(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 May 2018 (17.05.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/087425 A1

(51) International Patent Classification:
H04N 19/563 (2014.01) *G06T 17/30* (2006.01)
H04N 19/593 (2014.01) *G06T 17/10* (2006.01)
H04N 19/597 (2014.01)

(21) International Application Number:
PCT/FI2017/050756

(22) International Filing Date:
02 November 2017 (02.11.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1618784.1 08 November 2016 (08.11.2016) GB

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI]; Karaportti 3, 02610 Espoo (FI).

(72) Inventors: **HANNUKSELA, Miska**; Rusthollinrinne 2, 33610 Tampere (FI). **KAMMACHI SREEDHAR, Kashyap**; Pyykkiojankatu 2 B 18, 33710 Tampere (FI). **AMINLOU, Alireza**; Vaajakatu 5 G 152, 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,

WO 2018/087425 A1

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

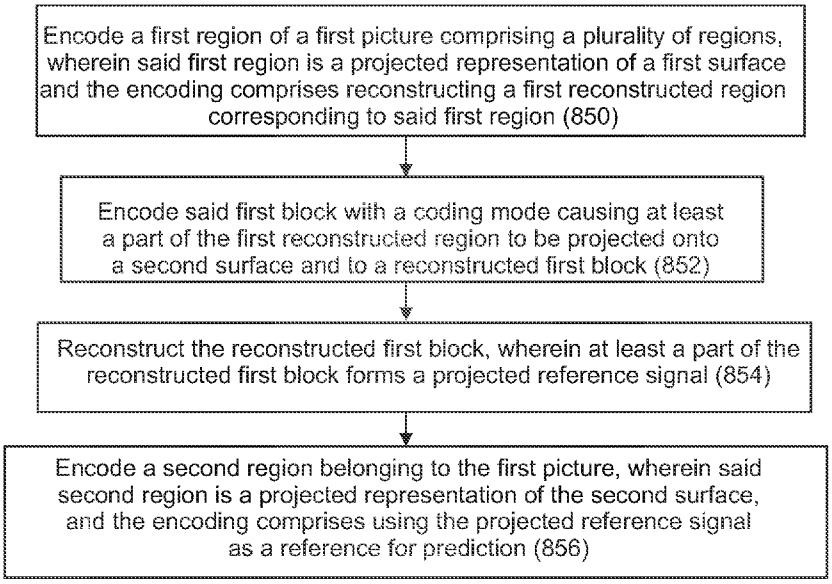


Fig. 8a

(57) Abstract: A method comprising: encoding a first region of a first picture comprising a plurality of regions, wherein said first region is a projected representation of a first surface and the encoding comprises reconstructing a first reconstructed region corresponding to said first region; encoding at least a first block of the first picture with an encoding mode causing at least a part of the first reconstructed region to be projected onto a second surface and further to a reconstructed first block; reconstructing the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal; and encoding at least a second region of the plurality of regions of the first picture, wherein said second region is a projected representation of the second surface and said encoding comprises using the projected reference signal as a reference for prediction.

[Continued on next page]

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))*

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

TECHNICAL FIELD

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

BACKGROUND

[0002] 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 video information at a lower bitrate than otherwise might be needed.

[0003] 360-degree panoramic content cover horizontally the full 360-degree field-of-view around the capturing position of an imaging device. A specific projection for mapping a panoramic image covering 360-degree field-of-view horizontally and 180-degree field-of-view vertically to a rectangular two-dimensional image plane is known as a monoscopic cubemap projection. Therein, 360-degree video/image data is projected onto six faces of a cube, and the cube faces can be unfolded to be represented as a 2D image on an image frame.

[0004] The cube faces may be arranged on frame in multiple ways, but whatever way is selected, there will always be intra prediction discontinuities over some cube face boundaries. When in-loop deblocking filtering is performed across such cube face boundaries, undesirable "leaking" of sample information from one cube face to another happens when these cube faces are actually not adjacent. It is difficult, maybe impossible, to avoid the introduction of such error, since the deblocking filtering is performed after reconstructing the block based on the reconstructed prediction error.

SUMMARY

[0005] Some embodiments provide a method for encoding and decoding video information. In some embodiments of the present invention there is provided a method, apparatus and computer program product for video coding.

[0006] A method according to a first aspect comprises encoding a first region of a first picture comprising a plurality of regions, wherein said first region is a projected representation of a first surface and the encoding comprises reconstructing a first

reconstructed region corresponding to said first region; encoding at least a first block of the first picture with an encoding mode causing at least a part of the first reconstructed region to be projected onto a second surface and further to a reconstructed first block; reconstructing the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal; and encoding at least a second region of the plurality of regions of the first picture, wherein said second region is a projected representation of the second surface and said encoding comprises using the projected reference signal as a reference for prediction.

