



(51) International Patent Classification:

H04N 19/503 (2014.01) *G06T 3/60* (2006.01)*H04N 19/597* (2014.01) *G06T 3/00* (2006.01)*H04N 19/105* (2014.01)

(21) International Application Number:

PCT/FI2017/050951

(22) International Filing Date:

29 December 2017 (29.12.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

20175007 03 January 2017 (03.01.2017) FI

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI];
Karaportti 3, 02610 Espoo (FI).(72) Inventors: **AMINLOU, Alireza**; Vaajakatu 5 G
152, 33720 Tampere (FI). **HANNUKSELA, Miska**;
Rusthollinrinne 2, 33610 Tampere (FI). **GHAZNAVI
YOUVALARI, Ramin**; Orivedenkatu 8 G 155, 33720
Tampere (FI).(74) Agent: **NOKIA TECHNOLOGIES OY** et al.; Ari Aarnio,
IPR Department, Karakaari 7, 02610 Espoo (FI).(81) Designated States (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, 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))
- 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))

(54) Title: AN APPARATUS, A METHOD AND A COMPUTER PROGRAM FOR VIDEO CODING AND DECODING

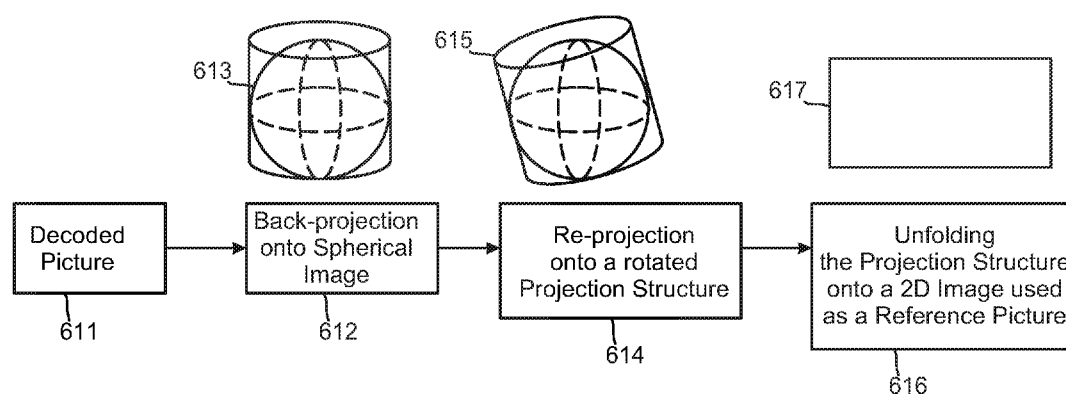


Fig. 6

(57) **Abstract:** There are disclosed various methods, apparatuses and computer program products for video encoding and decoding. In some embodiments a first reconstructed picture is interpreted as a first three-dimensional picture in a coordinate system. A rotation is obtained and the first three-dimensional picture is projected (612, 614) onto a first geometrical projection structure (613, 615), the geometrical projection structure having an orientation according to the rotation within the coordinate system. A first reference picture is formed (616) by unfolding the first geometrical projection structure into a second geometrical projection structure, and at least a block of a second reconstructed picture is predicted from the first reference picture.



AN APPARATUS, A METHOD AND A COMPUTER PROGRAM FOR VIDEO CODING AND DECODING

TECHNICAL FIELD

5 [0001] The present invention relates to an apparatus, a method and a computer program for video coding and decoding.

BACKGROUND

10 [0002] This section is intended to provide a background or context to the invention that is recited in the claims. The description herein may include concepts that could be pursued, but are not necessarily ones that have been previously conceived or pursued. Therefore, unless otherwise indicated herein, what is described in this section is not prior art to the description and claims in this application and is not admitted to be prior art by inclusion in this section.

15 [0003] A video coding system may comprise an encoder that transforms an input video into a compressed representation suited for storage/transmission and a decoder that can uncompress the compressed video representation back into a viewable form. The encoder may discard some information in the original video sequence in order to represent the video in a more compact form, for example, to enable the storage/transmission of the
20 video information at a lower bitrate than otherwise might be needed.

SUMMARY

[0004] Some embodiments provide a method for encoding and decoding video information. In some embodiments of the present invention there is provided a method,
25 apparatus and computer program product for video coding as well as decoding.

[0005] Various aspects of examples of the invention are provided in the detailed description.

[0006] According to a first aspect, there is provided a method comprising:
interpreting a first reconstructed picture as a first three-dimensional picture in a
30 coordinate system;
obtaining a rotation;
projecting the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

forming a first reference picture, said forming comprising unfolding the first geometrical projection structure into a second geometrical projection structure;
predicting at least a block of a second reconstructed picture from the first reference picture.

5 [0007] An apparatus according to a second aspect comprises at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:

interpret a first reconstructed picture as a first three-dimensional picture in a coordinate system;

10 obtain a rotation;

project the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

form a first reference picture, said forming comprising unfolding the first
15 geometrical projection structure into a second geometrical projection structure;
predict at least a block of a second reconstructed picture from the first reference picture.

[0008] A computer readable storage medium according to a third aspect comprises code
for use by an apparatus, which when executed by a processor, causes the apparatus to
20 perform:

interpret a first reconstructed picture as a first three-dimensional picture in a coordinate system;

obtain a rotation;

project the first three-dimensional picture onto a first geometrical projection
25 structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

form a first reference picture, said forming comprising unfolding the first
geometrical projection structure into a second geometrical projection structure;

predict at least a block of a second reconstructed picture from the first reference
30 picture.

[0009] An apparatus according to a fourth aspect comprises:

means for interpreting a first reconstructed picture as a first three-dimensional
picture in a coordinate system;

means for obtaining a rotation;

means for projecting the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

5 means for forming a first reference picture, said forming comprising unfolding the first geometrical projection structure into a second geometrical projection structure;

means for predicting at least a block of a second reconstructed picture from the first reference picture.

[0010] Further aspects include at least apparatuses and computer program products/code stored on a non-transitory memory medium arranged to carry out the above methods.

10

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] For a more complete understanding of example embodiments of the present invention, reference is now made to the following descriptions taken in connection with the accompanying drawings in which:

15 [0012] Figure 1a shows an example of a multi-camera system as a simplified block diagram, in accordance with an embodiment;

[0013] Figure 1b shows a perspective view of a multi-camera system, in accordance with an embodiment;

20 [0014] Figure 2a illustrates image stitching, projection, and mapping processes, in accordance with an embodiment;

[0015] Figure 2b illustrates a process of forming a monoscopic equirectangular panorama picture, in accordance with an embodiment;

[0016] Figure 3a shows an unprocessed reference frame having a regular grids, in accordance with an embodiment;

25 [0017] Figure 3b shows an unprocessed reference frame having a rotation angle of 1° , in accordance with an embodiment;

[0018] Figure 3c shows an unprocessed reference frame having a having a rotation angle of 5° , in accordance with an embodiment;

30 [0019] Figure 3d illustrates an example of indicating a displacement for each corner of a reference picture for temporal reference picture resampling;

[0020] Figure 4a shows a schematic diagram of an encoder suitable for implementing embodiments of the invention;

[0021] Figure 4b shows a schematic diagram of a decoder suitable for implementing embodiments of the invention;

- [0022] Figure 5a shows a video encoding method, in accordance with an embodiment;
- [0023] Figure 5b shows a video decoding method, in accordance with an embodiment;
- [0024] Figure 6 illustrates an example of manipulating/resampling reference frames based on camera orientation of a frame to be encoded for 360-degree video encoding, in accordance with an embodiment;
- [0025] Figure 7a shows an example of a three-dimensional coordinate system;
- [0026] Figure 7b shows another example of a three-dimensional coordinate system;
- [0027] Figure 8a shows an example of an out-of-the-loop approach, in accordance with an embodiment;
- [0028] Figure 8b shows another example of an out-of-the-loop approach, in accordance with an embodiment;
- [0029] Figure 9 shows an example of decoding images/frames of a video, in accordance with an embodiment;
- [0030] Figure 10a shows a flow chart of an encoding method, in accordance with an embodiment;
- [0031] Figure 10b shows a flow chart of a decoding method, in accordance with an embodiment;
- [0032] Figure 11a shows spatial candidate sources of the candidate motion vector predictor, in accordance with an embodiment;
- [0033] Figure 11b shows temporal candidate sources of the candidate motion vector predictor, in accordance with an embodiment;
- [0034] Figure 12 shows a schematic diagram of an example multimedia communication system within which various embodiments may be implemented;
- [0035] Figure 13 shows schematically an electronic device employing embodiments of the invention;
- [0036] Figure 14 shows schematically a user equipment suitable for employing embodiments of the invention;
- [0037] Figure 15 further shows schematically electronic devices employing embodiments of the invention connected using wireless and wired network connections.

DETAILED DESCRIPTION OF SOME EXAMPLE EMBODIMENTS

- [0038] Figures 1a and 1b illustrate an example of a camera having multiple lenses and imaging sensors but also other types of cameras may be used to capture wide view images and/or wide view video.

[0039] In the following, the terms wide view image and wide view video mean an image and a video, respectively, which comprise visual information having a relatively large viewing angle, larger than 100 degrees. Hence, a so called 360 panorama image/video as well as images/videos captured by using a fish eye lens may also be called as a wide view image/video in this specification. More generally, the wide view image/video may mean an image/video in which some kind of projection distortion may occur when a direction of view changes between successive images or frames of the video so that a transform may be needed to find out co-located pixels from a reference image or a reference frame. This will be described in more detail later in this specification.

10 [0040] The camera 100 of Figure 1a comprises two or more camera units 102 and is capable of capturing wide view images and/or wide view video. In this example the number of camera units 102 is eight, but may also be less than eight or more than eight. Each camera unit 102 is located at a different location in the multi-camera system and may have a different orientation with respect to other camera units 102. As an example, 15 the camera units 102 may have an omnidirectional constellation so that it has a 360 viewing angle in a 3D-space. In other words, such camera 100 may be able to see each direction of a scene so that each spot of the scene around the camera 100 can be viewed by at least one camera unit 102.

[0041] The camera 100 of Figure 1a may also comprise a processor 104 for controlling the 20 operations of the camera 100. There may also be a memory 106 for storing data and computer code to be executed by the processor 104, and a transceiver 108 for communicating with, for example, a communication network and/or other devices in a wireless and/or wired manner. The camera 100 may further comprise a user interface (UI) 110 for displaying information to the user, for generating audible signals and/or for 25 receiving user input. However, the camera 100 need not comprise each feature mentioned above, or may comprise other features as well. For example, there may be electric and/or mechanical elements for adjusting and/or controlling optics of the camera units 102 (not shown).

[0042] Figure 1a also illustrates some operational elements which may be implemented, 30 for example, as a computer code in the software of the processor, in a hardware, or both. A focus control element 114 may perform operations related to adjustment of the optical system of camera unit or units to obtain focus meeting target specifications or some other predetermined criteria. An optics adjustment element 116 may perform movements of the optical system or one or more parts of it according to instructions provided by the focus

control element 114. It should be noted here that the actual adjustment of the optical system need not be performed by the apparatus but it may be performed manually, wherein the focus control element 114 may provide information for the user interface 110 to indicate a user of the device how to adjust the optical system.

5 [0043] Figure 1b shows as a perspective view the camera 100 of Figure 1a. In Figure 1b seven camera units 102a-102g can be seen, but the camera 100 may comprise even more camera units which are not visible from this perspective. Figure 1b also shows two microphones 112a, 112b, but the apparatus may also comprise one or more than two microphones.

10 [0044] It should be noted here that embodiments disclosed in this specification may also be implemented with apparatuses having only one camera unit 102 or less or more than eight camera units 102a-102g.

[0045] In accordance with an embodiment, the camera 100 may be controlled by another device (not shown), wherein the camera 100 and the other device may communicate with
15 each other and a user may use a user interface of the other device for entering commands, parameters, etc. and the user may be provided information from the camera 100 via the user interface of the other device.

[0046] Terms 360-degree video or virtual reality (VR) video may be used interchangeably. They may generally refer to video content that provides such a large field of view that
20 only a part of the video is displayed at a single point of time in typical displaying arrangements. For example, a virtual reality video may be viewed on a head-mounted display (HMD) that may be capable of displaying e.g. about 100-degree field of view (FOV). The spatial subset of the virtual reality video content to be displayed may be selected based on the orientation of the head-mounted display. In another example, a flat-panel viewing environment is assumed, wherein e.g. up to 40-degree field-of-view may be
25 displayed. When displaying wide field of view content (e.g. fisheye) on such a display, it may be preferred to display a spatial subset rather than the entire picture.

[0047] 360-degree image or video content may be acquired and prepared for example as follows. Images or video can be captured by a set of cameras or a camera device with
30 multiple lenses and imaging sensors. The acquisition results in a set of digital image/video signals. The cameras/lenses may cover all directions around the center point of the camera set or camera device. The images of the same time instance are stitched, projected, and mapped onto a packed virtual reality frame. The breakdown of image stitching, projection, and mapping processes are illustrated with Figure 2a and described as follows. Input

images 201 are stitched and projected 202 onto a three-dimensional projection structure, such as a sphere or a cube. The projection structure may be considered to comprise one or more surfaces, such as plane(s) or part(s) thereof. A projection structure may be defined as a three-dimensional structure consisting of one or more surface(s) on which the captured
5 virtual reality image/video content may be projected, and from which a respective projected frame can be formed. The image data on the projection structure is further arranged onto a two-dimensional projected frame 203. The term projection may be defined as a process by which a set of input images are projected onto a projected frame. There may be a pre-defined set of representation formats of the projected frame, including
10 for example an equirectangular panorama and a cube map representation format.

[0048] Region-wise mapping 204 may be applied to map projected frames 203 onto one or more packed virtual reality frames 205. In some cases, the region-wise mapping may be understood to be equivalent to extracting two or more regions from the projected frame, optionally applying a geometric transformation (such as rotating, mirroring, and/or
15 resampling) to the regions, and placing the transformed regions in spatially non-overlapping areas, a.k.a. constituent frame partitions, within the packed virtual reality frame. If the region-wise mapping is not applied, the packed virtual reality frame 205 may be identical to the projected frame 203. Otherwise, regions of the projected frame are mapped onto a packed virtual reality frame by indicating the location, shape, and size of
20 each region in the packed virtual reality frame. The term mapping may be defined as a process by which a projected frame is mapped to a packed virtual reality frame. The term packed virtual reality frame may be defined as a frame that results from a mapping of a projected frame. In practice, the input images 201 may be converted to packed virtual reality frames 205 in one process without intermediate steps.

25 [0049] 360-degree panoramic content (i.e., images and video) cover horizontally the full 360-degree field-of-view around the capturing position of an imaging device. The vertical field-of-view may vary and can be e.g. 180 degrees. Panoramic image covering 360-degree field-of-view horizontally and 180-degree field-of-view vertically can be represented by a sphere that has been mapped to a two-dimensional image plane using
30 equirectangular projection. In this case, the horizontal coordinate may be considered equivalent to a longitude, and the vertical coordinate may be considered equivalent to a latitude, with no transformation or scaling applied. In some cases panoramic content with 360-degree horizontal field-of-view but with less than 180-degree vertical field-of-view may be considered special cases of equirectangular projection, where the polar areas of

the sphere have not been mapped onto the two-dimensional image plane. In some cases panoramic content may have less than 360-degree horizontal field-of-view and up to 180-degree vertical field-of-view, while otherwise have the characteristics of equirectangular projection format.

5 [0050] In cube map projection format, spherical video is projected onto the six faces (a.k.a. sides) of a cube. The cube map may be generated e.g. by first rendering the spherical scene six times from a viewpoint, with the views defined by an 90 degree view frustum representing each cube face. The cube sides may be frame-packed into the same frame or each cube side may be treated individually (e.g. in encoding). There are many possible
10 orders of locating cube sides onto a frame and/or cube sides may be rotated or mirrored. The frame width and height for frame-packing may be selected to fit the cube sides "tightly" e.g. at 3x2 cube side grid, or may include unused constituent frames e.g. at 4x3 cube side grid.

[0051] The process of forming a monoscopic equirectangular panorama picture is
15 illustrated in Figure 2b, in accordance with an embodiment. A set of input images 211, such as fisheye images of a camera array or a camera device 100 with multiple lenses and sensors 102, is stitched 212 onto a spherical image 213. The spherical image 213 is further projected 214 onto a cylinder 215 (without the top and bottom faces). The cylinder 215 is unfolded 216 to form a two-dimensional projected frame 217. In practice one or more of
20 the presented steps may be merged; for example, the input images 213 may be directly projected onto a cylinder 217 without an intermediate projection onto the sphere 213 and/or to the cylinder 215. The projection structure for equirectangular panorama may be considered to be a cylinder that comprises a single surface.

[0052] In general, 360-degree content can be mapped onto different types of solid
25 geometrical structures, such as polyhedron (i.e. a three-dimensional solid object containing flat polygonal faces, straight edges and sharp corners or vertices, e.g., a cube or a pyramid), cylinder (by projecting a spherical image onto the cylinder, as described above with the equirectangular projection), cylinder (directly without projecting onto a sphere first), cone, etc. and then unwrapped to a two-dimensional image plane. The two-
30 dimensional image plane can also be regarded as a geometrical structure. In other words, 360-degree content can be mapped onto a first geometrical structure and further unfolded to a second geometrical structure. However, it may be possible to directly obtain the transformation to the second geometrical structure from the original 360-degree content or from other wide view visual content.

[0053] In some cases panoramic content with 360-degree horizontal field-of-view but with less than 180-degree vertical field-of-view may be considered special cases of equirectangular projection, where the polar areas of the sphere have not been mapped onto the two-dimensional image plane. In some cases a panoramic image may have less than 360-degree horizontal field-of-view and up to 180-degree vertical field-of-view, while otherwise has the characteristics of equirectangular projection format.

[0054] Real-time Transport Protocol (RTP) is widely used for real-time transport of timed media such as audio and video. RTP may operate on top of the User Datagram Protocol (UDP), which in turn may operate on top of the Internet Protocol (IP). RTP is specified in Internet Engineering Task Force (IETF) Request for Comments (RFC) 3550, available from www.ietf.org/rfc/rfc3550.txt. In RTP transport, media data is encapsulated into RTP packets. Typically, each media type or media coding format has a dedicated RTP payload format.

[0055] An RTP session is an association among a group of participants communicating with RTP. It is a group communications channel which can potentially carry a number of RTP streams. An RTP stream is a stream of RTP packets comprising media data. An RTP stream is identified by an SSRC belonging to a particular RTP session. SSRC refers to either a synchronization source or a synchronization source identifier that is the 32-bit SSRC field in the RTP packet header. A synchronization source is characterized in that all packets from the synchronization source form part of the same timing and sequence number space, so a receiver may group packets by synchronization source for playback. Examples of synchronization sources include the sender of a stream of packets derived from a signal source such as a microphone or a camera, or an RTP mixer. Each RTP stream is identified by a SSRC that is unique within the RTP session.

[0056] Video codec may comprise an encoder that transforms the input video into a compressed representation suited for storage/transmission and a decoder that can uncompress the compressed video representation back into a viewable form. A video encoder and/or a video decoder may also be separate from each other, i.e. need not form a codec. Typically encoder discards some information in the original video sequence in order to represent the video in a more compact form (that is, at lower bitrate). A video encoder may be used to encode an image sequence, as defined subsequently, and a video decoder may be used to decode a coded image sequence. A video encoder or an intra coding part of a video encoder or an image encoder may be used to encode an image, and

a video decoder or an inter decoding part of a video decoder or an image decoder may be used to decode a coded image.

[0057] Some hybrid video encoders, for example many encoder implementations of ITU-T H.263 and H.264, encode the video information in two phases. Firstly pixel values in a certain picture area (or "block") are predicted for example by motion compensation means (finding and indicating an area in one of the previously coded video frames that corresponds closely to the block being coded) or by spatial means (using the pixel values around the block to be coded in a specified manner). Secondly the prediction error, i.e. the difference between the predicted block of pixels and the original block of pixels, is coded. This is typically done by transforming the difference in pixel values using a specified transform (e.g. Discrete Cosine Transform (DCT) or a variant of it), quantizing the coefficients and entropy coding the quantized coefficients. By varying the fidelity of the quantization process, encoder can control the balance between the accuracy of the pixel representation (picture quality) and size of the resulting coded video representation (file size or transmission bitrate).

[0058] In temporal prediction, the sources of prediction are previously decoded pictures (a.k.a. reference pictures). In intra block copy (a.k.a. intra-block-copy prediction), prediction is applied similarly to temporal prediction but the reference picture is the current picture and only previously decoded samples can be referred in the prediction process. Inter-layer or inter-view prediction may be applied similarly to temporal prediction, but the reference picture is a decoded picture from another scalable layer or from another view, respectively. In some cases, inter prediction may refer to temporal prediction only, while in other cases inter prediction may refer collectively to temporal prediction and any of intra block copy, inter-layer prediction, and inter-view prediction provided that they are performed with the same or similar process than temporal prediction. Inter prediction or temporal prediction may sometimes be referred to as motion compensation or motion-compensated prediction.

[0059] Intra prediction utilizes the fact that adjacent pixels within the same picture are likely to be correlated. Intra prediction can be performed in spatial or transform domain, i.e., either sample values or transform coefficients can be predicted. Intra prediction is typically exploited in intra coding, where no inter prediction is applied.

[0060] There may be different types of intra prediction modes available in a coding scheme, out of which an encoder can select and indicate the used one, e.g. on block or coding unit basis. A decoder may decode the indicated intra prediction mode and

reconstruct the prediction block accordingly. For example, several angular intra prediction modes, each for different angular direction, may be available. Angular intra prediction may be considered to extrapolate the border samples of adjacent blocks along a linear prediction direction. Additionally or alternatively, a planar prediction mode may be available. Planar prediction may be considered to essentially form a prediction block, in which each sample of a prediction block may be specified to be an average of vertically aligned sample in the adjacent sample column on the left of the current block and the horizontally aligned sample in the adjacent sample line above the current block.

Additionally or alternatively, a DC prediction mode may be available, in which the prediction block is essentially an average sample value of a neighboring block or blocks.

[0061] One outcome of the coding procedure is a set of coding parameters, such as motion vectors and quantized transform coefficients. Many parameters can be entropy-coded more efficiently if they are predicted first from spatially or temporally neighbouring parameters. For example, a motion vector may be predicted from spatially adjacent motion vectors and only the difference relative to the motion vector predictor may be coded. Prediction of coding parameters and intra prediction may be collectively referred to as in-picture prediction.

[0062] Figure 4a shows a block diagram of a video encoder suitable for employing embodiments of the invention. Figure 4a presents an encoder for two layers, but it would be appreciated that presented encoder could be similarly simplified to encode only one layer or extended to encode more than two layers. Figure 4a illustrates an embodiment of a video encoder comprising a first encoder section 500 for a base layer and a second encoder section 502 for an enhancement layer. Each of the first encoder section 500 and the second encoder section 502 may comprise similar elements for encoding incoming pictures. The encoder sections 500, 502 may comprise a pixel predictor 302, 402, prediction error encoder 303, 403 and prediction error decoder 304, 404. Figure 4a also shows an embodiment of the pixel predictor 302, 402 as comprising an inter-predictor 306, 406, an intra-predictor 308, 408, a mode selector 310, 410, a filter 316, 416, and a reference frame memory 318, 418. The pixel predictor 302 of the first encoder section 500 receives 300 base layer images of a video stream to be encoded at both the inter-predictor 306 (which determines the difference between the image and a motion compensated reference frame 318) and the intra-predictor 308 (which determines a prediction for an image block based only on the already processed parts of current frame or picture). The output of both the inter-predictor and the intra-predictor are passed to the mode selector

310. The intra-predictor 308 may have more than one intra-prediction modes. Hence, each mode may perform the intra-prediction and provide the predicted signal to the mode selector 310. The mode selector 310 also receives a copy of the base layer picture 300.

Correspondingly, the pixel predictor 402 of the second encoder section 502 receives 400 enhancement layer images of a video stream to be encoded at both the inter-predictor 406 (which determines the difference between the image and a motion compensated reference frame 418) and the intra-predictor 408 (which determines a prediction for an image block based only on the already processed parts of current frame or picture). The output of both the inter-predictor and the intra-predictor are passed to the mode selector 410. The intra-predictor 408 may have more than one intra-prediction modes. Hence, each mode may perform the intra-prediction and provide the predicted signal to the mode selector 410. The mode selector 410 also receives a copy of the enhancement layer picture 400.

