE, 11 December 2019 (2019-12-11), pages 1-8, XP033698831, DOI: 10.1109/IC3D48390.2019.8975991 [retrieved on 2020-01-29] the whole document ----- -/-	1-14
☒ 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 15 September 2022		Date of mailing of the international search report 26/09/2022
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 Nicolau, Stephane

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2022/064243

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2019/037150 A1 (SRIKANTH MANOHAR [US] ET AL) 31 January 2019 (2019-01-31) the whole document -----	1-14
A	CN 110 148 217 A (BEIJING HUAJIE IMI TECH CO LTD) 20 August 2019 (2019-08-20) the whole document -----	1-14

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2022/064243

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2021067501 A1	08-04-2021	CN 114342399 A DE 112020004716 T5 US 2022262041 A1 WO 2021067501 A1	12-04-2022 04-08-2022 18-08-2022 08-04-2021
CN 109147025 A	04-01-2019	NONE	
US 2019037150 A1	31-01-2019	US 2015254868 A1 US 2019037150 A1 WO 2015134996 A1	10-09-2015 31-01-2019 11-09-2015
CN 110148217 A	20-08-2019	NONE	