[0007] According to an embodiment, said coding mode is indicative of one or more of the following:

- the first reconstructed region or the at least a part of the first reconstructed region;
- the projection and/or transformation to be applied to the at least a part of the first reconstructed region;
- the first surface;
- the second surface.

[0008] According to an embodiment, the method further comprises specifying, by an encoder, with a first coding mode or a first parameter value of the coding mode, or inferring, by the encoder, that the first coding mode is applied in reconstruction in conventional order of processing blocks; or specifying, by an encoder, with a second coding mode or a second parameter value of the coding mode, or inferring, by the encoder, that the second coding mode is applied after reconstructing the picture otherwise.

[0009] According to an embodiment, the method further comprises obtaining a projected frame; and mapping a first region of the projected frame and a second region of the projected frame onto the first picture, wherein said first region of the projected frame is a projected representation of a first surface and said second region of the projected frame is a projected representation of a second surface, wherein said first region of the first picture corresponds to said first region of the projected frame and said second region of the first picture corresponds to said second region of the projected frame, and wherein said at least the first block of the first picture is spatially adjacent to the second region of the first picture, the first block being neither a part of the first region of the first picture nor the second region of the first picture.

[0010] According to an embodiment, the method further comprises indicating performed mapping and/or a location of the at least first block to an encoder; and in response to receiving the indication in the encoder, choosing a coding mode for the at least first block that causes at least a part of the first reconstructed region to be projected onto the second surface and further to a reconstructed first block.

[0011] According to an embodiment, the plurality of regions in the coded picture corresponds to only a part of the panorama image.

[0012] According to an embodiment, the first reconstructed region and the second reconstructed region are of different projection type.

[0013] According to an embodiment, the first reconstructed region is of cylindrical or equirectangular projection type, and the second reconstructed region is a top or bottom face of the cylinder or vertically truncated equirectangular panorama, and the second region is formed as a block-aligned bounding area covering the top or bottom face of the cylinder or truncated sphere.

[0014] According to an embodiment, said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises resampling the first reconstructed region.

[0015] According to an embodiment, said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises performing a geometric transform.

[0016] A second aspect relates to a method comprising: decoding, from a bitstream, a first encoded region of a plurality of regions of a first coded picture into the first reconstructed region, wherein said first reconstructed region is a projected representation of a first surface; decoding at least a first coded block of the first coded picture , the first coded block having a coding mode causing at least a part of the first reconstructed region to be projected onto a second surface and further to a reconstructed first block; reconstructing the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal; and decoding at least a second coded region of the plurality of regions of the first picture into a second reconstructed region, where said second reconstructed region is a projected representation of the second surface and said decoding comprises using the projected reference signal as a reference for prediction.

[0017] Further aspects relate to apparatuses and computer program products and/or computer readable storage medium stored with code thereon arranged to perform the above methods

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] For better understanding of the present invention, reference will now be made by way of example to the accompanying drawings in which:

[0019] Figure 1 shows schematically an electronic device employing embodiments of the invention;

[0020] Figure 2 shows schematically a user equipment suitable for employing embodiments of the invention;

[0021] Figure 3 further shows schematically electronic devices employing embodiments of the invention connected using wireless and wired network connections;

[0022] Figure 4 shows schematically an encoder suitable for implementing embodiments of the invention;

[0023] Figure 5a shows spatial candidate sources of the candidate motion vector predictor, in accordance with an embodiment;

[0024] Figure 5b shows temporal candidate sources of the candidate motion vector predictor, in accordance with an embodiment;

[0025] Figure 6a shows an example of stitching, projecting and mapped images of the same time instance onto a packed virtual reality frame;

[0026] Figure 6b shows a process of forming a monoscopic equirectangular panorama picture;

[0027] Figures 7a, 7b,7c show a 360 degree image projection on a monoscopic cubemap and two alternatives for representing the cube faces as a 2D image;

[0028] Figure 8a shows a flow chart of an encoding method involving reference sample projection in accordance with an embodiment;

[0029] Figure 8b shows a flow chart of a decoding method involving reference sample projection according to an embodiment;

[0030] Figure 9 illustrates an unfolded cubemap where reference samples are projected across cube face boundaries in accordance with an embodiment;