[0063] Depending on which encoding mode is selected to encode the current block, the output of the inter-predictor 306, 406 or the output of one of the optional intra-predictor modes or the output of a surface encoder within the mode selector is passed to the output of the mode selector 310, 410. The output of the mode selector is passed to a first summing device 321, 421. The first summing device may subtract the output of the pixel predictor 302, 402 from the base layer picture 300/enhancement layer picture 400 to produce a first prediction error signal 320, 420 which is input to the prediction error encoder 303, 403.

[0064] The pixel predictor 302, 402 further receives from a preliminary reconstructor 339, 439 the combination of the prediction representation of the image block 312, 412 and the output 338, 438 of the prediction error decoder 304, 404. The preliminary reconstructed image 314, 414 may be passed to the intra-predictor 308, 408 and to a filter 316, 416. The filter 316, 416 receiving the preliminary representation may filter the preliminary representation and output a final reconstructed image 340, 440 which may be saved in a reference frame memory 318, 418. The reference frame memory 318 may be connected to the inter-predictor 306 to be used as the reference image against which a future base layer picture 300 is compared in inter-prediction operations. Subject to the base layer being selected and indicated to be source for inter-layer sample prediction and/or inter-layer motion information prediction of the enhancement layer according to some embodiments, the reference frame memory 318 may also be connected to the inter-predictor 406 to be used as the reference image against which a future enhancement layer pictures 400 is compared in inter-prediction operations. Moreover, the reference frame memory 418 may

be connected to the inter-predictor 406 to be used as the reference image against which a future enhancement layer picture 400 is compared in inter-prediction operations.

[0065] Filtering parameters from the filter 316 of the first encoder section 500 may be provided to the second encoder section 502 subject to the base layer being selected and indicated to be source for predicting the filtering parameters of the enhancement layer according to some embodiments.

[0066] The prediction error encoder 303, 403 comprises a transform unit 342, 442 and a quantizer 344, 444. The transform unit 342, 442 transforms the first prediction error signal 320, 420 to a transform domain. The transform is, for example, the DCT transform. The quantizer 344, 444 quantizes the transform domain signal, e.g. the DCT coefficients, to form quantized coefficients.

[0067] The prediction error decoder 304, 404 receives the output from the prediction error encoder 303, 403 and performs the opposite processes of the prediction error encoder 303, 403 to produce a decoded prediction error signal 338, 438 which, when combined with the prediction representation of the image block 312, 412 at the second summing device 339, 439, produces the preliminary reconstructed image 314, 414. The prediction error decoder may be considered to comprise a dequantizer 361, 461, which dequantizes the quantized coefficient values, e.g. DCT coefficients, to reconstruct the transform signal and an inverse transformation unit 363, 463, which performs the inverse transformation to the reconstructed transform signal wherein the output of the inverse transformation unit 363, 463 contains reconstructed block(s). The prediction error decoder may also comprise a block filter which may filter the reconstructed block(s) according to further decoded information and filter parameters.

[0068] The entropy encoder 330, 430 receives the output of the prediction error encoder 303, 403 and may perform a suitable entropy encoding/variable length encoding on the signal to provide error detection and correction capability. The outputs of the entropy encoders 330, 430 may be inserted into a bitstream e.g. by a multiplexer 508.

[0069] Figure 4b shows a block diagram of a video decoder suitable for employing embodiments of the invention. Figure 8 depicts a structure of a two-layer decoder, but it would be appreciated that the decoding operations may similarly be employed in a single-layer decoder.

[0070] The video decoder 550 comprises a first decoder section 552 for base layer pictures and a second decoder section 554 for enhancement layer pictures. Block 556 illustrates a demultiplexer for delivering information regarding base layer pictures to the first decoder

section 552 and for delivering information regarding enhancement layer pictures to the second decoder section 554. Reference $P'n$ stands for a predicted representation of an image block. Reference $D'n$ stands for a reconstructed prediction error signal. Blocks 704, 804 illustrate preliminary reconstructed images ($I'n$). Reference $R'n$ stands for a final reconstructed image. Blocks 703, 803 illustrate inverse transform (T^{-1}). Blocks 702, 802 illustrate inverse quantization (Q^{-1}). Blocks 700, 800 illustrate entropy decoding (E^{-1}). Blocks 706, 806 illustrate a reference frame memory (RFM). Blocks 707, 807 illustrate prediction (P) (either inter prediction or intra prediction). Blocks 708, 808 illustrate filtering (F). Blocks 709, 809 may be used to combine decoded prediction error information with predicted base or enhancement layer pictures to obtain the preliminary reconstructed images ($I'n$). Preliminary reconstructed and filtered base layer pictures may be output 710 from the first decoder section 552 and preliminary reconstructed and filtered enhancement layer pictures may be output 810 from the second decoder section 554.

[0071] Herein, the decoder could be interpreted to cover any operational unit capable to carry out the decoding operations, such as a player, a receiver, a gateway, a demultiplexer and/or a decoder.

[0072] The H.264/AVC standard was developed by the Joint Video Team (JVT) of the Video Coding Experts Group (VCEG) of the Telecommunications Standardization Sector of International Telecommunication Union (ITU-T) and the Moving Picture Experts Group (MPEG) of International Organisation for Standardization (ISO) / International Electrotechnical Commission (IEC). The H.264/AVC standard is published by both parent standardization organizations, and it is referred to as ITU-T Recommendation H.264 and ISO/IEC International Standard 14496-10, also known as MPEG-4 Part 10 Advanced Video Coding (AVC). There have been multiple versions of the H.264/AVC standard, integrating new extensions or features to the specification. These extensions include Scalable Video Coding (SVC) and Multiview Video Coding (MVC).

[0073] Version 1 of the High Efficiency Video Coding (H.265/HEVC a.k.a. HEVC) standard was developed by the Joint Collaborative Team - Video Coding (JCT-VC) of VCEG and MPEG. The standard was published by both parent standardization organizations, and it is referred to as ITU-T Recommendation H.265 and ISO/IEC International Standard 23008-2, also known as MPEG-H Part 2 High Efficiency Video Coding (HEVC). Version 2 of H.265/HEVC included scalable, multiview, and fidelity range extensions, which may be abbreviated SHVC, MV-HEVC, and REXT, respectively.

Version 2 of H.265/HEVC was published as ITU-T Recommendation H.265 (10/2014) and as Edition 2 of ISO/IEC 23008-2. Further extensions to H.265/HEVC include three-dimensional and screen content coding extensions, which may be abbreviated 3D-HEVC and SCC, respectively.

5 [0074] SHVC, MV-HEVC, and 3D-HEVC use a common basis specification, specified in Annex F of the version 2 of the HEVC standard. This common basis comprises for example high-level syntax and semantics e.g. specifying some of the characteristics of the layers of the bitstream, such as inter-layer dependencies, as well as decoding processes, such as reference picture list construction including inter-layer reference pictures and
10 picture order count derivation for multi-layer bitstream. Annex F may also be used in potential subsequent multi-layer extensions of HEVC. It is to be understood that even though a video encoder, a video decoder, encoding methods, decoding methods, bitstream structures, and/or embodiments may be described in the following with reference to specific extensions, such as SHVC and/or MV-HEVC, they are generally applicable to
15 any multi-layer extensions of HEVC, and even more generally to any multi-layer video coding scheme.

[0075] Some key definitions, bitstream and coding structures, and concepts of H.264/AVC and HEVC are described in this section as an example of a video encoder, decoder, encoding method, decoding method, and a bitstream structure, wherein the embodiments
20 may be implemented. Some of the key definitions, bitstream and coding structures, and concepts of H.264/AVC are the same as in HEVC - hence, they are described below jointly. The aspects of the invention are not limited to H.264/AVC or HEVC, but rather the description is given for one possible basis on top of which the invention may be partly or fully realized.

25 [0076] Similarly to many earlier video coding standards, the bitstream syntax and semantics as well as the decoding process for error-free bitstreams are specified in H.264/AVC and HEVC. The encoding process is not specified, but encoders must generate conforming bitstreams. Bitstream and decoder conformance can be verified with the Hypothetical Reference Decoder (HRD). The standards contain coding tools that help
30 in coping with transmission errors and losses, but the use of the tools in encoding is optional and no decoding process has been specified for erroneous bitstreams.

[0077] In the description of existing standards as well as in the description of example embodiments, a syntax element may be defined as an element of data represented in the bitstream. A syntax structure may be defined as zero or more syntax elements present

together in the bitstream in a specified order. In the description of existing standards as well as in the description of example embodiments, a phrase "by external means" or "through external means" may be used. For example, an entity, such as a syntax structure or a value of a variable used in the decoding process, may be provided "by external means" to the decoding process. The phrase "by external means" may indicate that the entity is not included in the bitstream created by the encoder, but rather conveyed externally from the bitstream for example using a control protocol. It may alternatively or additionally mean that the entity is not created by the encoder, but may be created for example in the player or decoding control logic or alike that is using the decoder. The decoder may have an interface for inputting the external means, such as variable values.

[0078] The elementary unit for the input to an H.264/AVC or HEVC encoder and the output of an H.264/AVC or HEVC decoder, respectively, is a picture. A picture given as an input to an encoder may also referred to as a source picture, and a picture decoded by a decoder may be referred to as a decoded picture.

[0079] The source and decoded pictures are each comprised of one or more sample arrays, such as one of the following sets of sample arrays:

- Luma (Y) only (monochrome).
- Luma and two chroma (YCbCr or YCgCo).
- Green, Blue and Red (GBR, also known as RGB).
- Arrays representing other unspecified monochrome or tri-stimulus color samplings (for example, YZX, also known as XYZ).

[0080] In the following, these arrays may be referred to as luma (or L or Y) and chroma, where the two chroma arrays may be referred to as Cb and Cr; regardless of the actual color representation method in use. The actual color representation method in use can be indicated e.g. in a coded bitstream e.g. using the Video Usability Information (VUI) syntax of H.264/AVC and/or HEVC. A component may be defined as an array or single sample from one of the three sample arrays arrays (luma and two chroma) or the array or a single sample of the array that compose a picture in monochrome format.

[0081] In H.264/AVC and HEVC, a picture may either be a frame or a field. A frame comprises a matrix of luma samples and possibly the corresponding chroma samples. A field is a set of alternate sample rows of a frame and may be used as encoder input, when the source signal is interlaced. Chroma sample arrays may be absent (and hence monochrome sampling may be in use) or chroma sample arrays may be subsampled when compared to luma sample arrays. Chroma formats may be summarized as follows:

- In monochrome sampling there is only one sample array, which may be nominally considered the luma array.
- In 4:2:0 sampling, each of the two chroma arrays has half the height and half the width of the luma array.
- 5 – In 4:2:2 sampling, each of the two chroma arrays has the same height and half the width of the luma array.
- In 4:4:4 sampling when no separate color planes are in use, each of the two chroma arrays has the same height and width as the luma array.

[0082] In H.264/AVC and HEVC, it is possible to code sample arrays as separate color
10 planes into the bitstream and respectively decode separately coded color planes from the bitstream. When separate color planes are in use, each one of them is separately processed (by the encoder and/or the decoder) as a picture with monochrome sampling.

[0083] A partitioning may be defined as a division of a set into subsets such that each element of the set is in exactly one of the subsets.

15 [0084] In H.264/AVC, a macroblock is a 16x16 block of luma samples and the corresponding blocks of chroma samples. For example, in the 4:2:0 sampling pattern, a macroblock contains one 8x8 block of chroma samples per each chroma component. In H.264/AVC, a picture is partitioned to one or more slice groups, and a slice group contains one or more slices. In H.264/AVC, a slice consists of an integer number of
20 macroblocks ordered consecutively in the raster scan within a particular slice group.

[0085] When describing the operation of HEVC encoding and/or decoding, the following terms may be used. A coding block may be defined as an NxN block of samples for some value of N such that the division of a coding tree block into coding blocks is a partitioning. A coding tree block (CTB) may be defined as an NxN block of samples for
25 some value of N such that the division of a component into coding tree blocks is a partitioning. A coding tree unit (CTU) may be defined as a coding tree block of luma samples, two corresponding coding tree blocks of chroma samples of a picture that has three sample arrays, or a coding tree block of samples of a monochrome picture or a picture that is coded using three separate color planes and syntax structures used to code
30 the samples. A coding unit (CU) may be defined as a coding block of luma samples, two corresponding coding blocks of chroma samples of a picture that has three sample arrays, or a coding block of samples of a monochrome picture or a picture that is coded using three separate color planes and syntax structures used to code the samples.

[0086] In some video codecs, such as High Efficiency Video Coding (HEVC) codec, video pictures are divided into coding units (CU) covering the area of the picture. A CU consists of one or more prediction units (PU) defining the prediction process for the samples within the CU and one or more transform units (TU) defining the prediction error coding process for the samples in the said CU. Typically, a CU consists of a square block of samples with a size selectable from a predefined set of possible CU sizes. A CU with the maximum allowed size may be named as LCU (largest coding unit) or coding tree unit (CTU) and the video picture is divided into non-overlapping LCUs. An LCU can be further split into a combination of smaller CUs, e.g. by recursively splitting the LCU and resultant CUs. Each resulting CU typically has at least one PU and at least one TU associated with it. Each PU and TU can be further split into smaller PUs and TUs in order to increase granularity of the prediction and prediction error coding processes, respectively. Each PU has prediction information associated with it defining what kind of a prediction is to be applied for the pixels within that PU (e.g. motion vector information for inter predicted PUs and intra prediction directionality information for intra predicted PUs).

[0087] Each TU can be associated with information describing the prediction error decoding process for the samples within the said TU (including e.g. DCT coefficient information). It may be signalled at CU level whether prediction error coding is applied or not for each CU. In the case there is no prediction error residual associated with the CU, it can be considered there are no TUs for the said CU. The division of the image into CUs, and division of CUs into PUs and TUs may be signalled in the bitstream allowing the decoder to reproduce the intended structure of these units.

[0088] In HEVC, a picture can be partitioned in tiles, which are rectangular and contain an integer number of LCUs. In HEVC, the partitioning to tiles forms a grid comprising one or more tile columns and one or more tile rows. A coded tile is byte-aligned, which may be achieved by adding byte-alignment bits at the end of the coded tile.

[0089] In HEVC, the partitioning to tiles forms a regular grid, where heights and widths of tiles differ from each other by one LCU at the maximum. In HEVC, a slice is defined to be an integer number of coding tree units contained in one independent slice segment and all subsequent dependent slice segments (if any) that precede the next independent slice segment (if any) within the same access unit. In HEVC, a slice segment is defined to be an integer number of coding tree units ordered consecutively in the tile scan and contained in a single NAL unit. The division of each picture into slice segments is a partitioning. In

HEVC, an independent slice segment is defined to be a slice segment for which the values of the syntax elements of the slice segment header are not inferred from the values for a preceding slice segment, and a dependent slice segment is defined to be a slice segment for which the values of some syntax elements of the slice segment header are inferred from the values for the preceding independent slice segment in decoding order. In HEVC, a slice header is defined to be the slice segment header of the independent slice segment that is a current slice segment or is the independent slice segment that precedes a current dependent slice segment, and a slice segment header is defined to be a part of a coded slice segment containing the data elements pertaining to the first or all coding tree units represented in the slice segment. The CUs are scanned in the raster scan order of LCUs within tiles or within a picture, if tiles are not in use. Within an LCU, the CUs have a specific scan order.

[0090] In HEVC, a tile contains an integer number of coding tree units, and may consist of coding tree units contained in more than one slice. Similarly, a slice may consist of coding tree units contained in more than one tile. In HEVC, all coding tree units in a slice belong to the same tile and/or all coding tree units in a tile belong to the same slice. Furthermore, in HEVC, all coding tree units in a slice segment belong to the same tile and/or all coding tree units in a tile belong to the same slice segment.

[0091] A motion-constrained tile set is such that the inter prediction process is constrained in encoding such that no sample value outside the motion-constrained tile set, and no sample value at a fractional sample position that is derived using one or more sample values outside the motion-constrained tile set, is used for inter prediction of any sample within the motion-constrained tile set.

[0092] It is noted that sample locations used in inter prediction are saturated so that a location that would be outside the picture otherwise is saturated to point to the corresponding boundary sample of the picture. Hence, if a tile boundary is also a picture boundary, motion vectors may effectively cross that boundary or a motion vector may effectively cause fractional sample interpolation that would refer to a location outside that boundary, since the sample locations are saturated onto the boundary.

[0093] The temporal motion-constrained tile sets SEI message of HEVC can be used to indicate the presence of motion-constrained tile sets in the bitstream.

[0094] An inter-layer constrained tile set is such that the inter-layer prediction process is constrained in encoding such that no sample value outside each associated reference tile set, and no sample value at a fractional sample position that is derived using one or more

sample values outside each associated reference tile set, is used for inter-layer prediction of any sample within the inter-layer constrained tile set.

[0095] The inter-layer constrained tile sets SEI message of HEVC can be used to indicate the presence of inter-layer constrained tile sets in the bitstream.

5 [0096] The decoder reconstructs the output video by applying prediction means similar to the encoder to form a predicted representation of the pixel blocks (using the motion or spatial information created by the encoder and stored in the compressed representation) and prediction error decoding (inverse operation of the prediction error coding recovering the quantized prediction error signal in spatial pixel domain). After applying prediction
10 and prediction error decoding means the decoder sums up the prediction and prediction error signals (pixel values) to form the output video frame. The decoder (and encoder) can also apply additional filtering means to improve the quality of the output video before passing it for display and/or storing it as prediction reference for the forthcoming frames in the video sequence.

15 [0097] The filtering may for example include one more of the following: deblocking, sample adaptive offset (SAO), and/or adaptive loop filtering (ALF). H.264/AVC includes a deblocking, whereas HEVC includes both deblocking and SAO.

[0098] In typical video codecs the motion information is indicated with motion vectors associated with each motion compensated image block, such as a prediction unit. Each of
20 these motion vectors represents the displacement of the image block in the picture to be coded (in the encoder side) or decoded (in the decoder side) and the prediction source block in one of the previously coded or decoded pictures. In order to represent motion vectors efficiently those are typically coded differentially with respect to block specific predicted motion vectors. In typical video codecs the predicted motion vectors are created
25 in a predefined way, for example calculating the median of the encoded or decoded motion vectors of the adjacent blocks. Another way to create motion vector predictions is to generate a list of candidate predictions from adjacent blocks and/or co-located blocks in temporal reference pictures and signalling the chosen candidate as the motion vector predictor. In addition to predicting the motion vector values, it can be predicted which
30 reference picture(s) are used for motion-compensated prediction and this prediction information may be represented for example by a reference index of previously coded/decoded picture. The reference index is typically predicted from adjacent blocks and/or co-located blocks in temporal reference picture. Moreover, typical high efficiency video codecs employ an additional motion information coding/decoding mechanism, often

called merging/merge mode, where all the motion field information, which includes motion vector and corresponding reference picture index for each available reference picture list, is predicted and used without any modification/correction. Similarly, predicting the motion field information is carried out using the motion field information of adjacent blocks and/or co-located blocks in temporal reference pictures and the used motion field information is signalled among a list of motion field candidate list filled with motion field information of available adjacent/co-located blocks.

[0099] Typical video codecs enable the use of uni-prediction, where a single prediction block is used for a block being (de)coded, and bi-prediction, where two prediction blocks are combined to form the prediction for a block being (de)coded. Some video codecs enable weighted prediction, where the sample values of the prediction blocks are weighted prior to adding residual information. For example, multiplicative weighting factor and an additive offset which can be applied. In explicit weighted prediction, enabled by some video codecs, a weighting factor and offset may be coded for example in the slice header for each allowable reference picture index. In implicit weighted prediction, enabled by some video codecs, the weighting factors and/or offsets are not coded but are derived e.g. based on the relative picture order count (POC) distances of the reference pictures.

[0100] In typical video codecs the prediction residual after motion compensation is first transformed with a transform kernel (like DCT) and then coded. The reason for this is that often there still exists some correlation among the residual and transform can in many cases help reduce this correlation and provide more efficient coding.

[0101] Typical video encoders utilize Lagrangian cost functions to find optimal coding modes, e.g. the desired Macroblock mode and associated motion vectors. This kind of cost function uses a weighting factor λ to tie together the (exact or estimated) image distortion due to lossy coding methods and the (exact or estimated) amount of information that is required to represent the pixel values in an image area:

$$[0102] \quad C = D + \lambda R \quad (1)$$

[0103] where C is the Lagrangian cost to be minimized, D is the image distortion (e.g. Mean Squared Error) with the mode and motion vectors considered, and R the number of bits needed to represent the required data to reconstruct the image block in the decoder (including the amount of data to represent the candidate motion vectors).

[0104] Video coding standards and specifications may allow encoders to divide a coded picture to coded slices or alike. In-picture prediction is typically disabled across slice boundaries. Thus, slices can be regarded as a way to split a coded picture to independently

decodable pieces. In H.264/AVC and HEVC, in-picture prediction may be disabled across slice boundaries. Thus, slices can be regarded as a way to split a coded picture into independently decodable pieces, and slices are therefore often regarded as elementary units for transmission. In many cases, encoders may indicate in the bitstream which types of in-picture prediction are turned off across slice boundaries, and the decoder operation takes this information into account for example when concluding which prediction sources are available. For example, samples from a neighbouring macroblock or CU may be regarded as unavailable for intra prediction, if the neighbouring macroblock or CU resides in a different slice.

[0105] An elementary unit for the output of an H.264/AVC or HEVC encoder and the input of an H.264/AVC or HEVC decoder, respectively, is a Network Abstraction Layer (NAL) unit. For transport over packet-oriented networks or storage into structured files, NAL units may be encapsulated into packets or similar structures. A NAL unit may be defined as a syntax structure containing an indication of the type of data to follow and bytes containing that data in the form of an RBSP interspersed as necessary with startcode emulation prevention bytes. A raw byte sequence payload (RBSP) may be defined as a syntax structure containing an integer number of bytes that is encapsulated in a NAL unit. An RBSP is either empty or has the form of a string of data bits containing syntax elements followed by an RBSP stop bit and followed by zero or more subsequent bits equal to 0. NAL units consist of a header and payload.

[0106] In HEVC, a two-byte NAL unit header is used for all specified NAL unit types. The NAL unit header contains one reserved bit, a six-bit NAL unit type indication, a three-bit `nuh_temporal_id_plus1` indication for temporal level (may be required to be greater than or equal to 1) and a six-bit `nuh_layer_id` syntax element. The `temporal_id_plus1` syntax element may be regarded as a temporal identifier for the NAL unit, and a zero-based `TemporalId` variable may be derived as follows: $\text{TemporalId} = \text{temporal_id_plus1} - 1$. `TemporalId` equal to 0 corresponds to the lowest temporal level. The value of `temporal_id_plus1` is required to be non-zero in order to avoid start code emulation involving the two NAL unit header bytes. The bitstream created by excluding all VCL NAL units having a `TemporalId` greater than or equal to a selected value and including all other VCL NAL units remains conforming. Consequently, a picture having `TemporalId` equal to `TID` does not use any picture having a `TemporalId` greater than `TID` as inter prediction reference. A sub-layer or a temporal sub-layer may be defined to be a temporal scalable layer of a temporal scalable bitstream, consisting of VCL NAL units with a

particular value of the TemporalId variable and the associated non-VCL NAL units.

nuh_layer_id can be understood as a scalability layer identifier.

[0107] NAL units can be categorized into Video Coding Layer (VCL) NAL units and non-VCL NAL units. In H.264/AVC, coded slice NAL units contain syntax elements representing one or more coded macroblocks, each of which corresponds to a block of samples in the uncompressed picture. In HEVC, VCLNAL units contain syntax elements representing one or more CU.

[0108] In HEVC, a coded slice NAL unit can be indicated to be one of the following types:

nal_unit_type	Name of nal_unit_type	Content of NAL unit and RBSP syntax structure
0, 1	TRAIL_N, TRAIL_R	Coded slice segment of a non-TSA, non-STSA trailing picture slice_segment_layer_rbsp()
2, 3	TSA_N, TSA_R	Coded slice segment of a TSA picture slice_segment_layer_rbsp()
4, 5	STSA_N, STSA_R	Coded slice segment of an STSA picture slice_layer_rbsp()
6, 7	RADL_N, RADL_R	Coded slice segment of a RADL picture slice_layer_rbsp()
8, 9	RASL_N, RASL_R,	Coded slice segment of a RASL picture slice_layer_rbsp()
10, 12, 14	RSV_VCL_N10 RSV_VCL_N12 RSV_VCL_N14	Reserved // reserved non-RAP non-reference VCL NAL unit types
11, 13, 15	RSV_VCL_R11 RSV_VCL_R13 RSV_VCL_R15	Reserved // reserved non-RAP reference VCL NAL unit types
16, 17, 18	BLA_W_LP IDR_W_RADL BLA_N_LP	Coded slice segment of a BLA picture slice_segment_layer_rbsp()

19, 20	IDR_W_RADL IDR_N_LP	Coded slice segment of an IDR picture slice_segment_layer_rbsp()
21	CRA_NUT	Coded slice segment of a CRA picture slice_segment_layer_rbsp()
22, 23	RSV_IRAP_VCL22.. RSV_IRAP_VCL23	Reserved // reserved RAP VCL NAL unit types
24..31	RSV_VCL24.. RSV_VCL31	Reserved // reserved non-RAP VCL NAL unit types

[0109] In HEVC, abbreviations for picture types may be defined as follows: trailing (TRAIL) picture, Temporal Sub-layer Access (TSA), Step-wise Temporal Sub-layer Access (STSA), Random Access Decodable Leading (RADL) picture, Random Access Skipped Leading (RASL) picture, Broken Link Access (BLA) picture, Instantaneous Decoding Refresh (IDR) picture, Clean Random Access (CRA) picture.

[0110] A Random Access Point (RAP) picture, which may also be referred to as an intra random access point (IRAP) picture, is a picture where each slice or slice segment has nal_unit_type in the range of 16 to 23, inclusive. A IRAP picture in an independent layer does not refer to any pictures other than itself for inter prediction in its decoding process. When no intra block copy is in use, an IRAP picture in an independent layer contains only intra-coded slices. An IRAP picture belonging to a predicted layer with nuh_layer_id value currLayerId may contain P, B, and I slices, cannot use inter prediction from other pictures with nuh_layer_id equal to currLayerId, and may use inter-layer prediction from its direct reference layers. In the present version of HEVC, an IRAP picture may be a BLA picture, a CRA picture or an IDR picture. The first picture in a bitstream containing a base layer is an IRAP picture at the base layer. Provided the necessary parameter sets are available when they need to be activated, an IRAP picture at an independent layer and all subsequent non-RASL pictures at the independent layer in decoding order can be correctly decoded without performing the decoding process of any pictures that precede the IRAP picture in decoding order. The IRAP picture belonging to a predicted layer with nuh_layer_id value currLayerId and all subsequent non-RASL pictures with nuh_layer_id equal to currLayerId in decoding order can be correctly decoded without performing the decoding process of any pictures with nuh_layer_id equal to currLayerId that precede the

IRAP picture in decoding order, when the necessary parameter sets are available when they need to be activated and when the decoding of each direct reference layer of the layer with nuh_layer_id equal to currLayerId has been initialized (i.e. when LayerInitializedFlag[refLayerId] is equal to 1 for refLayerId equal to all nuh_layer_id values of the direct reference layers of the layer with nuh_layer_id equal to currLayerId). There may be pictures in a bitstream that contain only intra-coded slices that are not IRAP pictures.

[0111] In HEVC a CRA picture may be the first picture in the bitstream in decoding order, or may appear later in the bitstream. CRA pictures in HEVC allow so-called leading pictures that follow the CRA picture in decoding order but precede it in output order. Some of the leading pictures, so-called RASL pictures, may use pictures decoded before the CRA picture as a reference. Pictures that follow a CRA picture in both decoding and output order are decodable if random access is performed at the CRA picture, and hence clean random access is achieved similarly to the clean random access functionality of an IDR picture.

[0112] A CRA picture may have associated RADL or RASL pictures. When a CRA picture is the first picture in the bitstream in decoding order, the CRA picture is the first picture of a coded video sequence in decoding order, and any associated RASL pictures are not output by the decoder and may not be decodable, as they may contain references to pictures that are not present in the bitstream.

[0113] A leading picture is a picture that precedes the associated RAP picture in output order. The associated RAP picture is the previous RAP picture in decoding order (if present). A leading picture is either a RADL picture or a RASL picture.

[0114] All RASL pictures are leading pictures of an associated BLA or CRA picture.

When the associated RAP picture is a BLA picture or is the first coded picture in the bitstream, the RASL picture is not output and may not be correctly decodable, as the RASL picture may contain references to pictures that are not present in the bitstream. However, a RASL picture can be correctly decoded if the decoding had started from a RAP picture before the associated RAP picture of the RASL picture. RASL pictures are not used as reference pictures for the decoding process of non-RASL pictures. When present, all RASL pictures precede, in decoding order, all trailing pictures of the same associated RAP picture. In some drafts of the HEVC standard, a RASL picture was referred to a Tagged for Discard (TFD) picture.

[0115] All RADL pictures are leading pictures. RADL pictures are not used as reference pictures for the decoding process of trailing pictures of the same associated RAP picture. When present, all RADL pictures precede, in decoding order, all trailing pictures of the same associated RAP picture. RADL pictures do not refer to any picture preceding the associated RAP picture in decoding order and can therefore be correctly decoded when the decoding starts from the associated RAP picture.

[0116] When a part of a bitstream starting from a CRA picture is included in another bitstream, the RASL pictures associated with the CRA picture might not be correctly decodable, because some of their reference pictures might not be present in the combined bitstream. To make such a splicing operation straightforward, the NAL unit type of the CRA picture can be changed to indicate that it is a BLA picture. The RASL pictures associated with a BLA picture may not be correctly decodable hence are not be output/displayed. Furthermore, the RASL pictures associated with a BLA picture may be omitted from decoding.

[0117] A BLA picture may be the first picture in the bitstream in decoding order, or may appear later in the bitstream. Each BLA picture begins a new coded video sequence, and has similar effect on the decoding process as an IDR picture. However, a BLA picture contains syntax elements that specify a non-empty reference picture set. When a BLA picture has nal_unit_type equal to BLA_W_LP, it may have associated RASL pictures, which are not output by the decoder and may not be decodable, as they may contain references to pictures that are not present in the bitstream. When a BLA picture has nal_unit_type equal to BLA_W_LP, it may also have associated RADL pictures, which are specified to be decoded. When a BLA picture has nal_unit_type equal to BLA_W_RADL, it does not have associated RASL pictures but may have associated RADL pictures, which are specified to be decoded. When a BLA picture has nal_unit_type equal to BLA_N_LP, it does not have any associated leading pictures.

[0118] An IDR picture having nal_unit_type equal to IDR_N_LP does not have associated leading pictures present in the bitstream. An IDR picture having nal_unit_type equal to IDR_W_LP does not have associated RASL pictures present in the bitstream, but may have associated RADL pictures in the bitstream.

[0119] When the value of nal_unit_type is equal to TRAIL_N, TSA_N, STSA_N, RADL_N, RASL_N, RSV_VCL_N10, RSV_VCL_N12, or RSV_VCL_N14, the decoded picture is not used as a reference for any other picture of the same temporal sub-layer. That is, in HEVC, when the value of nal_unit_type is equal to TRAIL_N, TSA_N,

STSA_N, RADL_N, RASL_N, RSV_VCL_N10, RSV_VCL_N12, or RSV_VCL_N14, the decoded picture is not included in any of RefPicSetStCurrBefore, RefPicSetStCurrAfter and RefPicSetLtCurr of any picture with the same value of TemporalId. A coded picture with nal_unit_type equal to TRAIL_N, TSA_N, STSA_N, RADL_N, RASL_N, RSV_VCL_N10, RSV_VCL_N12, or RSV_VCL_N14 may be discarded without affecting the decodability of other pictures with the same value of TemporalId.

[0120] A trailing picture may be defined as a picture that follows the associated RAP picture in output order. Any picture that is a trailing picture does not have nal_unit_type equal to RADL_N, RADL_R, RASL_N or RASL_R. Any picture that is a leading picture may be constrained to precede, in decoding order, all trailing pictures that are associated with the same RAP picture. No RASL pictures are present in the bitstream that are associated with a BLA picture having nal_unit_type equal to BLA_W_RADL or BLA_N_LP. No RADL pictures are present in the bitstream that are associated with a BLA picture having nal_unit_type equal to BLA_N_LP or that are associated with an IDR picture having nal_unit_type equal to IDR_N_LP. Any RASL picture associated with a CRA or BLA picture may be constrained to precede any RADL picture associated with the CRA or BLA picture in output order. Any RASL picture associated with a CRA picture may be constrained to follow, in output order, any other RAP picture that precedes the CRA picture in decoding order.

[0121] In HEVC there are two picture types, the TSA and STSA picture types that can be used to indicate temporal sub-layer switching points. If temporal sub-layers with TemporalId up to N had been decoded until the TSA or STSA picture (exclusive) and the TSA or STSA picture has TemporalId equal to N+1, the TSA or STSA picture enables decoding of all subsequent pictures (in decoding order) having TemporalId equal to N+1. The TSA picture type may impose restrictions on the TSA picture itself and all pictures in the same sub-layer that follow the TSA picture in decoding order. None of these pictures is allowed to use inter prediction from any picture in the same sub-layer that precedes the TSA picture in decoding order. The TSA definition may further impose restrictions on the pictures in higher sub-layers that follow the TSA picture in decoding order. None of these pictures is allowed to refer a picture that precedes the TSA picture in decoding order if that picture belongs to the same or higher sub-layer as the TSA picture. TSA pictures have TemporalId greater than 0. The STSA is similar to the TSA picture but does not impose restrictions on the pictures in higher sub-layers that follow the STSA picture in decoding

order and hence enable up-switching only onto the sub-layer where the STSA picture resides.

[0122] A non-VCL NAL unit may be for example one of the following types: a sequence parameter set, a picture parameter set, a supplemental enhancement information (SEI) NAL unit, an access unit delimiter, an end of sequence NAL unit, an end of bitstream NAL unit, or a filler data NAL unit. Parameter sets may be needed for the reconstruction of decoded pictures, whereas many of the other non-VCL NAL units are not necessary for the reconstruction of decoded sample values.

[0123] Parameters that remain unchanged through a coded video sequence may be included in a sequence parameter set. In addition to the parameters that may be needed by the decoding process, the sequence parameter set may optionally contain video usability information (VUI), which includes parameters that may be important for buffering, picture output timing, rendering, and resource reservation. In HEVC a sequence parameter set RBSP includes parameters that can be referred to by one or more picture parameter set RBSPs or one or more SEI NAL units containing a buffering period SEI message. A picture parameter set contains such parameters that are likely to be unchanged in several coded pictures. A picture parameter set RBSP may include parameters that can be referred to by the coded slice NAL units of one or more coded pictures.

[0124] In HEVC, a video parameter set (VPS) may be defined as a syntax structure containing syntax elements that apply to zero or more entire coded video sequences as determined by the content of a syntax element found in the SPS referred to by a syntax element found in the PPS referred to by a syntax element found in each slice segment header. A video parameter set RBSP may include parameters that can be referred to by one or more sequence parameter set RBSPs.

[0125] The relationship and hierarchy between video parameter set (VPS), sequence parameter set (SPS), and picture parameter set (PPS) may be described as follows. VPS resides one level above SPS in the parameter set hierarchy and in the context of scalability and/or 3D video. VPS may include parameters that are common for all slices across all (scalability or view) layers in the entire coded video sequence. SPS includes the parameters that are common for all slices in a particular (scalability or view) layer in the entire coded video sequence, and may be shared by multiple (scalability or view) layers. PPS includes the parameters that are common for all slices in a particular layer representation (the representation of one scalability or view layer in one access unit) and are likely to be shared by all slices in multiple layer representations.

[0126] VPS may provide information about the dependency relationships of the layers in a bitstream, as well as many other information that are applicable to all slices across all (scalability or view) layers in the entire coded video sequence. VPS may be considered to comprise two parts, the base VPS and a VPS extension, where the VPS extension may be optionally present. In HEVC, the base VPS may be considered to comprise the video_parameter_set_rbsp() syntax structure without the vps_extension() syntax structure. The video_parameter_set_rbsp() syntax structure was primarily specified already for HEVC version 1 and includes syntax elements which may be of use for base layer decoding. In HEVC, the VPS extension may be considered to comprise the vps_extension() syntax structure. The vps_extension() syntax structure was specified in HEVC version 2 primarily for multi-layer extensions and comprises syntax elements which may be of use for decoding of one or more non-base layers, such as syntax elements indicating layer dependency relations.

[0127] H.264/AVC and HEVC syntax allows many instances of parameter sets, and each instance is identified with a unique identifier. In order to limit the memory usage needed for parameter sets, the value range for parameter set identifiers has been limited. In H.264/AVC and HEVC, each slice header includes the identifier of the picture parameter set that is active for the decoding of the picture that contains the slice, and each picture parameter set contains the identifier of the active sequence parameter set. Consequently, the transmission of picture and sequence parameter sets does not have to be accurately synchronized with the transmission of slices. Instead, it is sufficient that the active sequence and picture parameter sets are received at any moment before they are referenced, which allows transmission of parameter sets "out-of-band" using a more reliable transmission mechanism compared to the protocols used for the slice data. For example, parameter sets can be included as a parameter in the session description for Real-time Transport Protocol (RTP) sessions. If parameter sets are transmitted in-band, they can be repeated to improve error robustness.

[0128] Out-of-band transmission, signalling or storage can additionally or alternatively be used for other purposes than tolerance against transmission errors, such as ease of access or session negotiation. For example, a sample entry of a track in a file conforming to the ISOBMFF may comprise parameter sets, while the coded data in the bitstream is stored elsewhere in the file or in another file. The phrase along the bitstream (e.g. indicating along the bitstream) may be used in claims and described embodiments to refer to out-of-band transmission, signalling, or storage in a manner that the out-of-band data is

associated with the bitstream. The phrase decoding along the bitstream or alike may refer to decoding the referred out-of-band data (which may be obtained from out-of-band transmission, signalling, or storage) that is associated with the bitstream. A coded picture is a coded representation of a picture.

5 [0129] In HEVC, a coded picture may be defined as a coded representation of a picture containing all coding tree units of the picture. In HEVC, an access unit (AU) may be defined as a set of NAL units that are associated with each other according to a specified classification rule, are consecutive in decoding order, and contain at most one picture with any specific value of `nuh_layer_id`. In addition to containing the VCL NAL units of the
10 coded picture, an access unit may also contain non-VCL NAL units.

[0130] It may be required that coded pictures appear in certain order within an access unit. For example a coded picture with `nuh_layer_id` equal to `nuhLayerIdA` may be required to precede, in decoding order, all coded pictures with `nuh_layer_id` greater than `nuhLayerIdA` in the same access unit. An AU typically contains all the coded pictures that
15 represent the same output time and/or capturing time.

[0131] A bitstream may be defined as a sequence of bits, in the form of a NAL unit stream or a byte stream, that forms the representation of coded pictures and associated data forming one or more coded video sequences. A first bitstream may be followed by a second bitstream in the same logical channel, such as in the same file or in the same
20 connection of a communication protocol. An elementary stream (in the context of video coding) may be defined as a sequence of one or more bitstreams. The end of the first bitstream may be indicated by a specific NAL unit, which may be referred to as the end of bitstream (EOB) NAL unit and which is the last NAL unit of the bitstream. In HEVC and its current draft extensions, the EOB NAL unit is required to have `nuh_layer_id` equal to
25 0.

[0132] A byte stream format has been specified in H.264/AVC and HEVC for transmission or storage environments that do not provide framing structures. The byte stream format separates NAL units from each other by attaching a start code in front of each NAL unit. To avoid false detection of NAL unit boundaries, encoders run a byte-oriented start code emulation prevention algorithm, which adds an emulation prevention byte to the NAL unit
30 payload if a start code would have occurred otherwise. In order to, for example, enable straightforward gateway operation between packet- and stream-oriented systems, start code emulation prevention may always be performed regardless of whether the byte stream format is in use or not. The bit order for the byte stream format may be specified to

start with the most significant bit (MSB) of the first byte, proceed to the least significant bit (LSB) of the first byte, followed by the MSB of the second byte, etc. The byte stream format may be considered to consist of a sequence of byte stream NAL unit syntax structures. Each byte stream NAL unit syntax structure may be considered to contain one start code prefix followed by one NAL unit syntax structure, i.e. the nal_unit(NumBytesInNalUnit) syntax structure if syntax element names are referred to. A byte stream NAL unit may also contain an additional zero_byte syntax element. It may also contain one or more additional trailing_zero_8bits syntax elements. When a byte stream NAL unit is the first byte stream NAL unit in the bitstream, it may also contain one or more additional leading_zero_8bits syntax elements. The syntax of a byte stream NAL unit may be specified as follows:

byte_stream_nal_unit(NumBytesInNalUnit) {	Descriptor
while(next_bits(24) != 0x000001 && next_bits(32) != 0x00000001)	
leading_zero_8bits /* equal to 0x00 */	f(8)
if(next_bits(24) != 0x000001)	
zero_byte /* equal to 0x00 */	f(8)
start_code_prefix_one_3bytes /* equal to 0x000001 */	f(24)
nal_unit(NumBytesInNalUnit)	
while(more_data_in_byte_stream() && next_bits(24) != 0x000001 && next_bits(32) != 0x00000001)	
trailing_zero_8bits /* equal to 0x00 */	f(8)
}	

[0133] The order of byte stream NAL units in the byte stream may be required to follow the decoding order of the NAL units contained in the byte stream NAL units. The semantics of syntax elements may be specified as follows. leading_zero_8bits is a byte equal to 0x00. The leading_zero_8bits syntax element can only be present in the first byte stream NAL unit of the bitstream, because any bytes equal to 0x00 that follow a NAL unit syntax structure and precede the four-byte sequence 0x00000001 (which is to be interpreted as a zero_byte followed by a start_code_prefix_one_3bytes) will be considered to be trailing_zero_8bits syntax elements that are part of the preceding byte stream NAL unit. zero_byte is a single byte equal to 0x00. start_code_prefix_one_3bytes is a fixed-value sequence of 3 bytes equal to 0x000001. This syntax element may be called a start code prefix (or simply a start code). trailing_zero_8bits is a byte equal to 0x00.

[0134] A NAL unit may be defined as a syntax structure containing an indication of the type of data to follow and bytes containing that data in the form of an RBSP interspersed as necessary with emulation prevention bytes. A raw byte sequence payload (RBSP) may be defined as a syntax structure containing an integer number of bytes that is encapsulated in a NAL unit. An RBSP is either empty or has the form of a string of data bits containing syntax elements followed by an RBSP stop bit and followed by zero or more subsequent bits equal to 0.

[0135] NAL units consist of a header and payload. In H.264/AVC and HEVC, the NAL unit header indicates the type of the NAL unit.

10 [0136] The HEVC syntax of the nal_unit(NumBytesInNalUnit) syntax structure are provided next as an example of a syntax of NAL unit.

nal_unit(NumBytesInNalUnit) {	Descriptor
nal_unit_header()	
NumBytesInRbsp = 0	
for(i = 2; i < NumBytesInNalUnit; i++)	
if(i + 2 < NumBytesInNalUnit && next_bits(24) == 0x000003) {	
rbsp_byte [NumBytesInRbsp++]	b(8)
rbsp_byte [NumBytesInRbsp++]	b(8)
i += 2	
emulation_prevention_three_byte /* equal to 0x03 */	f(8)
} else	
rbsp_byte [NumBytesInRbsp++]	b(8)
}	

[0137] In HEVC, a coded video sequence (CVS) may be defined, for example, as a sequence of access units that consists, in decoding order, of an IRAP access unit with NoRaslOutputFlag equal to 1, followed by zero or more access units that are not IRAP access units with NoRaslOutputFlag equal to 1, including all subsequent access units up to but not including any subsequent access unit that is an IRAP access unit with NoRaslOutputFlag equal to 1. An IRAP access unit may be defined as an access unit in which the base layer picture is an IRAP picture. The value of NoRaslOutputFlag is equal to 1 for each IDR picture, each BLA picture, and each IRAP picture that is the first picture in that particular layer in the bitstream in decoding order, is the first IRAP picture that follows an end of sequence NAL unit having the same value of nuh_layer_id in decoding order. In multi-layer HEVC, the value of NoRaslOutputFlag is equal to 1 for each IRAP picture when its nuh_layer_id is such that LayerInitializedFlag[nuh_layer_id] is equal to 0 and LayerInitializedFlag[refLayerId] is equal to 1 for all values of refLayerId equal to IdDirectRefLayer[nuh_layer_id][j], where j is in the range of 0 to

NumDirectRefLayers[nuh_layer_id] ? 1, inclusive. Otherwise, the value of NoRaslOutputFlag is equal to HandleCraAsBlaFlag. NoRaslOutputFlag equal to 1 has an impact that the RASL pictures associated with the IRAP picture for which the NoRaslOutputFlag is set are not output by the decoder. There may be means to provide the value of HandleCraAsBlaFlag to the decoder from an external entity, such as a player or a receiver, which may control the decoder. HandleCraAsBlaFlag may be set to 1 for example by a player that seeks to a new position in a bitstream or tunes into a broadcast and starts decoding and then starts decoding from a CRA picture. When HandleCraAsBlaFlag is equal to 1 for a CRA picture, the CRA picture is handled and decoded as if it were a BLA picture.

[0138] In HEVC, a coded video sequence may additionally or alternatively (to the specification above) be specified to end, when a specific NAL unit, which may be referred to as an end of sequence (EOS) NAL unit, appears in the bitstream and has nuh_layer_id equal to 0.

[0139] A group of pictures (GOP) and its characteristics may be defined as follows. A GOP can be decoded regardless of whether any previous pictures were decoded. An open GOP is such a group of pictures in which pictures preceding the initial intra picture in output order might not be correctly decodable when the decoding starts from the initial intra picture of the open GOP. In other words, pictures of an open GOP may refer (in inter prediction) to pictures belonging to a previous GOP. An HEVC decoder can recognize an intra picture starting an open GOP, because a specific NAL unit type, CRA NAL unit type, may be used for its coded slices. A closed GOP is such a group of pictures in which all pictures can be correctly decoded when the decoding starts from the initial intra picture of the closed GOP. In other words, no picture in a closed GOP refers to any pictures in previous GOPs. In H.264/AVC and HEVC, a closed GOP may start from an IDR picture. In HEVC a closed GOP may also start from a BLA_W_RADL or a BLA_N_LP picture. An open GOP coding structure is potentially more efficient in the compression compared to a closed GOP coding structure, due to a larger flexibility in selection of reference pictures.

[0140] A Structure of Pictures (SOP) may be defined as one or more coded pictures consecutive in decoding order, in which the first coded picture in decoding order is a reference picture at the lowest temporal sub-layer and no coded picture except potentially the first coded picture in decoding order is a RAP picture. All pictures in the previous SOP precede in decoding order all pictures in the current SOP and all pictures in the next

SOP succeed in decoding order all pictures in the current SOP. A SOP may represent a hierarchical and repetitive inter prediction structure. The term group of pictures (GOP) may sometimes be used interchangeably with the term SOP and having the same semantics as the semantics of SOP.

5 [0141] The bitstream syntax of H.264/AVC and HEVC indicates whether a particular picture is a reference picture for inter prediction of any other picture. Pictures of any coding type (I, P, B) can be reference pictures or non-reference pictures in H.264/AVC and HEVC.

[0142] In HEVC, a reference picture set (RPS) syntax structure and decoding process are
10 used. A reference picture set valid or active for a picture includes all the reference pictures used as reference for the picture and all the reference pictures that are kept marked as "used for reference" for any subsequent pictures in decoding order. There are six subsets of the reference picture set, which are referred to as namely RefPicSetStCurr0 (a.k.a. RefPicSetStCurrBefore), RefPicSetStCurr1 (a.k.a. RefPicSetStCurrAfter),
15 RefPicSetStFoll0, RefPicSetStFoll1, RefPicSetLtCurr, and RefPicSetLtFoll. RefPicSetStFoll0 and RefPicSetStFoll1 may also be considered to form jointly one subset RefPicSetStFoll. The notation of the six subsets is as follows. "Curr" refers to reference pictures that are included in the reference picture lists of the current picture and hence may be used as inter prediction reference for the current picture. "Foll" refers to reference
20 pictures that are not included in the reference picture lists of the current picture but may be used in subsequent pictures in decoding order as reference pictures. "St" refers to short-term reference pictures, which may generally be identified through a certain number of least significant bits of their POC value. "Lt" refers to long-term reference pictures, which are specifically identified and generally have a greater difference of POC values relative
25 to the current picture than what can be represented by the mentioned certain number of least significant bits. "0" refers to those reference pictures that have a smaller POC value than that of the current picture. "1" refers to those reference pictures that have a greater POC value than that of the current picture. RefPicSetStCurr0, RefPicSetStCurr1, RefPicSetStFoll0 and RefPicSetStFoll1 are collectively referred to as the short-term
30 subset of the reference picture set. RefPicSetLtCurr and RefPicSetLtFoll are collectively referred to as the long-term subset of the reference picture set.

[0143] In HEVC, a reference picture set may be specified in a sequence parameter set and taken into use in the slice header through an index to the reference picture set. A reference picture set may also be specified in a slice header. A reference picture set may be coded

independently or may be predicted from another reference picture set (known as inter-RPS prediction). In both types of reference picture set coding, a flag (used_by_curr_pic_X_flag) is additionally sent for each reference picture indicating whether the reference picture is used for reference by the current picture (included in a *Curr list) or not (included in a *Foll list). Pictures that are included in the reference picture set used by the current slice are marked as "used for reference", and pictures that are not in the reference picture set used by the current slice are marked as "unused for reference". If the current picture is an IDR picture, RefPicSetStCurr0, RefPicSetStCurr1, RefPicSetStFoll0, RefPicSetStFoll1, RefPicSetLtCurr, and RefPicSetLtFoll are all set to empty.

[0144] A Decoded Picture Buffer (DPB) may be used in the encoder and/or in the decoder. There are two reasons to buffer decoded pictures, for references in inter prediction and for reordering decoded pictures into output order. As H.264/AVC and HEVC provide a great deal of flexibility for both reference picture marking and output reordering, separate buffers for reference picture buffering and output picture buffering may waste memory resources. Hence, the DPB may include a unified decoded picture buffering process for reference pictures and output reordering. A decoded picture may be removed from the DPB when it is no longer used as a reference and is not needed for output.

[0145] In many coding modes of H.264/AVC and HEVC, the reference picture for inter prediction is indicated with an index to a reference picture list. The index may be coded with variable length coding, which usually causes a smaller index to have a shorter value for the corresponding syntax element. In H.264/AVC and HEVC, two reference picture lists (reference picture list 0 and reference picture list 1) are generated for each bi-predictive (B) slice, and one reference picture list (reference picture list 0) is formed for each inter-coded (P) slice.

[0146] A reference picture list, such as reference picture list 0 and reference picture list 1, is typically constructed in two steps: First, an initial reference picture list is generated. The initial reference picture list may be generated for example on the basis of frame_num, POC, temporal_id (or TemporalId or alike), or information on the prediction hierarchy such as GOP structure, or any combination thereof. Second, the initial reference picture list may be reordered by reference picture list reordering (RPLR) commands, also known as reference picture list modification syntax structure, which may be contained in slice headers. If reference picture sets are used, the reference picture list 0 may be initialized to contain RefPicSetStCurr0 first, followed by RefPicSetStCurr1, followed by

RefPicSetLtCurr. Reference picture list 1 may be initialized to contain RefPicSetStCurr1 first, followed by RefPicSetStCurr0. In HEVC, the initial reference picture lists may be modified through the reference picture list modification syntax structure, where pictures in the initial reference picture lists may be identified through an entry index to the list. In other words, in HEVC, reference picture list modification is encoded into a syntax structure comprising a loop over each entry in the final reference picture list, where each loop entry is a fixed-length coded index to the initial reference picture list and indicates the picture in ascending position order in the final reference picture list.

[0147] Many coding standards, including H.264/AVC and HEVC, may have decoding process to derive a reference picture index to a reference picture list, which may be used to indicate which one of the multiple reference pictures is used for inter prediction for a particular block. A reference picture index may be coded by an encoder into the bitstream in some inter coding modes or it may be derived (by an encoder and a decoder) for example using neighbouring blocks in some other inter coding modes.

[0148] In order to represent motion vectors efficiently in bitstreams, motion vectors may be coded differentially with respect to a block-specific predicted motion vector. In many video codecs, the predicted motion vectors are created in a predefined way, for example by calculating the median of the encoded or decoded motion vectors of the adjacent blocks. Another way to create motion vector predictions, sometimes referred to as advanced motion vector prediction (AMVP), is to generate a list of candidate predictions from adjacent blocks and/or co-located blocks in temporal reference pictures and signalling the chosen candidate as the motion vector predictor. In addition to predicting the motion vector values, the reference index of previously coded/decoded picture can be predicted. The reference index is typically predicted from adjacent blocks and/or co-located blocks in temporal reference picture. Differential coding of motion vectors is typically disabled across slice boundaries.

[0149] The width and height of a decoded picture may have certain constraints, e.g. so that the width and height are multiples of a (minimum) coding unit size. For example, HEVC the width and height of a decoded picture are multiples of 8 luma samples. If the encoded picture has extents that do not fulfil such constraints, the (de)coding may still be performed with a picture size complying with the constraints but the output may be performed by cropping the unnecessary sample lines and columns. In HEVC, this cropping can be controlled by the encoder using the so-called conformance cropping window feature. The conformance cropping window is specified (by the encoder) in the

SPS and when outputting the pictures the decoder is required to crop the decoded pictures according to the conformance cropping window.

[0150] Scalable video coding may refer to coding structure where one bitstream can contain multiple representations of the content, for example, at different bitrates, resolutions or frame rates. In these cases the receiver can extract the desired representation depending on its characteristics (e.g. resolution that matches best the display device). Alternatively, a server or a network element can extract the portions of the bitstream to be transmitted to the receiver depending on e.g. the network characteristics or processing capabilities of the receiver. A meaningful decoded representation can be produced by decoding only certain parts of a scalable bitstream. A scalable bitstream typically consists of a "base layer" providing the lowest quality video available and one or more enhancement layers that enhance the video quality when received and decoded together with the lower layers. In order to improve coding efficiency for the enhancement layers, the coded representation of that layer typically depends on the lower layers. E.g. the motion and mode information of the enhancement layer can be predicted from lower layers. Similarly the pixel data of the lower layers can be used to create prediction for the enhancement layer.

[0151] In some scalable video coding schemes, a video signal can be encoded into a base layer and one or more enhancement layers. An enhancement layer may enhance, for example, the temporal resolution (i.e., the frame rate), the spatial resolution, or simply the quality of the video content represented by another layer or part thereof. Each layer together with all its dependent layers is one representation of the video signal, for example, at a certain spatial resolution, temporal resolution and quality level. In this document, we refer to a scalable layer together with all of its dependent layers as a "scalable layer representation". The portion of a scalable bitstream corresponding to a scalable layer representation can be extracted and decoded to produce a representation of the original signal at certain fidelity.

[0152] Scalability modes or scalability dimensions may include but are not limited to the following:

- Quality scalability: Base layer pictures are coded at a lower quality than enhancement layer pictures, which may be achieved for example using a greater quantization parameter value (i.e., a greater quantization step size for transform coefficient quantization) in the base layer than in the enhancement layer.

- Spatial scalability: Base layer pictures are coded at a lower resolution (i.e. have fewer samples) than enhancement layer pictures. Spatial scalability and quality scalability, particularly its coarse-grain scalability type, may sometimes be considered the same type of scalability.
- 5 – Bit-depth scalability: Base layer pictures are coded at lower bit-depth (e.g. 8 bits) than enhancement layer pictures (e.g. 10 or 12 bits).
- Dynamic range scalability: Scalable layers represent a different dynamic range and/or images obtained using a different tone mapping function and/or a different optical transfer function.
- 10 – Chroma format scalability: Base layer pictures provide lower spatial resolution in chroma sample arrays (e.g. coded in 4:2:0 chroma format) than enhancement layer pictures (e.g. 4:4:4 format).
- Color gamut scalability: enhancement layer pictures have a richer/broader color representation range than that of the base layer pictures - for example the enhancement
- 15 layer may have UHD TV (ITU-R BT.2020) color gamut and the base layer may have the ITU-R BT.709 color gamut.
- View scalability, which may also be referred to as multiview coding. The base layer represents a first view, whereas an enhancement layer represents a second view.
- Depth scalability, which may also be referred to as depth-enhanced coding. A layer or
- 20 some layers of a bitstream may represent texture view(s), while other layer or layers may represent depth view(s).
- Region-of-interest scalability (as described below).
- Interlaced-to-progressive scalability (also known as field-to-frame scalability): coded interlaced source content material of the base layer is enhanced with an enhancement
- 25 layer to represent progressive source content.
- Hybrid codec scalability (also known as coding standard scalability): In hybrid codec scalability, the bitstream syntax, semantics and decoding process of the base layer and the enhancement layer are specified in different video coding standards. Thus, base
- 30 layer pictures are coded according to a different coding standard or format than enhancement layer pictures. For example, the base layer may be coded with H.264/AVC and an enhancement layer may be coded with an HEVC multi-layer extension.

[0153] It should be understood that many of the scalability types may be combined and applied together. For example color gamut scalability and bit-depth scalability may be combined.

[0154] The term layer may be used in context of any type of scalability, including view scalability and depth enhancements. An enhancement layer may refer to any type of an enhancement, such as SNR, spatial, multiview, depth, bit-depth, chroma format, and/or color gamut enhancement. A base layer may refer to any type of a base video sequence, such as a base view, a base layer for SNR/spatial scalability, or a texture base view for depth-enhanced video coding.

[0155] Various technologies for providing three-dimensional (3D) video content are currently investigated and developed. It may be considered that in stereoscopic or two-view video, one video sequence or view is presented for the left eye while a parallel view is presented for the right eye. More than two parallel views may be needed for applications which enable viewpoint switching or for autostereoscopic displays which may present a large number of views simultaneously and let the viewers to observe the content from different viewpoints.

[0156] A view may be defined as a sequence of pictures representing one camera or viewpoint. The pictures representing a view may also be called view components. In other words, a view component may be defined as a coded representation of a view in a single access unit. In multiview video coding, more than one view is coded in a bitstream. Since views are typically intended to be displayed on stereoscopic or multiview autostereoscopic display or to be used for other 3D arrangements, they typically represent the same scene and are content-wise partly overlapping although representing different viewpoints to the content. Hence, inter-view prediction may be utilized in multiview video coding to take advantage of inter-view correlation and improve compression efficiency. One way to realize inter-view prediction is to include one or more decoded pictures of one or more other views in the reference picture list(s) of a picture being coded or decoded residing within a first view. View scalability may refer to such multiview video coding or multiview video bitstreams, which enable removal or omission of one or more coded views, while the resulting bitstream remains conforming and represents video with a smaller number of views than originally. Region of Interest (ROI) coding may be defined to refer to coding a particular region within a video at a higher fidelity.

[0157] ROI scalability may be defined as a type of scalability wherein an enhancement layer enhances only part of a reference-layer picture e.g. spatially, quality-wise, in bit-

depth, and/or along other scalability dimensions. As ROI scalability may be used together with other types of scalabilities, it may be considered to form a different categorization of scalability types. There exists several different applications for ROI coding with different requirements, which may be realized by using ROI scalability. For example, an
5 enhancement layer can be transmitted to enhance the quality and/or a resolution of a region in the base layer. A decoder receiving both enhancement and base layer bitstream might decode both layers and overlay the decoded pictures on top of each other and display the final picture.

[0158] One branch of research for obtaining compression improvement in stereoscopic
10 video is known as asymmetric stereoscopic video coding. Asymmetric stereoscopic video coding is based a theory that the Human Visual System (HVS) fuses the stereoscopic image pair such that the perceived quality is close to that of the higher quality view. Thus, compression improvement is obtained by providing a quality difference between the two coded views. In mixed-resolution (MR) stereoscopic video coding, also referred to as
15 resolution-asymmetric stereoscopic video coding, one of the views has lower spatial resolution and/or has been low-pass filtered compared to the other view.

[0159] In signal processing, resampling of images is usually understood as changing the sampling rate of the current image in horizontal or/and vertical directions. Resampling results in a new image which is represented with different number of pixels in horizontal
20 or/and vertical direction. In some applications, the process of image resampling is equal to image resizing. In general, resampling is classified in two processes: downsampling and upsampling.

[0160] Downsampling or subsampling process may be defined as reducing the sampling rate of a signal, and it typically results in reducing of the image sizes in horizontal and/or
25 vertical directions. In image downsampling, the spatial resolution of the output image, i.e. the number of pixels in the output image, is reduced compared to the spatial resolution of the input image. Downsampling ratio may be defined as the horizontal or vertical resolution of the downsampled image divided by the respective resolution of the input image for downsampling. Downsampling ratio may alternatively be defined as the number
30 of samples in the downsampled image divided by the number of samples in the input image for downsampling. As the two definitions differ, the term downsampling ratio may further be characterized by indicating whether it is indicated along one coordinate axis or both coordinate axes (and hence as a ratio of number of pixels in the images). Image downsampling may be performed for example by decimation, i.e. by selecting a specific

number of pixels, based on the downsampling ratio, out of the total number of pixels in the original image. In some embodiments downsampling may include low-pass filtering or other filtering operations, which may be performed before or after image decimation. Any low-pass filtering method may be used, including but not limited to linear averaging.

5 [0161] Upsampling process may be defined as increasing the sampling rate of the signal, and it typically results in increasing of the image sizes in horizontal and/or vertical directions. In image upsampling, the spatial resolution of the output image, i.e. the number of pixels in the output image, is increased compared to the spatial resolution of the input image. Upsampling ratio may be defined as the horizontal or vertical resolution of the
10 upsampled image divided by the respective resolution of the input image. Upsampling ratio may alternatively be defined as the number of samples in the upsampled image divided by the number of samples in the input image. As the two definitions differ, the term upsampling ratio may further be characterized by indicating whether it is indicated along one coordinate axis or both coordinate axes (and hence as a ratio of number of
15 pixels in the images). Image upsampling may be performed for example by copying or interpolating pixel values such that the total number of pixels is increased. In some embodiments, upsampling may include filtering operations, such as edge enhancement filtering.

[0162] Frame packing may be defined to comprise arranging more than one input picture, which may be referred to as (input) constituent frames, into an output picture. In general, frame packing is not limited to any particular type of constituent frames or the constituent frames need not have a particular relation with each other. In many cases, frame packing is used for arranging constituent frames of a stereoscopic video clip into a single picture sequence, as explained in more details in the next paragraph. The arranging may include
20 placing the input pictures in spatially non-overlapping areas within the output picture. For example, in a side-by-side arrangement, two input pictures are placed within an output picture horizontally adjacently to each other. The arranging may also include partitioning of one or more input pictures into two or more constituent frame partitions and placing the constituent frame partitions in spatially non-overlapping areas within the output picture.
25 The output picture or a sequence of frame-packed output pictures may be encoded into a bitstream e.g. by a video encoder. The bitstream may be decoded e.g. by a video decoder. The decoder or a post-processing operation after decoding may extract the decoded constituent frames from the decoded picture(s) e.g. for displaying.
30

[0163] In frame-compatible stereoscopic video (a.k.a. frame packing of stereoscopic video), a spatial packing of a stereo pair into a single frame is performed at the encoder side as a pre-processing step for encoding and then the frame-packed frames are encoded with a conventional 2D video coding scheme. The output frames produced by the decoder
5 contain constituent frames of a stereo pair.

[0164] In a typical operation mode, the spatial resolution of the original frames of each view and the packaged single frame have the same resolution. In this case the encoder downsamples the two views of the stereoscopic video before the packing operation. The spatial packing may use for example a side-by-side or top-bottom format, and the
10 downsampling should be performed accordingly.

[0165] Frame packing may be preferred over multiview video coding (e.g. MVC extension of H.264/AVC or MV-HEVC extension of H.265/HEVC) for example due to the following reasons:

[0166] The post-production workflows might be tailored for a single video signal. Some post-production tools might not be able to handle two separate picture sequences and/or might not be able to keep the separate picture sequences in synchrony with each other.
15

[0167] The distribution system, such as transmission protocols, might be such that support single coded sequence only and/or might not be able to keep separate coded sequences in synchrony with each other and/or may require more buffering or latency to keep the
20 separate coded sequences in synchrony with each other.

[0168] The decoding of bitstreams with multiview video coding tools may require support of specific coding modes, which might not be available in players. For example, many smartphones support H.265/HEVC Main profile decoding but are not able to handle H.265/HEVC Multiview Main profile decoding even though it only requires high-level
25 additions compared to the Main profile.

[0169] Frame packing may be inferior to multiview video coding in terms of compression performance (a.k.a. rate-distortion performance) due to, for example, the following reasons. In frame packing, inter-view sample prediction and inter-view motion prediction are not enabled between the views. Furthermore, in frame packing, motion vectors
30 pointing outside the boundaries of the constituent frame (to another constituent frame) or causing sub-pixel interpolation using samples outside the boundaries of the constituent frame (within another constituent frame) may be sub-optimally handled. In conventional multiview video coding, the sample locations used in inter prediction and sub-pixel interpolation may be saturated to be within the picture boundaries or equivalently areas

outside the picture boundary in the reconstructed pictures may be padded with border sample values.

[0170] Capturing process of 360-degree panoramic video may include camera rotation.

This camera rotation causes change in the position and scale of the objects in each picture compared to the previous pictures and hence may make the motion compensation inefficient in the compression.

[0171] Small amounts of rotation may be caused by shaking and other small movements when the content is shot with a handheld camera. Intentional rotation may be used in 360-degree video for example to keep a moving region-of-interest (ROI) in the center point of viewing (e.g. in the middle of an equirectangular panorama picture). In content occupying less than 360-degree field-of-view, rotation may be used similarly to keep moving regions-of-interest within the picture area. The camera rotation may be virtual, i.e. a director may choose the rotation at a post-production stage.

[0172] Figures 3a—3c show a rectangular grid 241 of an Equirectangular panoramic image and the corresponding resulted camera rotation effect. The camera rotation in this example is 1 degree in Figure 3b and 5 degrees in Figure 3c along x, y and z axis. The unprocessed reference frame has the regular grid as show in Figure 3a. If the camera is rotated in the current frame with respect to the reference frame (e.g. 1 or 5 degree), the unprocessed reference frame should be rotated accordingly which results in, for example, one of processed reference frames illustrated in Figures 3b and 3c.

[0173] The examples demonstrate that block-based translational motion compensation is likely to fail when camera rotation takes place. The examples demonstrate that even small amounts of rotation, which could e.g. be caused by unintentional movements of a handheld camera, may cause severe transformations in the image. In other words, if a frame to be motion predicted (a current frame) and the reference frame do not have the same capturing position e.g. due to the movement of the camera between capturing moments of the current frame and the reference frame, pixels in the current frame and co-located pixels in the unprocessed reference frame do not necessarily represent the same location in the captured scene. Thus, a motion vector might point to an incorrect location in the reference frame if no deformation between the reference frame and the current frame were made before determining motion vector candidate(s).

[0174] Camera orientation may characterize the orientation of a camera device or a camera rig relative to a coordinate system. Camera orientation may for example be indicated by

rotation angles, sometimes e.g. referred to as yaw, pitch and roll, around orthogonal coordinate axes.

[0175] The optional reference picture resampling by H.263 Annex P may be used to resample a temporal reference picture by indicating a displacement for each corner of the reference picture, as illustrated in Figure 3d. Bilinear interpolation is used for deriving the resampled sample values. This coding mode may be used for compensation of global motion. However, the warping enabled by H.263 Annex P may not be capable of modeling the transformations in 360-degree video that are caused by camera rotation.

[0176] An elastic motion model uses 2-D discrete cosine basis functions to represent a motion field. A reference frame may be generated by applying elastic motion model to a decoded frame. The generated reference frame is then used as a reference for prediction in a conventional manner. A similar approach could be used with other sophisticated motion models, such as the affine motion model.

[0177] While sophisticated motion models are more capable than the method of H.263 Annex P to reproduce different types of geometric transformations, they may not be able to capture the exact transformation caused by camera rotation to 360-degree video.

[0178] In the following, an example of manipulating/resampling the reference frames based on the camera orientation of the frame to be encoded for 360-degree video encoding will be explained with reference to Figure 6, in accordance with an embodiment. A decoded picture 611 (or equivalently a reconstructed picture in an encoder) is back-projected 612 onto a sphere. Back-projecting may alternatively be called mapping or projecting. Back-projecting may comprise projecting onto a first projection structure as an intermediate step. For example, if the decoded picture 611 is an equirectangular panorama picture, the decoded picture may first be mapped onto a cylinder and from the cylinder mapped onto a sphere. The orientation of the first projection structure 613 may be selected based on camera orientation when the decoded picture was captured, or alternatively the first projection structure may have a default orientation. A spherical image may for example be represented by a set of samples, each having spherical coordinates, such as yaw and pitch, and a sample value. In an example, a yaw value and a pitch value are directly proportional to the x and y coordinate, respectively, of a sample in a decoded equirectangular panorama picture.

[0179] The spherical image is then mapped 614 onto a second projection structure 615. If the first projection structure 613 has an orientation according to the camera orientation when the decoded picture was captured, the second projection structure may have an

orientation matching that of the camera orientation of the picture being encoded or decoded. If the first projection structure has a default orientation, the second projection structure may have an orientation matching the difference of the camera orientations for current picture being encoded or decoded and the decoded picture. Camera orientation
5 may be acquired directly from the camera (e.g. using a gyroscope and/or an accelerometer built in or attached to the camera) or can be estimated based on the reference frames or it may be retrieved from a bitstream or information about the camera orientation may have been attached with the frames.

[0180] When the equirectangular panorama format is used, the projection structure is a
10 cylinder. However, the invention is not limited to the equirectangular projection or the usage of a cylinder as the projection structure. For example cube map projection and a cube as a projection structure could be used instead.

[0181] The second projection structure 615 is then unfolded 616 to form a two-
dimensional image 617 that can be used as a reference picture for the picture being
15 encoded or decoded. The projected reference picture may be temporarily stored into a memory so that the motion prediction may utilize the projected reference picture. The unmodified reference picture may also be stored into the frame memory for example as long as that reference picture will be used as a reference. It should be noted that when the same reference picture is used as a reference for more than one picture to be
20 encoded/decoded, different projections may be needed for different pictures to be encoded/decoded, if they have different camera positions when the pictures have been captured.

[0182] Two or more of the above-described stages may be merged into a single process.
For example, forming of the spherical image may be omitted and back-projection directly
25 to a rotated projection structure may be applied.

[0183] It may not be necessary to transmit information of the geometry of the mapping for each picture but it may be sufficient to send information of the geometry once for each bitstream or coded video sequence or some other entity in which the geometry remains unchanged, or a fixed format may be used which is known by the encoder and the
30 decoder, wherein information of the geometry may not be transmitted at all.

[0184] In accordance with an embodiment, the rotation information may be transmitted for each picture so that the rotation information indicates the (absolute) rotation of the picture with a reference rotation (e.g. 0 degrees in each of the x, y and z direction). The difference between the rotation of a reference picture and the rotation of a current picture may be

obtained for example by subtracting the respective rotation angles in a particular order or by performing a reverse projection of the first angle, followed by the (forward) projection of the second angle.

[0185] The video encoding method according to an example embodiment will now be described with reference to the simplified block diagram of Figure 5a and the flow diagram of Figure 10a. The elements of Figure 5a may, for example, be implemented by the first encoder section 500 of the encoder of Figure 4a, or they may be separate from the first encoder section 500.

[0186] An uncompressed picture 221 (U0) is encoded 222 first as an intra-coded picture. A conventional intra picture encoding process can be used. The reconstructed picture 223 is then stored 224 in the decoded picture buffer (DPB) to be used as a reference in inter prediction.

[0187] For encoding the inter frames 225 (uncompressed picture Un , $n>0$, where n indicates the ((de)coding order of pictures), rotation information of a current frame to be encoded and one or more reference frames are examined (block 1002 in Figure 10a) to find out whether there is a difference in the rotation of the current frame and the one or more reference frames. If so, the one or more of the reference frames are rotated 227 and resampled 1003 based on the camera rotation parameters, as described earlier, to form manipulated reference pictures (frames) 228 so that the rotation of the manipulated reference pictures 228 correspond with the rotation of the current frame 225. The manipulated reference picture(s) 228 may be stored 1004 to a memory for the inter picture encoding process 229. The camera rotation parameters for each picture can be acquired 1001 directly from the camera or can be estimated from the previous pictures during the encoding or in a preprocessing step prior to encoding (block 226 in Figure 5a). Then the current frame is encoded 229, 1005 using the rotated reference frames. Original reference frames may additionally be used in the encoding 229 of the current frame. The encoding process may also perform decoding 1006 to form reconstructed picture for the current picture and possibly to be used as a reference picture for some subsequent picture(s). The reconstructed picture 230 (Rn , $n>0$) may be stored 1007 in the decoded picture buffer 224 (DPB).

[0188] The camera rotation information (for example, yaw, pitch and roll) for each picture can be transmitted to the decoder by encoding them into the bitstream 231.

[0189] The video decoding method according to the invention may be described with reference to the simplified block diagram of Figure 5b and the flow diagram of Figure

10b. The elements of Figure 5b may, for example, be implemented in the first decoder section 552 of the decoder of Figure 4b, or they may be separate from the first decoder section 552.

[0190] As an input, a bitstream 231 comprising coded pictures is obtained 1020. When a
5 coded picture is an intra-coded picture, intra picture decoding process 232 may be used, resulting into a reconstructed picture 233 which is stored in the decoded picture buffer 234.

[0191] When a coded picture is an inter-coded picture, the decoder may apply reference
10 picture rotation/resampling operation 235 to the reference picture(s) of the current decoded picture. For that, rotation information of the current picture and reference frames may be obtained 1021, for example, from the bitstream 231 or from some other appropriate source. The reference picture rotation/resampling operation 235 may examine
1022 rotation information of the current frame and the reference frame(s) to find out whether there is a difference in the rotation of the current frame and the reference
15 frame(s). If so, the reference frame(s) is/are rotated and resampled 1023 to form manipulated reference pictures (frames) 236 so that the rotation of the manipulated reference pictures 236 correspond with the rotation of the current frame. The manipulated reference pictures 236 may be stored 1024 to a memory for an inter picture decoding process.

[0192] The inter picture decoding process 237, 1025 may be used where at least one
20 reference picture that is or may be used as a reference for prediction is the picture R0. The decoding may result into a reconstructed picture 238 (Rn), which may be included 1026 in the decoded picture buffer 234.

[0193] Another embodiment for encoding utilizing an out-of-the-loop approach is
25 illustrated with reference to Figure 8a. Images are input 811 for encoding and changes of the camera orientation 812 are pre-compensated in the stitching and projection step 813 in which a projected frame 814 is formed. In other words, the orientation of the coordinate system and/or the projection structure used in stitching is kept unchanged through a video sequence, regardless of the camera orientation. The projected frame may then be
30 introduced to region-wise mapping 815 to form packed frames 816. The packed frames may then be encoded 817 and included 818 in a bitstream 819.

[0194] The camera orientation may be included in the encoded bitstream in the bitstream multiplexing stage 818. The bitstream multiplexing 818 may be regarded as part of encoding or may be regarded as a separate stage.

[0195] Another embodiment for encoding is illustrated with reference to Figure 8b. In this embodiment, the input 821 to the process is a sequence of projected frames. Rotation compensation 820 is applied to the projected frames, resulting into projected frames 814 (from projection structures of different orientations than those used originally in stitching and projection). The rotation compensation 820 may be implemented e.g. in the same way than what was explained in connection with Figure 6 above. Otherwise, this embodiment is similar to the embodiment of Figure 8a explained above.

[0196] In accordance with yet another embodiment, a fixed rotation angle (e.g. 0 degrees) may be assumed as follows. For example, there are several captured frames which may have different rotation angles. Hence, each frame having rotation angle different from the fixed rotation angle, may be rotated so that the rotation angle becomes the fixed rotation angle. After that, motion prediction may be performed in a straightforward manner as described above with Figure 8a or Figure 8b assuming that the rotation angle of each image corresponds with the fixed rotation angle. In order to enable decoders to reconstruct camera orientation, the fixed rotation angle as well as the camera orientations for captured frames may be included in the encoded bitstream in the bitstream multiplexing stage 818.

[0197] An embodiment for decoding is illustrated with reference to Figure 9. A bitstream is input 911 to the decoder. The bitstream may comprise encoded projected frames and/or encoded packed VR frames. In the bitstream demultiplexing stage 912, the camera orientation 913 is extracted from the bitstream. The bitstream demultiplexing 912 may be regarded as part of decoding or may be regarded as a separate stage. The bitstream demultiplexing stage 912 also extracts image information from the bitstream and provides it to a decoding stage 914. The output of the decoding stage 914 comprises packed VR frames 915; however, in case region-wise packing had not been applied in the encoding side, the output of the decoding stage may be considered to comprise projected frames. If the output of the decoding stage comprises packed VR frames, region-wise back-mapping 916 may be performed for the packed VR frames to form projected frames. If the packed frames already correspond with projected frames, the region-wise back-mapping 916 need not be performed. The projected frames 917 may be provided to rotation compensation 918 to produce decoded images 919 for rendering on a display, storing to a memory (e.g. to a decoded picture buffer and/or to a reference frame memory), retransmitting further, and/or for some other purposes.

[0198] Region-wise back-mapping may be specified or implemented as a process that maps regions of a packed VR frame to a projected frame. Metadata may be included in or

along the bitstream that describes the region-wise mapping from a projected frame to a packed VR frame. For example, a mapping of a source rectangle of a projected frame to a destination rectangle in a packed VR frame may be included in such metadata. The width and height of the source rectangle in relation to the width and height of the destination rectangle, respectively, may indicate a horizontal and vertical resampling ratio, respectively. A back-mapping process maps samples of the destination rectangle (as indicated in the metadata) of the packed VR frame to the source rectangle (as indicated in the metadata) of an output projected frame. The back-mapping process may include resampling according to the width and height ratios of the source and destination rectangles.

[0199] In an example, an encoder or any other entity includes back-mapping metadata into or along a bitstream in addition to or instead of mapping metadata. Back-mapping metadata may be indicative of the process to apply to the packed VR frame, e.g. resulting from the decoding stage 914, to achieve an output projected frame (e.g. 917). Back-mapping metadata may for example comprise source and destination rectangles, as described above, and rotation and mirroring to be applied to a region of a packed VR frame to obtain a region in the output projected frame.

[0200] The rotation compensation may be considered to be a part of the decoding process, e.g. similarly to cropping according to a conformance cropping window in HEVC.

Alternatively, the rotation compensation may be considered as a step outside the decoder.

[0201] The rotation compensation may be combined with subsequent steps in the processing pipeline, such as YUV to RGB conversion and rendering onto a display viewport.

[0202] The embodiments are not limited to any particular coordinate system. The paragraphs below describe some examples of coordinate systems that can be used.

[0203] Figure 7a specifies the coordinate axes used for defining yaw, pitch, and roll angles. Yaw is applied prior to pitch, and pitch is applied prior to roll. Yaw rotates around the Y (vertical, up) axis, pitch around the X (lateral, side-to-side) axis, and roll around the Z (back-to-front) axis. Rotations are extrinsic, i.e., around the X, Y, and Z fixed reference axes. The angles increase counter-clockwise when looking towards the origin.

[0204] Another coordinate system is illustrated in Figure 7b, which represents the rotation on a 3D space along each axis. The camera is located in the center i.e., (0, 0, 0) location, and its rotation can be along at least one axis. The rotation along Y, X and Z axes are defined as Yaw, Roll, and Pitch, respectively.

[0205] In the presented coordinate systems or any similar coordinate system, yaw, pitch, and roll may be indicated e.g. in degrees as floating point decimal values. Value ranges may be defined for yaw, pitch, and roll. For example, yaw may be required to be in the range of 0, inclusive, to 360, exclusive; pitch may be required to be in the range of -90 to 90, inclusive; and roll may be required to be in the range of 0, inclusive, to 360, exclusive.

[0206] According to an embodiment, a decoded motion field (or equivalently a reconstructed motion field in an encoder) is back-projected onto a sphere, e.g. based on associated block coordinates for each set of motion information. Back-projecting may comprise projecting onto a first projection structure as an intermediate step. For example, if a motion field is for an equirectangular panorama picture, the motion field may first be mapped onto a cylinder and from the cylinder mapped onto a sphere. The orientation of the first projection structure may be selected based on camera orientation when the decoded picture corresponding to the motion field was captured, or alternatively the first projection structure may have a default orientation. The spherically mapped motion field image is then mapped onto a second projection structure. If the first projection structure has an orientation according to the camera orientation when the decoded picture was captured, the second projection structure may have an orientation matching that of the camera orientation of the picture being encoded or decoded. If the first projection structure has a default orientation, the second projection structure may have an orientation matching the difference of the camera orientations for current picture being encoded or decoded and the decoded picture. Camera orientation may be acquired directly from the camera (e.g. using a gyroscope and/or an accelerometer built in or attached to the camera) or can be estimated based on the reference frames or it may be retrieved from a bitstream or information about the camera orientation may have been attached with the frames. The motion field mapped onto the second projection structure is then mapped onto a reference motion field of a two-dimensional image, essentially by unfolding the second projection structure onto the two-dimensional image. Decimation or resampling may be a part of said mapping. For example, if two or more sets of motion information are mapped onto the same block of the reference motion field, one of them may be selected, e.g. on the basis which set is mapped closer to a reference point (e.g. mid-most sample) of the block, or motion information may be averaged or interpolated particularly if same reference picture(s) are used in those sets of motion information that are mapped to the same block of the reference motion field. The reference motion field is or may be used as a reference

for TMVP of HEVC or a similar process that uses a motion field of a reference picture as a source for motion information prediction of a current picture.

[0207] Motion vector prediction of H.265/HEVC is described below as an example of a system or method where embodiments may be applied.

5 [0208] H.265/HEVC includes two motion vector prediction schemes, namely the advanced motion vector prediction (AMVP) and the merge mode. In the AMVP or the merge mode, a list of motion vector candidates is derived for a PU. There are two kinds of candidates: spatial candidates and temporal candidates, where temporal candidates may also be referred to as TMVP candidates. The sources of the candidate motion vector predictors are presented in Figures 11a and 11b. X stands for the current prediction unit. A0, A1, B0, B1, B2 in Figure 11a are spatial candidates while C0, C1 in Figure 11b are temporal candidates. The block comprising or corresponding to the candidate C0 or C1 in Figure 11b, whichever is the source for the temporal candidate, may be referred to as the collocated block.

15 [0209] A candidate list derivation may be performed for example as follows, while it should be understood that other possibilities may exist for candidate list derivation. If the occupancy of the candidate list is not at maximum, the spatial candidates are included in the candidate list first if they are available and not already exist in the candidate list. After that, if occupancy of the candidate list is not yet at maximum, a temporal candidate is included in the candidate list. If the number of candidates still does not reach the maximum allowed number, the combined bi-predictive candidates (for B slices) and a zero motion vector are added in. After the candidate list has been constructed, the encoder decides the final motion information from candidates for example based on a rate-distortion optimization (RDO) decision and encodes the index of the selected candidate into the bitstream. Likewise, the decoder decodes the index of the selected candidate from the bitstream, constructs the candidate list, and uses the decoded index to select a motion vector predictor from the candidate list.

25 [0210] In H.265/HEVC, AMVP and the merge mode may be characterized as follows. In AMVP, the encoder indicates whether uni-prediction or bi-prediction is used and which reference pictures are used as well as encodes a motion vector difference. In the merge mode, only the chosen candidate from the candidate list is encoded into the bitstream indicating the current prediction unit has the same motion information as that of the indicated predictor. Thus, the merge mode creates regions composed of neighbouring prediction blocks sharing identical motion information, which is only signalled once for

each region. Another difference between AMVP and the merge mode in H.265/HEVC is that the maximum number of candidates of AMVP is 2 while that of the merge mode is 5.

[0211] The advanced motion vector prediction may operate for example as follows, while other similar realizations of advanced motion vector prediction are also possible for example with different candidate position sets and candidate locations with candidate position sets. Two spatial motion vector predictors (MVPs) may be derived and a temporal motion vector predictor (TMVP) may be derived. They may be selected among the positions: three spatial motion vector predictor candidate positions located above the current prediction block (B0, B1, B2) and two on the left (A0, A1). The first motion vector predictor that is available (e.g. resides in the same slice, is inter-coded, etc.) in a pre-defined order of each candidate position set, (B0, B1, B2) or (A0, A1), may be selected to represent that prediction direction (up or left) in the motion vector competition. A reference index for the temporal motion vector predictor may be indicated by the encoder in the slice header (e.g. as a `collocated_ref_idx` syntax element). The first motion vector predictor that is available (e.g. is inter-coded) in a pre-defined order of potential temporal candidate locations, e.g. in the order (C0, C1), may be selected as a source for a temporal motion vector predictor. The motion vector obtained from the first available candidate location in the co-located picture may be scaled according to the proportions of the picture order count differences of the reference picture of the temporal motion vector predictor, the co-located picture, and the current picture. Moreover, a redundancy check may be performed among the candidates to remove identical candidates, which can lead to the inclusion of a zero motion vector in the candidate list. The motion vector predictor may be indicated in the bitstream for example by indicating the direction of the spatial motion vector predictor (up or left) or the selection of the temporal motion vector predictor candidate. The co-located picture may also be referred to as the collocated picture, the source for motion vector prediction, or the source picture for motion vector prediction.

[0212] The merging/merge mode/process/mechanism may operate for example as follows, while other similar realizations of the merge mode are also possible for example with different candidate position sets and candidate locations with candidate position sets.

[0213] In the merging/merge mode/process/mechanism, where all the motion information of a block/PU is predicted and used without any modification/correction. The aforementioned motion information for a PU may comprise one or more of the following:

- 1) The information whether 'the PU is uni-predicted using only reference picture list0' or

'the PU is uni-predicted using only reference picture list1' or 'the PU is bi-predicted using both reference picture list0 and list1'; 2) Motion vector value corresponding to the reference picture list0, which may comprise a horizontal and vertical motion vector component; 3) Reference picture index in the reference picture list0 and/or an identifier of a reference picture pointed to by the Motion vector corresponding to reference picture list 0, where the identifier of a reference picture may be for example a picture order count value, a layer identifier value (for inter-layer prediction), or a pair of a picture order count value and a layer identifier value; 4) Information of the reference picture marking of the reference picture, e.g. information whether the reference picture was marked as "used for short-term reference" or "used for long-term reference"; 5) - 7) The same as 2) - 4), respectively, but for reference picture list1.

[0214] Similarly, predicting the motion information is carried out using the motion information of adjacent blocks and/or co-located blocks in temporal reference pictures. A list, often called as a merge list, may be constructed by including motion prediction candidates associated with available adjacent/co-located blocks and the index of selected motion prediction candidate in the list is signalled and the motion information of the selected candidate is copied to the motion information of the current PU. When the merge mechanism is employed for a whole CU and the prediction signal for the CU is used as the reconstruction signal, i.e. prediction residual is not processed, this type of coding/decoding the CU is typically named as skip mode or merge based skip mode. In addition to the skip mode, the merge mechanism may also be employed for individual PUs (not necessarily the whole CU as in skip mode) and in this case, prediction residual may be utilized to improve prediction quality. This type of prediction mode is typically named as an inter-merge mode.

[0215] One of the candidates in the merge list and/or the candidate list for AMVP or any similar motion vector candidate list may be a TMVP candidate or alike, which may be derived from the collocated block within an indicated or inferred reference picture, such as the reference picture indicated for example in the slice header. In HEVC, the reference picture list to be used for obtaining a collocated partition is chosen according to the `collocated_from_l0_flag` syntax element in the slice header. When the flag is equal to 1, it specifies that the picture that contains the collocated partition is derived from list 0, otherwise the picture is derived from list 1. When `collocated_from_l0_flag` is not present, it is inferred to be equal to 1. The `collocated_ref_idx` in the slice header specifies the reference index of the picture that contains the collocated partition. When the current slice

is a P slice, `collocated_ref_idx` refers to a picture in list 0. When the current slice is a B slice, `collocated_ref_idx` refers to a picture in list 0 if `collocated_from_l0` is 1, otherwise it refers to a picture in list 1. `collocated_ref_idx` always refers to a valid list entry, and the resulting picture is the same for all slices of a coded picture. When `collocated_ref_idx` is not present, it is inferred to be equal to 0.

[0216] In HEVC the so-called target reference index for temporal motion vector prediction in the merge list is set as 0 when the motion coding mode is the merge mode. When the motion coding mode in HEVC utilizing the temporal motion vector prediction is the advanced motion vector prediction mode, the target reference index values are explicitly indicated (e.g. per each PU).

[0217] In HEVC, the availability of a candidate predicted motion vector (PMV) may be determined as follows (both for spatial and temporal candidates) (SRTP = short-term reference picture, LRTP = long-term reference picture):

reference picture for target reference index	reference picture for candidate PMV	candidate PMV availability
STRP	STRP	“available” (and scaled)
STRP	LRTP	“unavailable”
LRTP	STRP	“unavailable”
LRTP	LRTP	“available” (but not scaled)

[0218] In HEVC, when the target reference index value has been determined, the motion vector value of the temporal motion vector prediction may be derived as follows: The motion vector PMV at the block that is collocated with the bottom-right neighbor (location C0 in Figure 11b) of the current prediction unit is obtained. The picture where the collocated block resides may be e.g. determined according to the signalled reference index in the slice header as described above. If the PMV at location C0 is not available, the motion vector PMV at location C1 (see Figure 11b) of the collocated picture is obtained. The determined available motion vector PMV at the co-located block is scaled with respect to the ratio of a first picture order count difference and a second picture order count difference. The first picture order count difference is derived between the picture containing the co-located block and the reference picture of the motion vector of the co-located block. The second picture order count difference is derived between the current picture and the target reference picture. If one but not both of the target reference picture

and the reference picture of the motion vector of the collocated block is a long-term reference picture (while the other is a short-term reference picture), the TMVP candidate may be considered unavailable. If both of the target reference picture and the reference picture of the motion vector of the collocated block are long-term reference pictures, no POC-based motion vector scaling may be applied.

[0219] Motion parameter types or motion information may include but are not limited to one or more of the following types:

- an indication of a prediction type (e.g. intra prediction, uni-prediction, bi-prediction) and/or a number of reference pictures;
- 10 — an indication of a prediction direction, such as inter (a.k.a. temporal) prediction, inter-layer prediction, inter-view prediction, view synthesis prediction (VSP), and inter-component prediction (which may be indicated per reference picture and/or per prediction type and where in some embodiments inter-view and view-synthesis prediction may be jointly considered as one prediction direction) and/or
- 15 — an indication of a reference picture type, such as a short-term reference picture and/or a long-term reference picture and/or an inter-layer reference picture (which may be indicated e.g. per reference picture)
- a reference index to a reference picture list and/or any other identifier of a reference picture (which may be indicated e.g. per reference picture and the type of which may depend on the prediction direction and/or the reference picture type and which may be
- 20 accompanied by other relevant pieces of information, such as the reference picture list or alike to which reference index applies);
- a horizontal motion vector component (which may be indicated e.g. per prediction block or per reference index or alike);
- 25 — a vertical motion vector component (which may be indicated e.g. per prediction block or per reference index or alike);
- one or more parameters, such as picture order count difference and/or a relative camera separation between the picture containing or associated with the motion parameters and its reference picture, which may be used for scaling of the horizontal motion vector
- 30 component and/or the vertical motion vector component in one or more motion vector prediction processes (where said one or more parameters may be indicated e.g. per each reference picture or each reference index or alike);
- coordinates of a block to which the motion parameters and/or motion information applies, e.g. coordinates of the top-left sample of the block in luma sample units;

- extents (e.g. a width and a height) of a block to which the motion parameters and/or motion information applies.

[0220] In general, motion vector prediction mechanisms, such as those motion vector prediction mechanisms presented above as examples, may include prediction or inheritance of certain pre-defined or indicated motion parameters.

[0221] A motion field associated with a picture may be considered to comprise of a set of motion information produced for every coded block of the picture. A motion field may be accessible by coordinates of a block, for example. A set of motion information associated with a block may for example correspond to the top-left or midmost sample location of the block. A motion field may be used for example in TMVP or any other motion prediction mechanism where a source or a reference for prediction other than the current (de)coded picture is used.

[0222] Figure 12 is a graphical representation of an example multimedia communication system within which various embodiments may be implemented. A data source 1510 provides a source signal in an analog, uncompressed digital, or compressed digital format, or any combination of these formats. An encoder 1520 may include or be connected with a pre-processing, such as data format conversion and/or filtering of the source signal. The encoder 1520 encodes the source signal into a coded media bitstream. It should be noted that a bitstream to be decoded may be received directly or indirectly from a remote device located within virtually any type of network. Additionally, the bitstream may be received from local hardware or software. The encoder 1520 may be capable of encoding more than one media type, such as audio and video, or more than one encoder 1520 may be required to code different media types of the source signal. The encoder 1520 may also get synthetically produced input, such as graphics and text, or it may be capable of producing coded bitstreams of synthetic media. In the following, only processing of one coded media bitstream of one media type is considered to simplify the description. It should be noted, however, that typically real-time broadcast services comprise several streams (typically at least one audio, video and text sub-titling stream). It should also be noted that the system may include many encoders, but in the figure only one encoder 1520 is represented to simplify the description without a lack of generality. It should be further understood that, although text and examples contained herein may specifically describe an encoding process, one skilled in the art would understand that the same concepts and principles also apply to the corresponding decoding process and vice versa.

[0223] The coded media bitstream may be transferred to a storage 1530. The storage 1530 may comprise any type of mass memory to store the coded media bitstream. The format of the coded media bitstream in the storage 1530 may be an elementary self-contained bitstream format, or one or more coded media bitstreams may be encapsulated into a container file, or the coded media bitstream may be encapsulated into a Segment format suitable for DASH (or a similar streaming system) and stored as a sequence of Segments. If one or more media bitstreams are encapsulated in a container file, a file generator (not shown in the figure) may be used to store the one more media bitstreams in the file and create file format metadata, which may also be stored in the file. The encoder 1520 or the storage 1530 may comprise the file generator, or the file generator is operationally attached to either the encoder 1520 or the storage 1530. Some systems operate “live”, i.e. omit storage and transfer coded media bitstream from the encoder 1520 directly to the sender 1540. The coded media bitstream may then be transferred to the sender 1540, also referred to as the server, on a need basis. The format used in the transmission may be an elementary self-contained bitstream format, a packet stream format, a Segment format suitable for DASH (or a similar streaming system), or one or more coded media bitstreams may be encapsulated into a container file. The encoder 1520, the storage 1530, and the server 1540 may reside in the same physical device or they may be included in separate devices. The encoder 1520 and server 1540 may operate with live real-time content, in which case the coded media bitstream is typically not stored permanently, but rather buffered for small periods of time in the content encoder 1520 and/or in the server 1540 to smooth out variations in processing delay, transfer delay, and coded media bitrate.

[0224] The server 1540 sends the coded media bitstream using a communication protocol stack. The stack may include but is not limited to one or more of Real-Time Transport Protocol (RTP), User Datagram Protocol (UDP), Hypertext Transfer Protocol (HTTP), Transmission Control Protocol (TCP), and Internet Protocol (IP). When the communication protocol stack is packet-oriented, the server 1540 encapsulates the coded media bitstream into packets. For example, when RTP is used, the server 1540 encapsulates the coded media bitstream into RTP packets according to an RTP payload format. Typically, each media type has a dedicated RTP payload format. It should be again noted that a system may contain more than one server 1540, but for the sake of simplicity, the following description only considers one server 1540.

[0225] If the media content is encapsulated in a container file for the storage 1530 or for inputting the data to the sender 1540, the sender 1540 may comprise or be operationally

attached to a "sending file parser" (not shown in the figure). In particular, if the container file is not transmitted as such but at least one of the contained coded media bitstream is encapsulated for transport over a communication protocol, a sending file parser locates appropriate parts of the coded media bitstream to be conveyed over the communication
5 protocol. The sending file parser may also help in creating the correct format for the communication protocol, such as packet headers and payloads. The multimedia container file may contain encapsulation instructions, such as hint tracks in the ISOBMFF, for encapsulation of the at least one of the contained media bitstream on the communication protocol.

10 [0226] The server 1540 may or may not be connected to a gateway 1550 through a communication network, which may e.g. be a combination of a CDN, the Internet and/or one or more access networks. The gateway may also or alternatively be referred to as a middle-box. For DASH, the gateway may be an edge server (of a CDN) or a web proxy. It is noted that the system may generally comprise any number gateways or alike, but for the
15 sake of simplicity, the following description only considers one gateway 1550. The gateway 1550 may perform different types of functions, such as translation of a packet stream according to one communication protocol stack to another communication protocol stack, merging and forking of data streams, and manipulation of data stream according to the downlink and/or receiver capabilities, such as controlling the bit rate of the forwarded
20 stream according to prevailing downlink network conditions.

[0227] The system includes one or more receivers 1560, typically capable of receiving, demodulating, and de-capsulating the transmitted signal into a coded media bitstream. The coded media bitstream may be transferred to a recording storage 1570. The recording storage 1570 may comprise any type of mass memory to store the coded media bitstream.
25 The recording storage 1570 may alternatively or additively comprise computation memory, such as random access memory. The format of the coded media bitstream in the recording storage 1570 may be an elementary self-contained bitstream format, or one or more coded media bitstreams may be encapsulated into a container file. If there are multiple coded media bitstreams, such as an audio stream and a video stream, associated
30 with each other, a container file is typically used and the receiver 1560 comprises or is attached to a container file generator producing a container file from input streams. Some systems operate "live," i.e. omit the recording storage 1570 and transfer coded media bitstream from the receiver 1560 directly to the decoder 1580. In some systems, only the most recent part of the recorded stream, e.g., the most recent 10-minute excerpt of the

recorded stream, is maintained in the recording storage 1570, while any earlier recorded data is discarded from the recording storage 1570.

[0228] The coded media bitstream may be transferred from the recording storage 1570 to the decoder 1580. If there are many coded media bitstreams, such as an audio stream and a video stream, associated with each other and encapsulated into a container file or a single media bitstream is encapsulated in a container file e.g. for easier access, a file parser (not shown in the figure) is used to decapsulate each coded media bitstream from the container file. The recording storage 1570 or a decoder 1580 may comprise the file parser, or the file parser is attached to either recording storage 1570 or the decoder 1580. It should also be noted that the system may include many decoders, but here only one decoder 1570 is discussed to simplify the description without a lack of generality

[0229] The coded media bitstream may be processed further by a decoder 1570, whose output is one or more uncompressed media streams. Finally, a renderer 1590 may reproduce the uncompressed media streams with a loudspeaker or a display, for example. The receiver 1560, recording storage 1570, decoder 1570, and renderer 1590 may reside in the same physical device or they may be included in separate devices.

[0230] A sender 1540 and/or a gateway 1550 may be configured to perform switching between different representations e.g. for view switching, bitrate adaptation and/or fast start-up, and/or a sender 1540 and/or a gateway 1550 may be configured to select the transmitted representation(s). Switching between different representations may take place for multiple reasons, such as to respond to requests of the receiver 1560 or prevailing conditions, such as throughput, of the network over which the bitstream is conveyed. A request from the receiver can be, e.g., a request for a Segment or a Subsegment from a different representation than earlier, a request for a change of transmitted scalability layers and/or sub-layers, or a change of a rendering device having different capabilities compared to the previous one. A request for a Segment may be an HTTP GET request. A request for a Subsegment may be an HTTP GET request with a byte range. Additionally or alternatively, bitrate adjustment or bitrate adaptation may be used for example for providing so-called fast start-up in streaming services, where the bitrate of the transmitted stream is lower than the channel bitrate after starting or random-accessing the streaming in order to start playback immediately and to achieve a buffer occupancy level that tolerates occasional packet delays and/or retransmissions. Bitrate adaptation may include multiple representation or layer up-switching and representation or layer down-switching operations taking place in various orders.

[0231] A decoder 1580 may be configured to perform switching between different representations e.g. for view switching, bitrate adaptation and/or fast start-up, and/or a decoder 1580 may be configured to select the transmitted representation(s). Switching between different representations may take place for multiple reasons, such as to achieve faster decoding operation or to adapt the transmitted bitstream, e.g. in terms of bitrate, to prevailing conditions, such as throughput, of the network over which the bitstream is conveyed. Faster decoding operation might be needed for example if the device including the decoder 580 is multi-tasking and uses computing resources for other purposes than decoding the scalable video bitstream. In another example, faster decoding operation might be needed when content is played back at a faster pace than the normal playback speed, e.g. twice or three times faster than conventional real-time playback rate. The speed of decoder operation may be changed during the decoding or playback for example as response to changing from a fast-forward play from normal playback rate or vice versa, and consequently multiple layer up-switching and layer down-switching operations may take place in various orders.

[0232] In the above, some embodiments have been described with reference to the term block. It needs to be understood that the term block may be interpreted in the context of the terminology used in a particular codec or coding format. For example, the term block may be interpreted as a prediction unit in HEVC. It needs to be understood that the term block may be interpreted differently based on the context it is used. For example, when the term block is used in the context of motion fields, it may be interpreted to match to the block grid of the motion field.

[0233] In the above, some embodiments have been described with reference to back-projecting on a sphere, e.g. in step 612 of Fig. 6. It needs to be understood that another projection structure than a sphere may be likewise used in the back-projection.

[0234] In the above, some embodiments have been described with reference to projected frames that may have resulted from stitching and projection of source frames. It needs to be understood that embodiments may be similarly realized with any non-rectilinear frames, such as fisheye frames, instead of projected frames. As an example, a fisheye frame may be back-projected onto a projection structure. E.g. if a fisheye frame covers 180 degrees in field of view, it may be mapped onto a projection structure that is a hemisphere.

[0235] The phrase along the bitstream (e.g. indicating along the bitstream) may be used in claims and described embodiments to refer to out-of-band transmission, signaling, or

storage in a manner that the out-of-band data is associated with the bitstream. The phrase decoding along the bitstream or alike may refer to decoding the referred out-of-band data (which may be obtained from out-of-band transmission, signaling, or storage) that is associated with the bitstream. In the above, some embodiments have been described with reference to encoding or including indications or metadata in the bitstream and/or decoding indications or metadata from the bitstream. It needs to be understood that indications or metadata may additionally or alternatively be encoded or included along the bitstream and/or decoded along the bitstream. For example, indications or metadata may be included in or decoded from a container file that encapsulates the bitstream.

[0236] Some embodiments have been described with reference to the phrase camera and/or the orientation of the camera and/or the camera rotation. It needs to be understood that the phrase camera equally applies to a camera rig or alike multi-device capturing system. It also needs to be understood that the camera may be virtual e.g. in computer-generated content, where the camera orientation or such may be obtained from the modeling parameters used in creating the computer-generated content.

[0237] The following describes in further detail suitable apparatus and possible mechanisms for implementing the embodiments of the invention. In this regard reference is first made to Figure 13 which shows a schematic block diagram of an exemplary apparatus or electronic device 50 depicted in Figure 14, which may incorporate a transmitter according to an embodiment of the invention.

[0238] The electronic device 50 may for example be a mobile terminal or user equipment of a wireless communication system. However, it would be appreciated that embodiments of the invention may be implemented within any electronic device or apparatus which may require transmission of radio frequency signals.

[0239] The apparatus 50 may comprise a housing 30 for incorporating and protecting the device. The apparatus 50 further may comprise a display 32 in the form of a liquid crystal display. In other embodiments of the invention the display may be any suitable display technology suitable to display an image or video. The apparatus 50 may further comprise a keypad 34. In other embodiments of the invention any suitable data or user interface mechanism may be employed. For example the user interface may be implemented as a virtual keyboard or data entry system as part of a touch-sensitive display. The apparatus may comprise a microphone 36 or any suitable audio input which may be a digital or analogue signal input. The apparatus 50 may further comprise an audio output device which in embodiments of the invention may be any one of: an earpiece 38, speaker, or an

analogue audio or digital audio output connection. The apparatus 50 may also comprise a battery 40 (or in other embodiments of the invention the device may be powered by any suitable mobile energy device such as solar cell, fuel cell or clockwork generator). The term battery discussed in connection with the embodiments may also be one of these mobile energy devices. Further, the apparatus 50 may comprise a combination of different kinds of energy devices, for example a rechargeable battery and a solar cell. The apparatus may further comprise an infrared port 41 for short range line of sight communication to other devices. In other embodiments the apparatus 50 may further comprise any suitable short range communication solution such as for example a Bluetooth wireless connection or a USB/FireWire wired connection.

[0240] The apparatus 50 may comprise a controller 56 or processor for controlling the apparatus 50. The controller 56 may be connected to memory 58 which in embodiments of the invention may store both data and/or may also store instructions for implementation on the controller 56. The controller 56 may further be connected to codec circuitry 54 suitable for carrying out coding and decoding of audio and/or video data or assisting in coding and decoding carried out by the controller 56.

[0241] The apparatus 50 may further comprise a card reader 48 and a smart card 46, for example a universal integrated circuit card (UICC) reader and a universal integrated circuit card for providing user information and being suitable for providing authentication information for authentication and authorization of the user at a network.

[0242] The apparatus 50 may comprise radio interface circuitry 52 connected to the controller and suitable for generating wireless communication signals for example for communication with a cellular communications network, a wireless communications system or a wireless local area network. The apparatus 50 may further comprise an antenna 60 connected to the radio interface circuitry 52 for transmitting radio frequency signals generated at the radio interface circuitry 52 to other apparatus(es) and for receiving radio frequency signals from other apparatus(es).

[0243] In some embodiments of the invention, the apparatus 50 comprises a camera 42 capable of recording or detecting imaging.

[0244] With respect to Figure 15, an example of a system within which embodiments of the present invention can be utilized is shown. The system 10 comprises multiple communication devices which can communicate through one or more networks. The system 10 may comprise any combination of wired and/or wireless networks including, but not limited to a wireless cellular telephone network (such as a global systems for

mobile communications (GSM), universal mobile telecommunications system (UMTS), long term evolution (LTE) based network, code division multiple access (CDMA) network etc.), a wireless local area network (WLAN) such as defined by any of the IEEE 802.x standards, a Bluetooth personal area network, an Ethernet local area network, a
5 token ring local area network, a wide area network, and the Internet.

[0245] For example, the system shown in Figure 15 shows a mobile telephone network 11 and a representation of the internet 28. Connectivity to the internet 28 may include, but is not limited to, long range wireless connections, short range wireless connections, and various wired connections including, but not limited to, telephone lines, cable lines, power
10 lines, and similar communication pathways.

[0246] The example communication devices shown in the system 10 may include, but are not limited to, an electronic device or apparatus 50, a combination of a personal digital assistant (PDA) and a mobile telephone 14, a PDA 16, an integrated messaging device (IMD) 18, a desktop computer 20, a notebook computer 22, a tablet computer. The
15 apparatus 50 may be stationary or mobile when carried by an individual who is moving. The apparatus 50 may also be located in a mode of transport including, but not limited to, a car, a truck, a taxi, a bus, a train, a boat, an airplane, a bicycle, a motorcycle or any similar suitable mode of transport.

[0247] Some or further apparatus may send and receive calls and messages and
20 communicate with service providers through a wireless connection 25 to a base station 24. The base station 24 may be connected to a network server 26 that allows communication between the mobile telephone network 11 and the internet 28. The system may include additional communication devices and communication devices of various types.

[0248] The communication devices may communicate using various transmission
25 technologies including, but not limited to, code division multiple access (CDMA), global systems for mobile communications (GSM), universal mobile telecommunications system (UMTS), time divisional multiple access (TDMA), frequency division multiple access (FDMA), transmission control protocol-internet protocol (TCP-IP), short messaging service (SMS), multimedia messaging service (MMS), email, instant messaging service
30 (IMS), Bluetooth, IEEE 802.11, Long Term Evolution wireless communication technique (LTE) and any similar wireless communication technology. A communications device involved in implementing various embodiments of the present invention may communicate using various media including, but not limited to, radio, infrared, laser, cable connections, and any suitable connection.

[0249] Although the above examples describe embodiments of the invention operating within a wireless communication device, it would be appreciated that the invention as described above may be implemented as a part of any apparatus comprising a circuitry in which radio frequency signals are transmitted and received. Thus, for example, 5
embodiments of the invention may be implemented in a mobile phone, in a base station, in a computer such as a desktop computer or a tablet computer comprising radio frequency communication means (e.g. wireless local area network, cellular radio, etc.).

[0250] In general, the various embodiments of the invention may be implemented in hardware or special purpose circuits or any combination thereof. While various aspects of 10
the invention may be illustrated and described as block diagrams or using some other pictorial representation, it is well understood that these blocks, apparatus, systems, techniques or methods described herein may be implemented in, as non-limiting examples, hardware, software, firmware, special purpose circuits or logic, general purpose hardware or controller or other computing devices, or some combination thereof.

15 [0251] Embodiments of the inventions may be practiced in various components such as integrated circuit modules. The design of integrated circuits is by and large a highly automated process. Complex and powerful software tools are available for converting a logic level design into a semiconductor circuit design ready to be etched and formed on a semiconductor substrate.

20 [0252] Programs, such as those provided by Synopsys, Inc. of Mountain View, California and Cadence Design, of San Jose, California automatically route conductors and locate components on a semiconductor chip using well established rules of design as well as libraries of pre stored design modules. Once the design for a semiconductor circuit has been completed, the resultant design, in a standardized electronic format (e.g., Opus, 25
GDSII, or the like) may be transmitted to a semiconductor fabrication facility or "fab" for fabrication.

[0253] The foregoing description has provided by way of exemplary and non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those 30
skilled in the relevant arts in view of the foregoing description, when read in conjunction with the accompanying drawings and the appended claims. However, all such and similar modifications of the teachings of this invention will still fall within the scope of this invention.

[0254] In the following some examples will be provided.

[0255] According to a first example, there is provided a method comprising:

interpreting a first reconstructed picture as a first three-dimensional picture in a coordinate system;

5 obtaining a rotation;

projecting the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

forming a first reference picture, said forming comprising unfolding the first
10 geometrical projection structure into a second geometrical projection structure;
predicting at least a block of a second reconstructed picture from the first reference picture.

[0256] In some embodiments the method further comprises performing two or more of said interpreting, projecting, and forming as a single process.

15 [0257] In some embodiments of the method the first and second reconstructed pictures comply with an equirectangular panorama representation format.

[0258] In some embodiments the method further comprises:

decoding a first coded picture into a first reconstructed picture; and

20 decoding a second coded picture into the second reconstructed picture; wherein the decoding comprises said predicting.

[0259] In some embodiments the method further comprises:

decoding one or more syntax elements indicative of the rotation.

[0260] In some embodiments the method further comprises:

25 encoding a first picture into a first coded picture, wherein the encoding comprises reconstructing the first reconstructed picture; and

encoding a second picture into a second coded picture; wherein the encoding comprises reconstructing the second reconstructed picture and said predicting.

[0261] In some embodiments the method further comprises:

30 obtaining a first orientation of the apparatus when capturing a first set of input images from which the first picture originates;

obtaining a second orientation of the apparatus when capturing a second set of input images from which the second picture originates;

deriving the rotation on the basis of the first orientation and the second orientation.

[0262] In some embodiments the method further comprises:

estimating the rotation based on the first picture and the second picture.

[0263] According to a second example there is provided an apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon,

5 which when executed by said at least one processor, causes the apparatus to perform at least:

interpret a first reconstructed picture as a first three-dimensional picture in a coordinate system;

obtain a rotation;

10 project the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

form a first reference picture, said forming comprising unfolding the first geometrical projection structure into a second geometrical projection structure;

15 predict at least a block of a second reconstructed picture from the first reference picture.

[0264] According to a third example there is provided a computer readable storage medium comprising code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:

20 interpret a first reconstructed picture as a first three-dimensional picture in a coordinate system;

obtain a rotation;

project the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

25 form a first reference picture, said forming comprising unfolding the first geometrical projection structure into a second geometrical projection structure;

predict at least a block of a second reconstructed picture from the first reference picture.

[0265] According to a fourth example there is provided an apparatus comprising:

30 means for interpreting a first reconstructed picture as a first three-dimensional picture in a coordinate system;

means for obtaining a rotation;

means for projecting the first three-dimensional picture onto a first geometrical projection structure, the geometrical projection structure having an orientation according to the rotation within the coordinate system;

5 means for forming a first reference picture, said forming comprising unfolding the first geometrical projection structure into a second geometrical projection structure;

means for predicting at least a block of a second reconstructed picture from the first reference picture.

[0266] According to a fifth example there is provided a method comprising:

obtaining images captured by a camera;

10 obtaining information of orientation of the camera;

using the orientation information to compensate the orientation of the camera in the image with reference to a coordinate system; and

forming a projected frame from the orientation compensated image by using a projection structure.

15 [0267] In some embodiments the method further comprises:

keeping an orientation of the coordinate system unchanged.

[0268] In some embodiments the method further comprises:

keeping the projection structure unchanged.

[0269] In some embodiments the method further comprises:

20 region-wise mapping the projected frame to form packed frames.

[0270] In some embodiments the method further comprises:

including information of orientation of the camera in a bitstream.

[0271] According to a sixth example there is provided a method comprising:

25 obtaining projected images which have been formed on the basis of images captured by a camera;

obtaining information of orientation of the camera; and

using the orientation information to rotate the projected images with reference to a coordinate system.

[0272] In some embodiments the method further comprises:

30 region-wise mapping the projected frame to form packed frames.

[0273] In some embodiments the method further comprises:

including information of orientation of the camera in a bitstream.

[0274] According to a seventh example there is provided a method comprising:
receiving encoded projected images which have been formed on the basis of images
captured by a camera;

decoding the encoded images to form reconstructed projected images;
5 obtaining information of orientation of the camera; and
using the orientation information to rotate the reconstructed projected images with
reference to a coordinate system.

[0275] In some embodiments wherein the encoded projected images have also been
region-wise mapped, and, the decoding further comprises:

10 decoding the encoded images to form reconstructed region-wise mapped images; and
region-wise back-mapping the reconstructed region-wise mapped images into the
reconstructed projected images.

[0276] In some embodiments the method further comprises:

obtaining the information of orientation of the camera from a bitstream.

15 [0277] According to an eighth example there is provided a method comprising:

back-projecting a motion field onto a first projection structure;

back-projecting the motion field from the first projection structure to a sphere to form
a spherically mapped motion field image;

mapping the spherically mapped motion field image onto a second projection
20 structure; and

mapping the motion field mapped onto the second projection structure onto a
reference motion field of a two-dimensional image.

[0278] In some embodiments the method further comprises using the reference motion
field in motion information prediction.

25 [0279] In some embodiments the method further comprises one of:

selecting the orientation of the first projection structure based on camera orientation
when the decoded picture corresponding to the motion field was captured;

using a default orientation with the first projection structure.

[0280] In some embodiments of the method:

30 the first projection structure has an orientation according to the camera orientation
when the decoded picture was captured; and

the second projection structure has an orientation matching that of the camera
orientation of the picture being encoded or decoded.

[0281] In some embodiments of the method:

the first projection structure has a default orientation; and

the second projection structure has an orientation matching the difference of the camera orientations for current picture being encoded or decoded and the decoded picture.

5 [0282] In some embodiments of the method the motion field is for an equirectangular panorama picture, wherein the method further comprises:

mapping the motion field onto a cylinder; and

mapping the motion field from the cylinder onto a sphere.

CLAIMS

1. A method for video coding comprising:

obtaining a first reconstructed picture of the video as a first three-dimensional

5 picture in a coordinate system;

obtaining a first rotation angle, wherein the first rotation angle is the absolute
rotation of the first reconstructed picture with a reference rotation;

obtaining a second rotation angle;

10 projecting the first three-dimensional picture onto a first projected picture on a first
geometrical projection structure;

rotating the first projected picture to the reference rotation based on the first
rotation angle to create a second projected picture;

rotating the second projected picture based on the second rotation angle to create a
third projected picture;

15 forming a first reference picture, said forming comprising unfolding the third
projected picture on the first geometrical projection structure into a second geometrical
projection structure;

predicting at least a block of a second reconstructed picture from the first reference
picture.

20

2. The method of claim 1, further comprising performing two or more of said
interpreting, projecting, and forming as a single process.

25 3. The method of claim 1, wherein the first and second reconstructed pictures
comply with an equirectangular panorama representation format.

4. The method according to any of claims 1 to 3 further comprising:

decoding a first coded picture into a first reconstructed picture; and

30 decoding a second coded picture into the second reconstructed picture; wherein the
decoding comprises said predicting.

5. The method of claim 4 further comprising:

decoding one or more syntax elements indicative of the rotation.

6. The method according to any of claims 1 to 3 further comprising:

encoding a first picture into a first coded picture, wherein the encoding comprises reconstructing the first reconstructed picture; and

5 encoding a second picture into a second coded picture; wherein the encoding comprises reconstructing the second reconstructed picture and said predicting.

7. The method of claim 6 further comprising:

obtaining a first orientation of an apparatus when capturing a first set of input images from which the first picture originates;

10 obtaining a second orientation of the apparatus when capturing a second set of input images from which the second picture originates;

deriving the rotation on the basis of the first orientation and the second orientation.

8. The method of claim 6 further comprising:

15 estimating the rotation based on the first picture and the second picture.

9. An apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:

20 obtain a first reconstructed picture of a video as a first three-dimensional picture in a coordinate system;

obtain a first rotation angle, wherein the first rotation angle is the absolute rotation of the first reconstructed picture with a reference rotation;

obtain a second rotation angle;

25 project the first three-dimensional picture onto a first projected picture on a first geometrical projection structure;

rotate the first projected picture to the reference rotation based on the first rotation angle to create a second projected picture;

30 rotate the second projected picture based on the second rotation angle to create a third projected picture;

form a first reference picture, said forming comprising unfolding the third projected picture on the first geometrical projection structure into a second geometrical projection structure;

predict at least a block of a second reconstructed picture from the first reference picture

10 11. The apparatus of claim 9, said at least one memory stored with code thereon,
which when executed by said at least one processor, causes the apparatus to perform at least:
perform two or more of said interpreting, projecting, and forming as a single
process.

10 11. The apparatus of claim 9, wherein the first and second reconstructed pictures
comply with an equirectangular panorama representation format.

12. The apparatus according to any of claims 9 to 11, said at least one memory
stored with code thereon, which when executed by said at least one processor, causes the
apparatus to perform at least:
15 decode a first coded picture into a first reconstructed picture; and
 decode a second coded picture into the second reconstructed picture; wherein the
decoding comprises said predicting.

13. The apparatus of claim 12, said at least one memory stored with code thereon,
20 which when executed by said at least one processor, causes the apparatus to perform at least:
 decode one or more syntax elements indicative of the rotation.

14. The apparatus according to any of claims 9 to 11, said at least one memory
stored with code thereon, which when executed by said at least one processor, causes the
25 apparatus to perform at least:
 encode a first picture into a first coded picture, wherein the encoding comprises
reconstructing the first reconstructed picture; and
 encode a second picture into a second coded picture; wherein the encoding
comprises reconstructing the second reconstructed picture and said predicting.

30 15. The apparatus of claim 14, said at least one memory stored with code thereon,
which when executed by said at least one processor, causes the apparatus to perform at least:
 obtain a first orientation of the apparatus when capturing a first set of input images
from which the first picture originates;

obtain a second orientation of the apparatus when capturing a second set of input images from which the second picture originates;

derive the rotation on the basis of the first orientation and the second orientation.

5 16. The apparatus of claim 14, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least: estimate the rotation based on the first picture and the second picture.

10 17. A computer readable storage medium comprising code for use by an apparatus, which when executed by a processor, causes the apparatus to perform at least:

obtain a first reconstructed picture of a video as a first three-dimensional picture in a coordinate system;

obtain a first rotation angle, wherein the first rotation angle is the absolute rotation of the first reconstructed picture with a reference rotation;

15 obtain a second rotation angle;

project the first three-dimensional picture onto a first projected picture on a first geometrical projection structure;

rotate the first projected picture to the reference rotation based on the first rotation angle to create a second projected picture;

20 rotate the second projected picture based on the second rotation angle to create a third projected picture;

form a first reference picture, said forming comprising unfolding the third projected picture on the first geometrical projection structure into a second geometrical projection structure;

25 predict at least a block of a second reconstructed picture from the first reference picture.

18. An apparatus comprising:

30 means for obtaining a first reconstructed picture of a video as a first three-dimensional picture in a coordinate system;

means for obtaining a first rotation angle, wherein the first rotation angle is the absolute rotation of the first reconstructed picture with a reference rotation;

means for obtaining a second rotation angle;

means for projecting the first three-dimensional picture onto a first projected picture on a first geometrical projection structure;

means for rotating the first projected picture to the reference rotation based on the first rotation angle to create a second projected picture;

5 means for rotating the second projected picture based on the second rotation angle to create a third projected picture;

means for forming a first reference picture, said forming comprising unfolding the third projected picture on the first geometrical projection structure into a second geometrical projection structure;

10 means for predicting at least a block of a second reconstructed picture from the first reference picture.

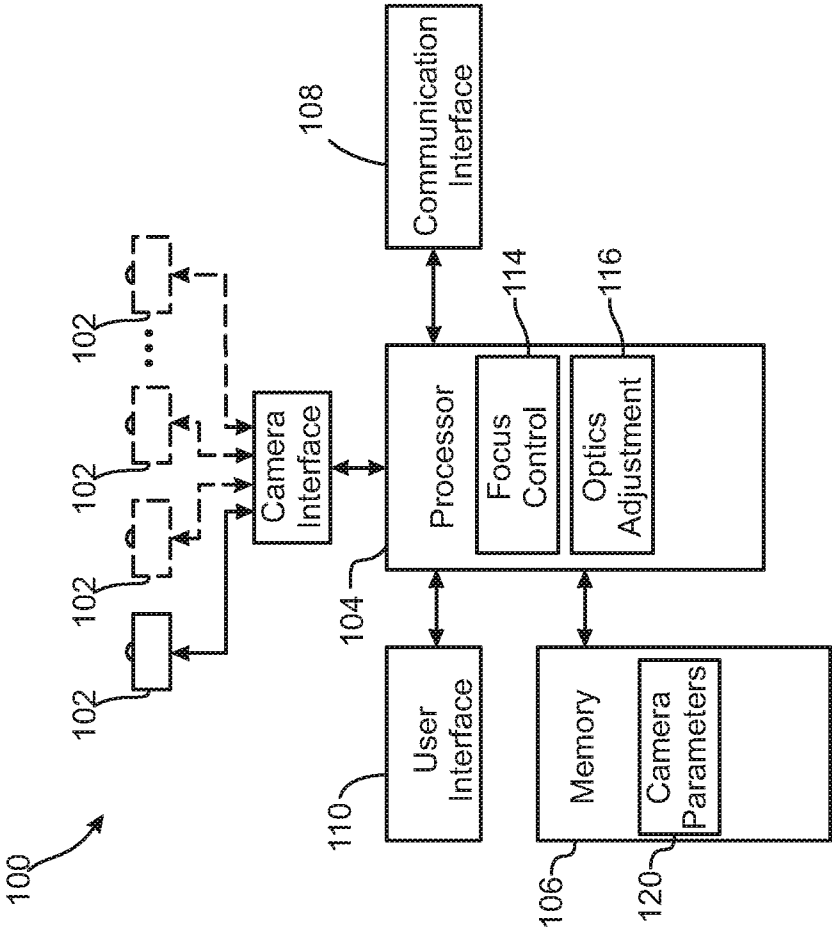


Fig. 1a

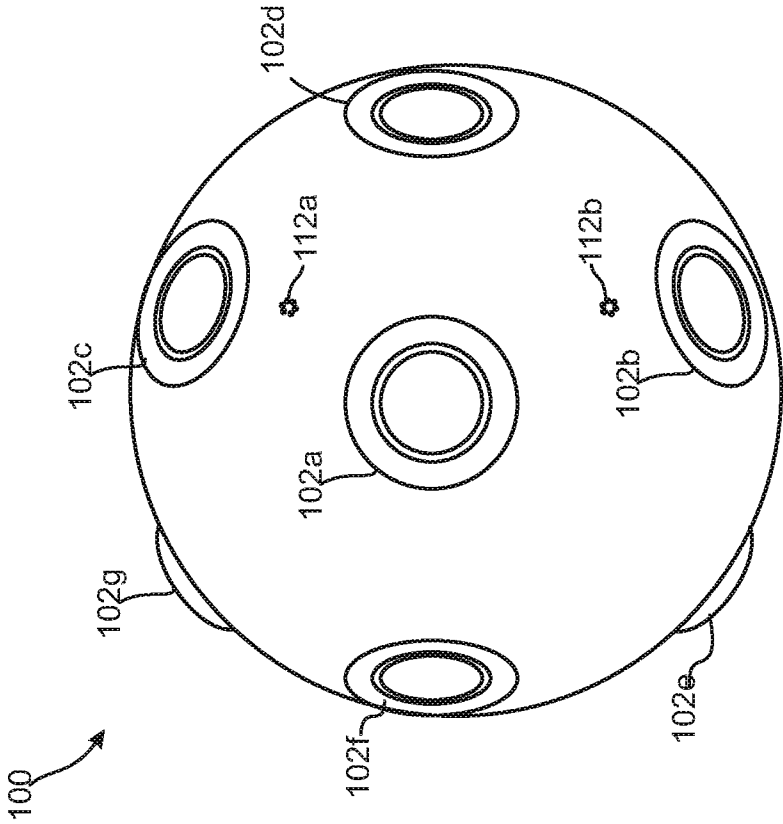


Fig. 1b

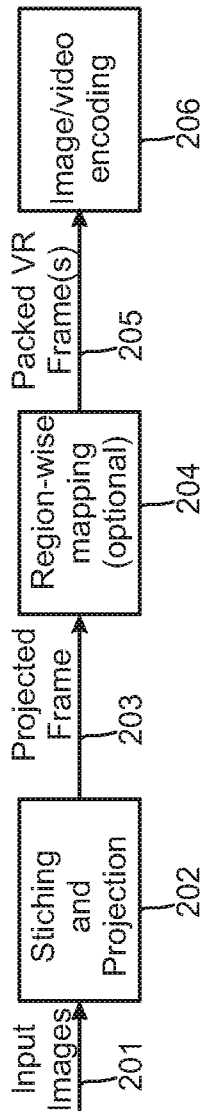


Fig. 2a

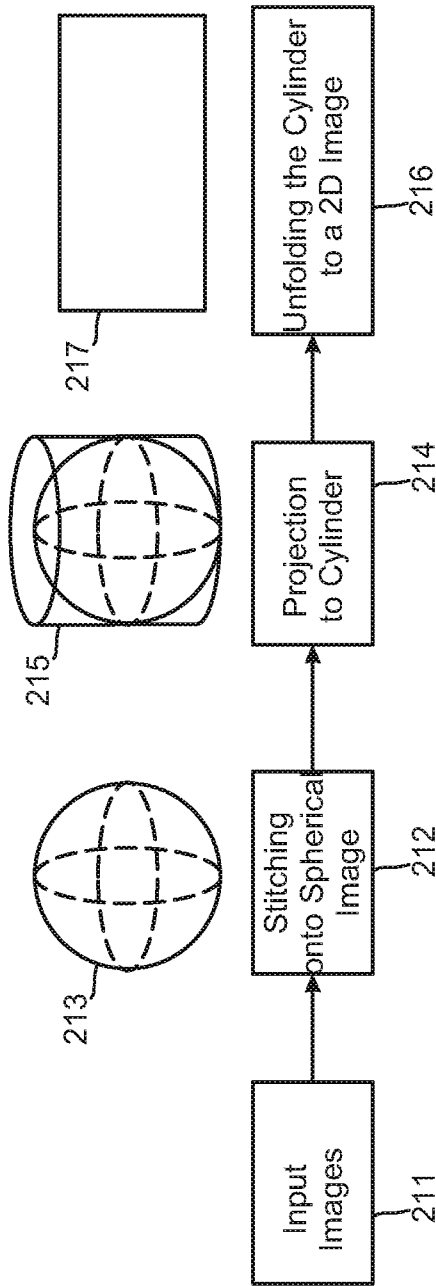


Fig. 2b

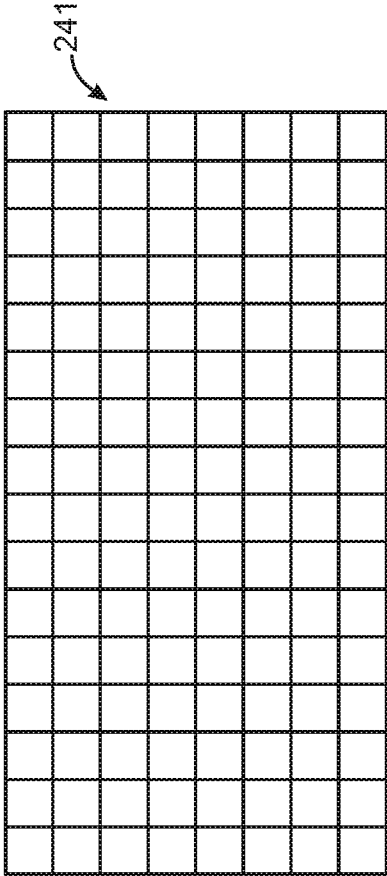


Fig. 3a

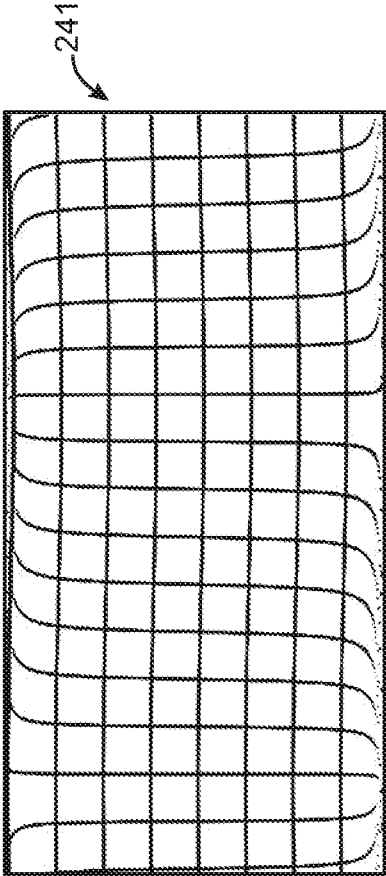


Fig. 3b

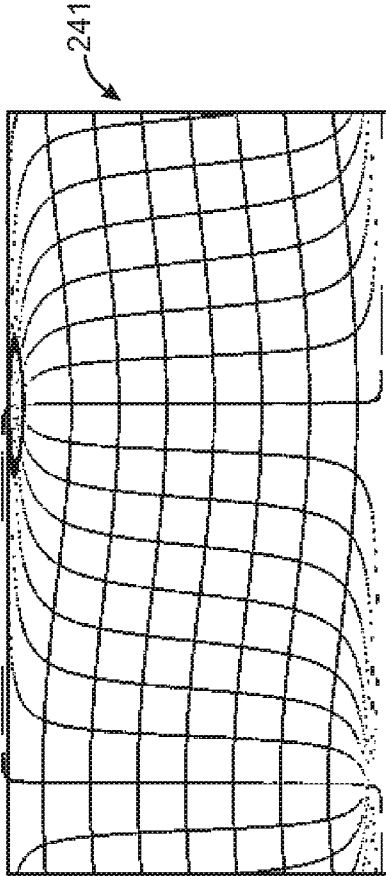


Fig. 3c

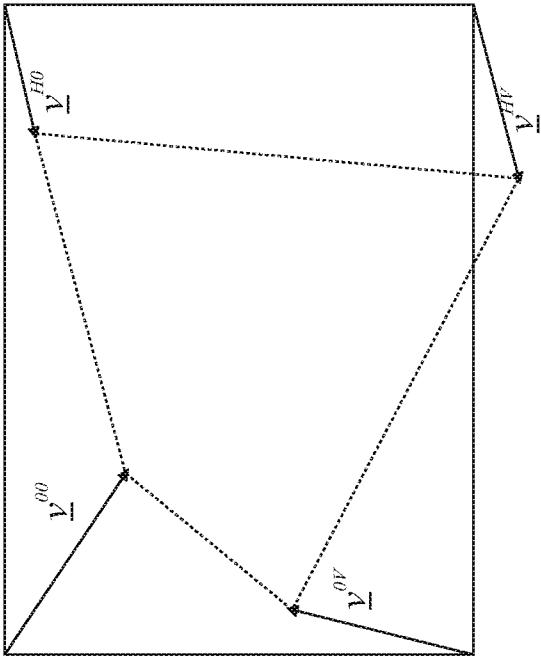


Fig. 3d

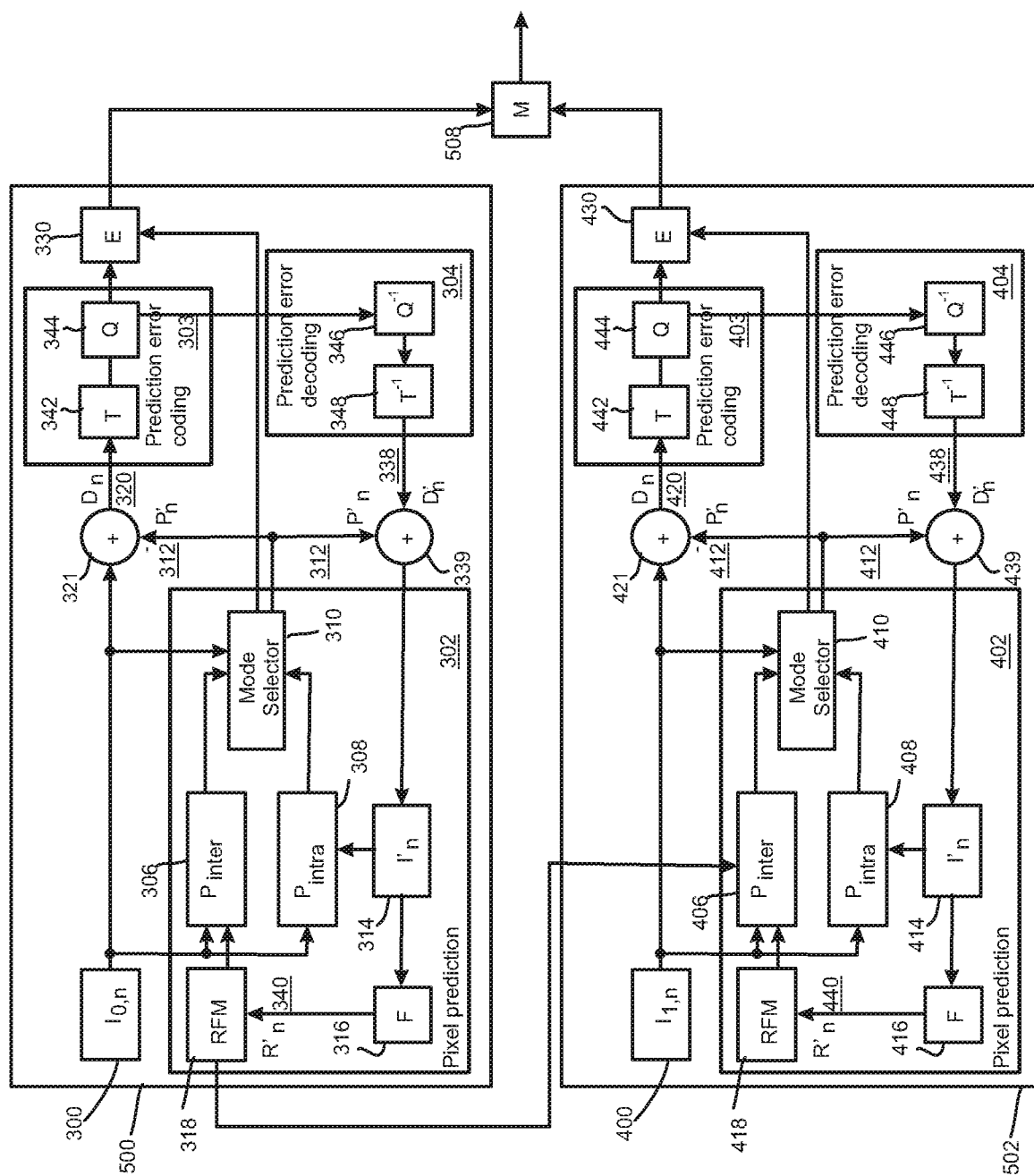


Fig. 4a

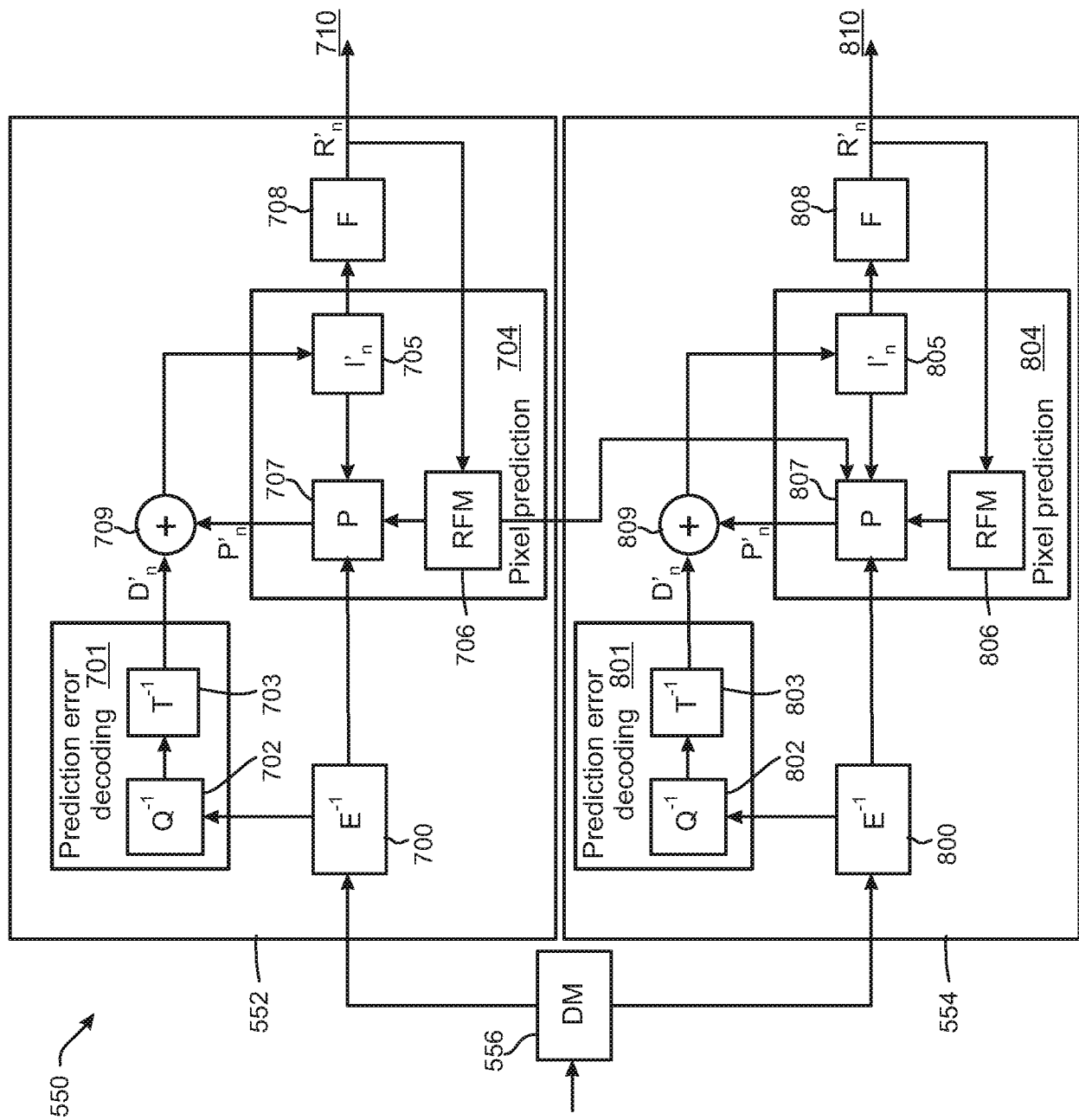


Fig. 4b

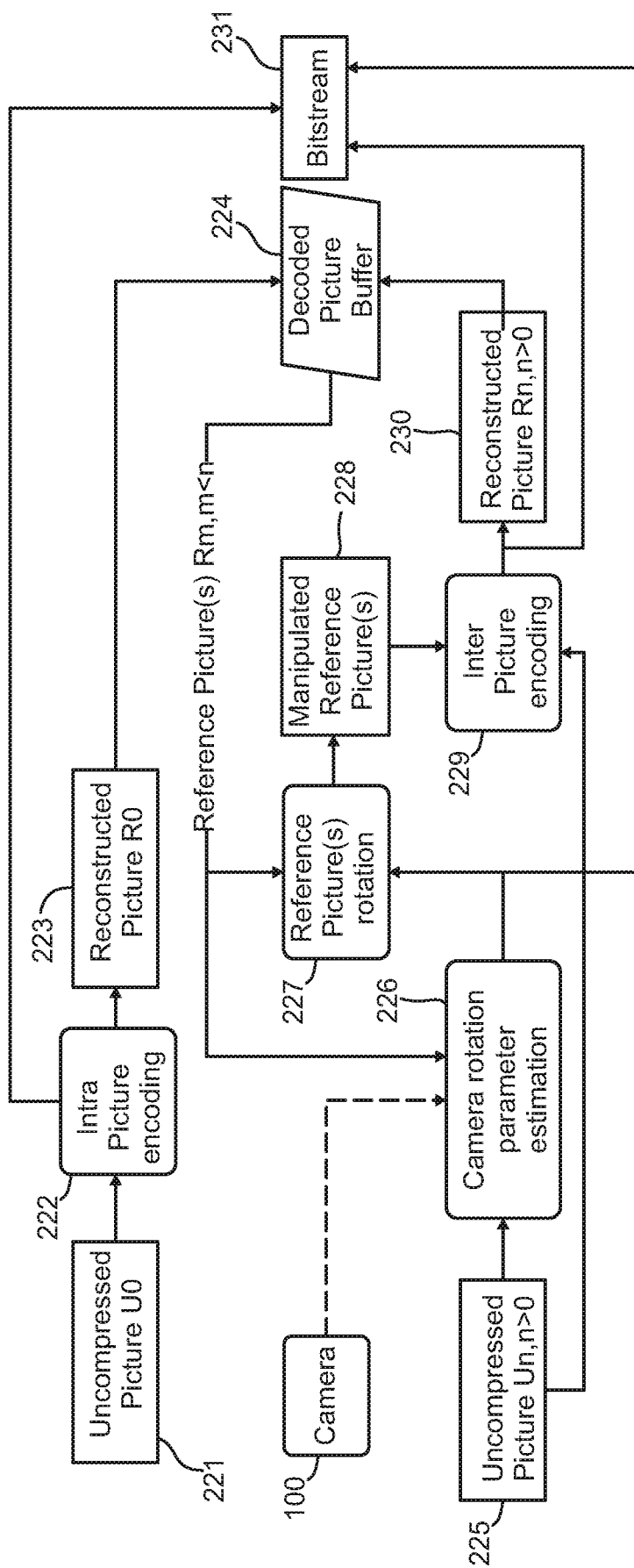


Fig. 5a

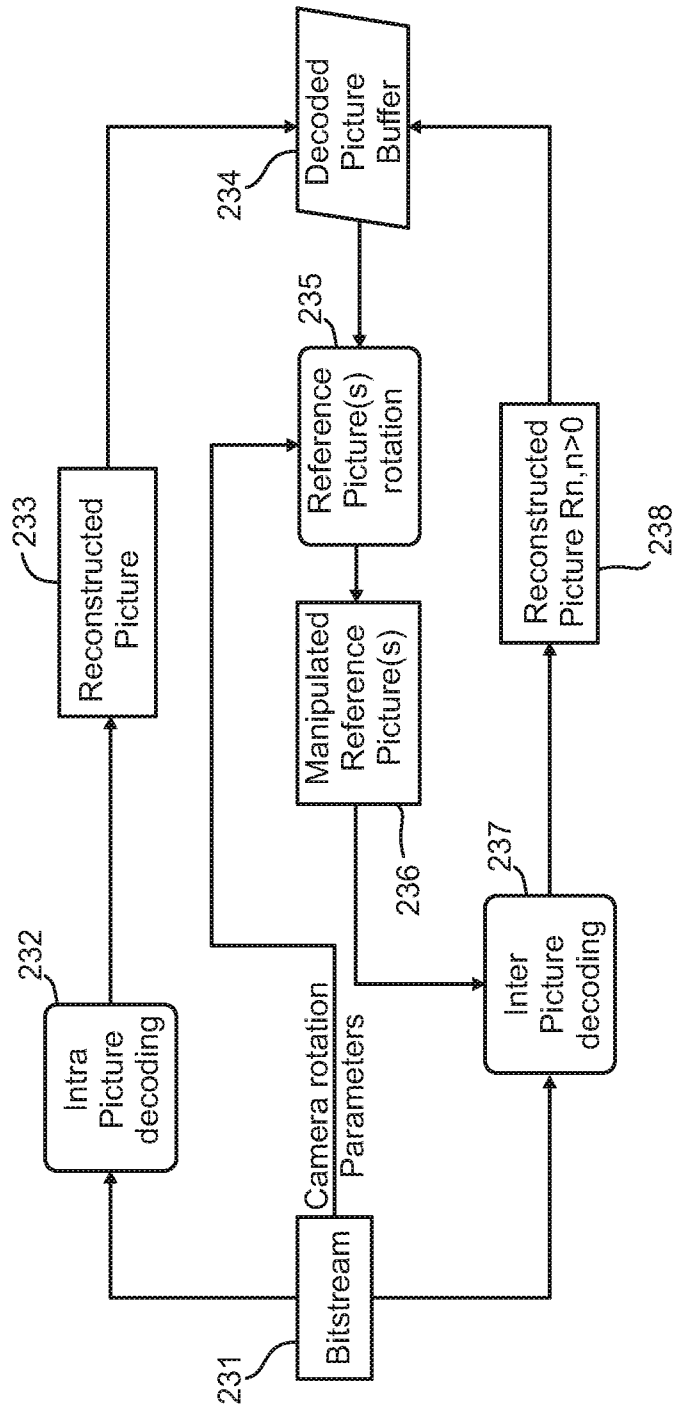


Fig. 5b

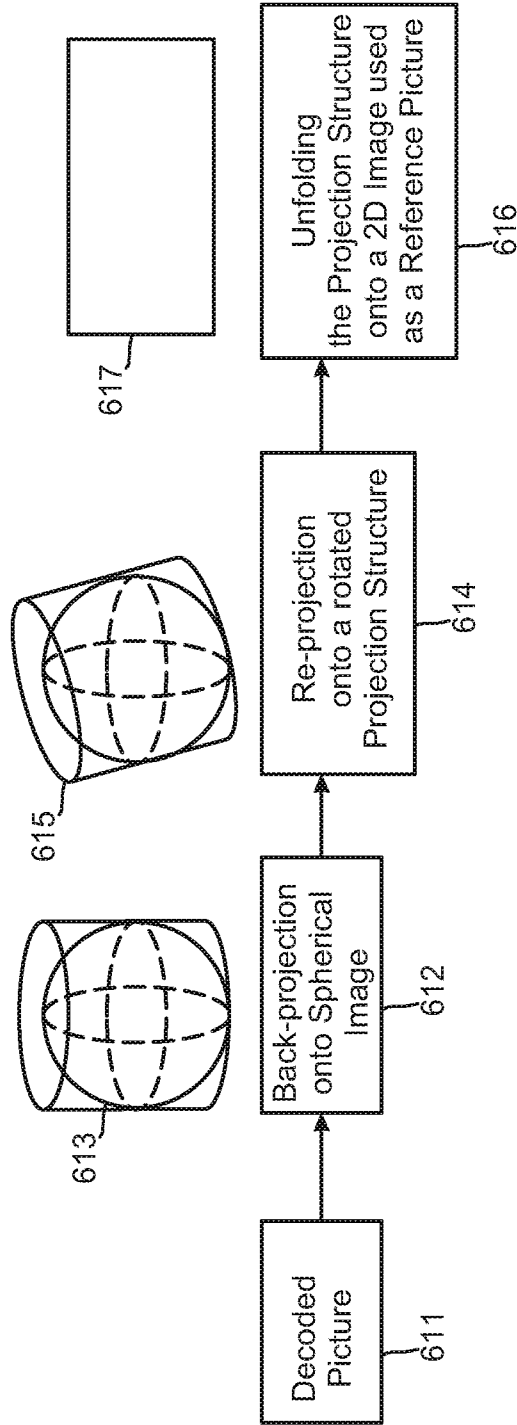


Fig. 6

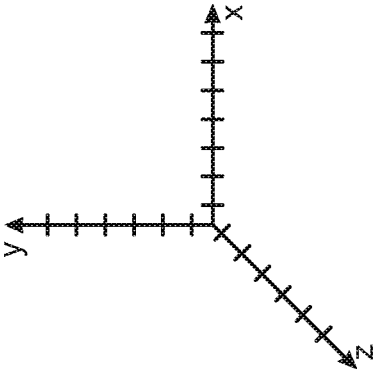


Fig. 7a

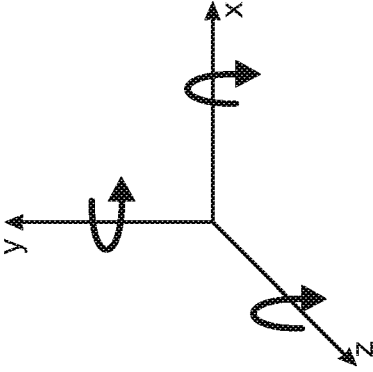


Fig. 7b

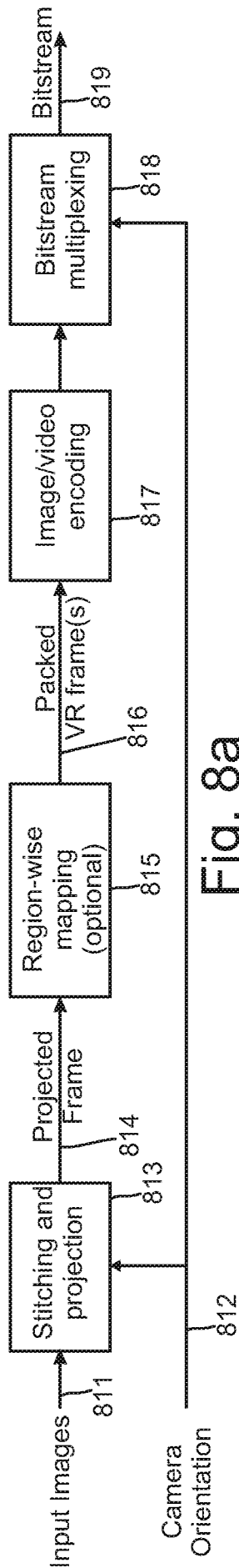


Fig. 8a

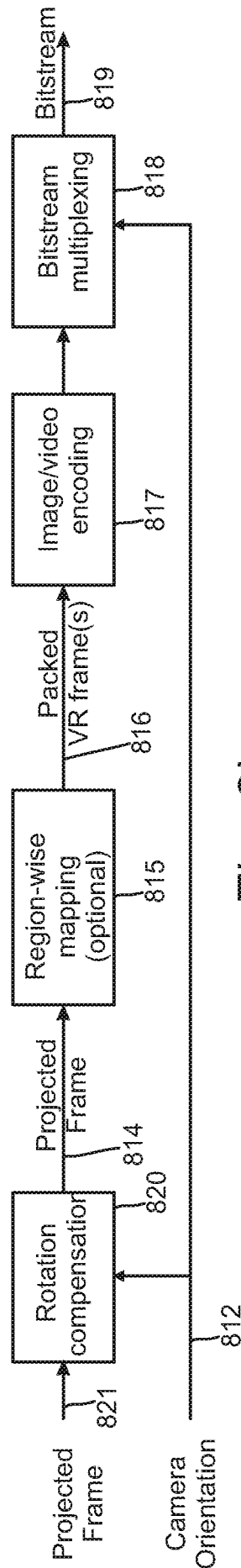


Fig. 8b

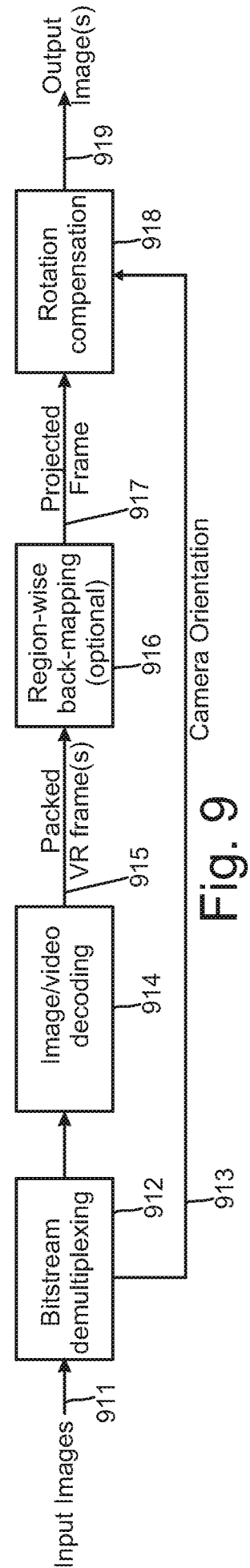


Fig. 9

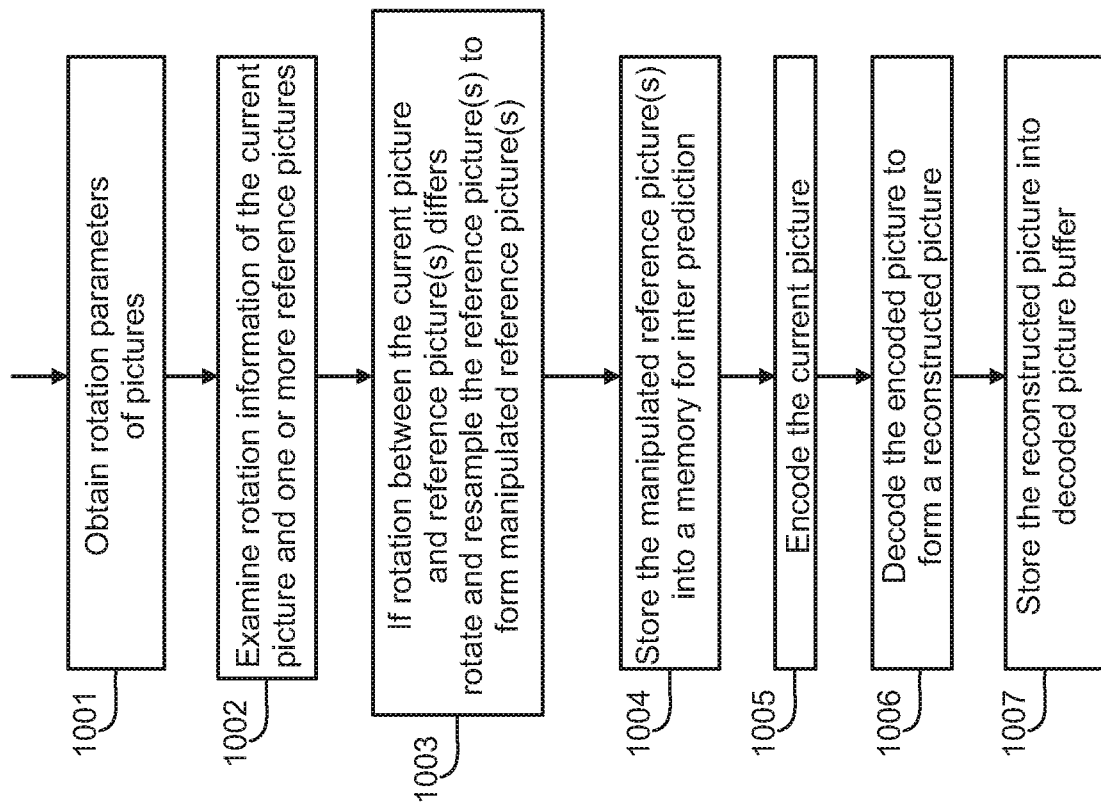


Fig. 10a

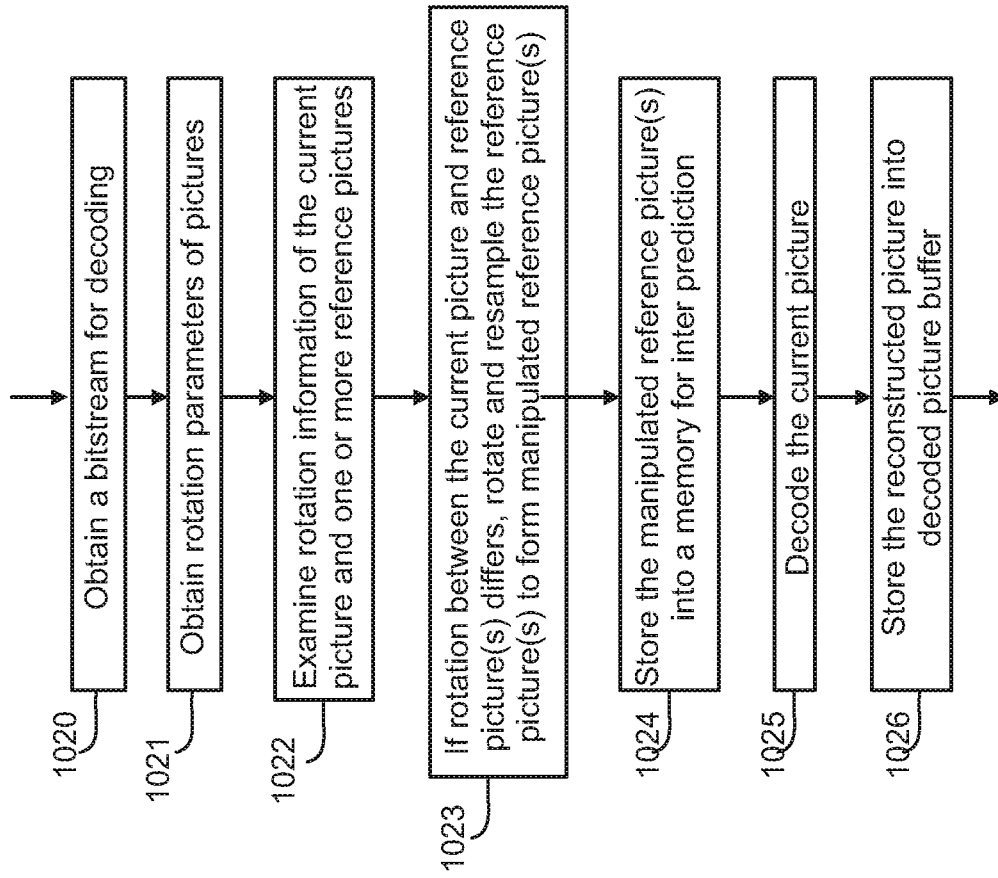


Fig. 10b

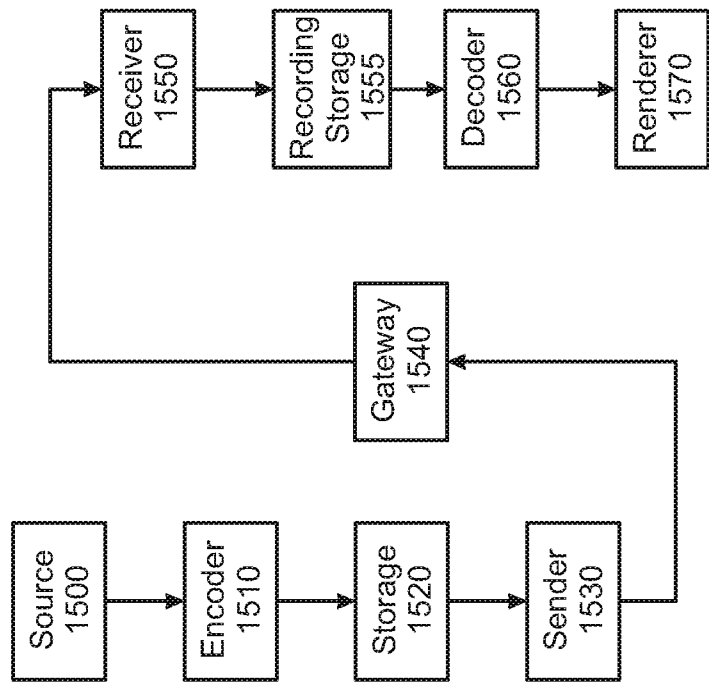


Fig. 12

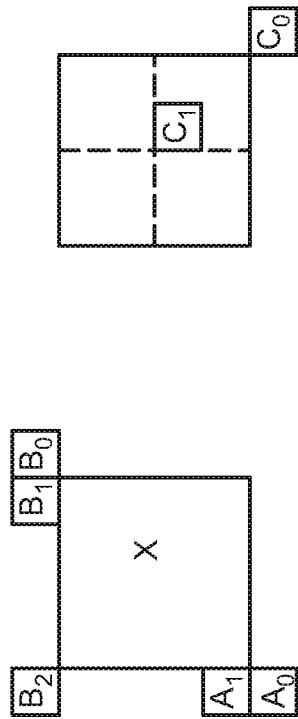


Fig. 11b

Fig. 11a

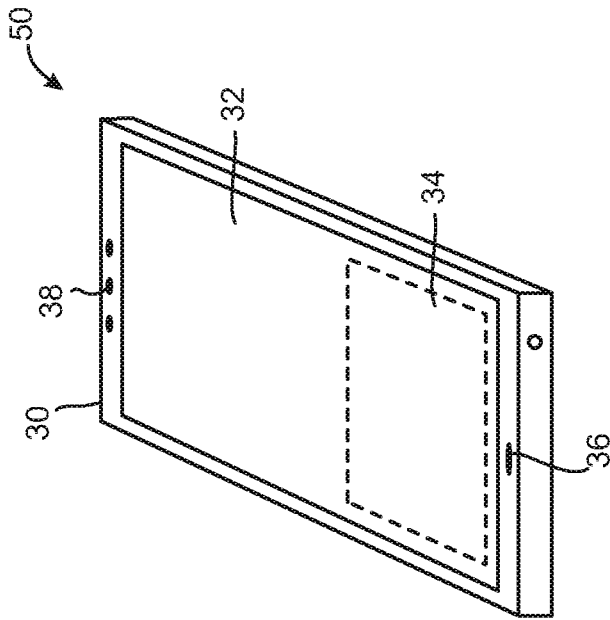


Fig. 14

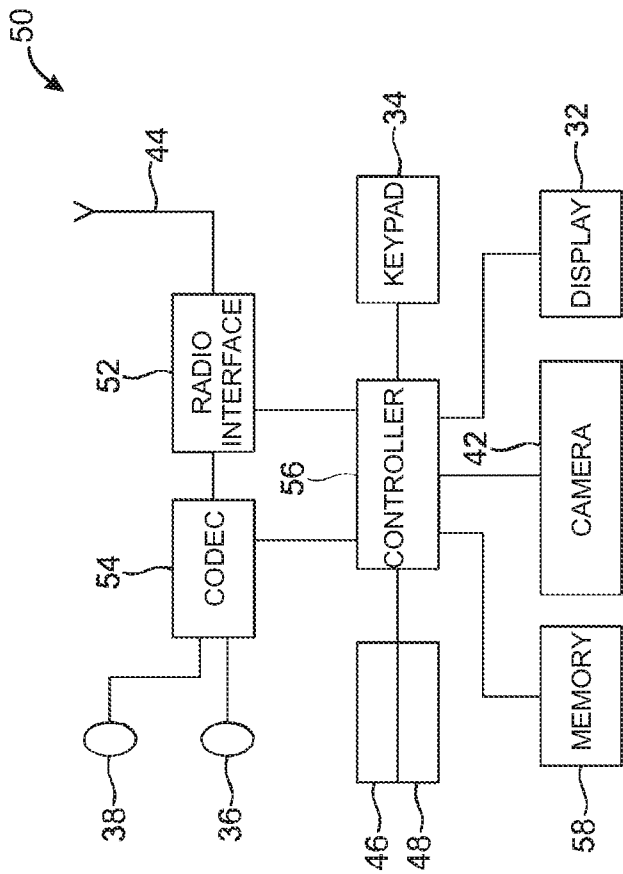


Fig. 13

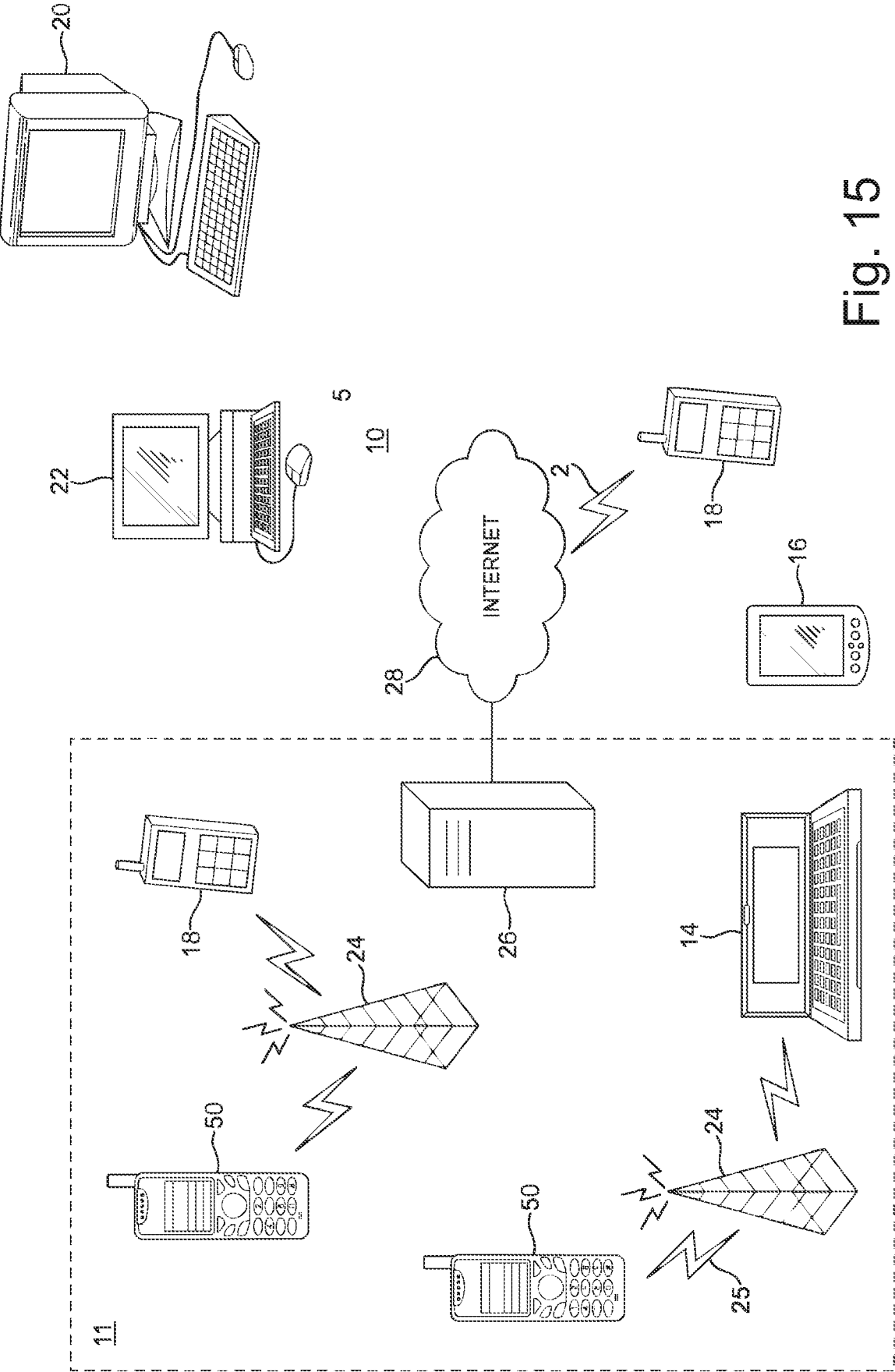


Fig. 15

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI2017/050951

A. CLASSIFICATION OF SUBJECT MATTER

See extra sheet

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)

IPC: H04N, G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
FI, SE, NO, DK

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

EPODOC, EPO-Internal full-text databases, WPIAP, XP3GPP, XPAIP, XPESP, XPETSI, XPI3E, XPIEE, XPIETF, XPIOP, XPIPCOM, XPJPEG, XPMISC, XPOAC, XPRD XPTK, COMPDX, EMBASE, INSPEC, TDB, NPL, Google

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CHAN YUL PARK et al. A hybrid motion compensation for wide and fixed field-of-view image sequences. 2012 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Seoul, South Korea: IEEE, 2012-06-27, ISSN 2155-504, ISBN 978-1-4673-0295-1, <DOI:10.1109/BMSB.2012.6264322>. Retrieved from < https://ieeexplore.ieee.org/abstract/document/6264322/ > [retrieved on 2018-05-02]. abstract; sections I, II.B	1-18
Y	US 9277122 B1 (IMURA KEN [US] et al.) 01 March 2016 (01.03.2016) abstract; column 1, line 58 – column 2, line 23; column 4, lines 43–45; column 6, lines 4–20; column 9, lines 4–25; claim 1; Figs. 3, 8	1-18

☒ Further documents are listed in the continuation of Box C.
 ☒ See patent family annex.

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search

02 May 2018 (02.05.2018)

Date of mailing of the international search report

03 May 2018 (03.05.2018)

Name and mailing address of the ISA/FI
Finnish Patent and Registration Office
FI-00091 PRH, FINLAND

Facsimile No. +358 29 509 5328

Authorized officer

Ari Peltola

Telephone No. +358 29 509 5000

INTERNATIONAL SEARCH REPORT

International application No.

PCT/FI2017/050951

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	HO HYEONG RYU et al. Prefiltering and Postfiltering Based on Global Motion Compensation for Improving Coding Efficiency in H.264 and HEVC Codecs. IEICE TRANSACTIONS on Information and Systems 2016-10-07, Vol. E100-D, No. 1, pp. 160-165, ISSN 1745-1361, <DOI:10.1587/transinf.2016EDP7273>. Retrieved from https://search.ieice.org/bin/summary.php?id=e100-d_1_160 [retrieved on 2018-04-27]. abstract; sections 1, 2.1, 2.2; Fig. 2	1-18
A	US 2008143748 A1 (JIN HAILIN [US] et al.) 19 June 2008 (19.06.2008) the whole document	1-18

INTERNATIONAL SEARCH REPORT
Information on Patent Family Members

International application No.
PCT/FI2017/050951

Patent document cited in search report	Publication date	Patent family members(s)	Publication date
US 9277122 B1	01/03/2016	WO 2017027884 A1	16/02/2017
.....			
US 2008143748 A1	19/06/2008	US 8988466 B2	24/03/2015
		WO 2008076766 A1	26/06/2008
.....			

CLASSIFICATION OF SUBJECT MATTER

IPC
H04N 19/503 (2014.01)
H04N 19/597 (2014.01)
H04N 19/105 (2014.01)
G06T 3/60 (2006.01)
G06T 3/00 (2006.01)