[0031] Figure 10 illustrates an unfolded cubemap where reference samples are projected across cube face boundaries in accordance with another embodiment;

[0032] Figures 11a, 11b show an example of using different projection types for different regions of an image in accordance with an embodiment;

[0033] Figures 12a, 12b show examples of block-aligned projections of regions in accordance with various embodiments;

[0034] Figure 13 shows an example of resampling cube face into different spatial resolution in accordance with an embodiment;

[0035] Figures 14a, 14b show an example of an equirectangular panorama divided into three stripes and arranging the resampled top and bottom stripes and the middle stripe into constituent frame partitions in accordance with an embodiment;

[0036] Figures 15a, 15b show an example of logically partitioning an equirectangular panorama divided into two constituent partitions and arranging the stripes into constituent frame partitions in accordance with an embodiment;

[0037] Figure 16 shows a schematic diagram of a decoder suitable for implementing embodiments of the invention; and

[0038] Figure 17 shows a schematic diagram of an example multimedia communication system within which various embodiments may be implemented.

DETAILED DESCRIPTON OF SOME EXAMPLE EMBODIMENTS

[0039] In the following, several embodiments of the invention will be described in the context of one video coding arrangement. It is to be noted, however, that the invention is not limited to this particular arrangement. In fact, the different embodiments have applications widely in any environment where improvement of non-scalable, scalable and/or multiview video coding is required. For example, the invention may be applicable to video coding systems like streaming systems, DVD players, digital television receivers, personal video recorders, systems and computer programs on personal computers, handheld computers and communication devices, as well as network elements such as transcoders and cloud computing arrangements where video data is handled.

[0040] The following describes in further detail suitable apparatus and possible mechanisms for implementing some embodiments. In this regard reference is first made to Figures 1 and 2, where Figure 1 shows a block diagram of a video coding system according to an example embodiment as a schematic block diagram of an exemplary apparatus or electronic device 50, which may incorporate a codec according to an embodiment of the invention. Figure 2 shows a layout of an apparatus according to an example embodiment. The elements of Figs. 1 and 2 will be explained next.

[0041] 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 encoding and decoding or encoding or decoding video images.

[0042] 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.

[0043] 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 apparatus may further comprise a camera 42 capable of recording or capturing images and/or video. The apparatus 50 may further comprise an infrared port 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.

[0044] 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 in the form of image and audio 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.

[0045] The apparatus 50 may further comprise a card reader 48 and a smart card 46, for example a UICC and UICC reader for providing user information and being suitable for providing authentication information for authentication and authorization of the user at a network.

[0046] 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 44 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).

[0047] The apparatus 50 may comprise a camera capable of recording or detecting individual frames which are then passed to the codec 54 or the controller for processing. The apparatus may receive the video image data for processing from another device prior to transmission and/or storage. The apparatus 50 may also receive either wirelessly or by a wired connection the image for coding/decoding.

[0048] With respect to Figure 3, 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 or wireless networks including, but not limited to a wireless cellular telephone network (such as a GSM, UMTS, CDMA network etc), a wireless local area network (WLAN) such as defined by any of the IEEE 802.x standards, a Bluetooth personal area network, an Ethernet local area network, a token ring local area network, a wide area network, and the Internet.

[0049] The system 10 may include both wired and wireless communication devices and/or apparatus 50 suitable for implementing embodiments of the invention.

[0050] For example, the system shown in Figure 3 shows a mobile telephone network 11 and a representation of the internet 28. Connectivity to the internet 28 may include, but is not limited to, long range wireless connections, short range wireless connections, and various wired connections including, but not limited to, telephone lines, cable lines, power lines, and similar communication pathways.

[0051] 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. The apparatus 50 may be stationary or mobile when carried by an individual who is moving. The apparatus 50 may also be located in a mode of transport including, but not limited to, a car, a truck, a taxi, a bus, a train, a boat, an airplane, a bicycle, a motorcycle or any similar suitable mode of transport.

[0052] The embodiments may also be implemented in a set-top box; i.e. a digital TV receiver, which may/may not have a display or wireless capabilities, in tablets or (laptop) personal computers (PC), which have hardware or software or combination of the encoder/decoder implementations, in various operating systems, and in chipsets, processors, DSPs and/or embedded systems offering hardware/software based coding.

[0053] Some or further apparatus may send and receive calls and messages and communicate with service providers through a wireless connection 25 to a base station 24.

The base station 24 may be connected to a network server 26 that allows communication between the mobile telephone network 11 and the internet 28. The system may include additional communication devices and communication devices of various types.

[0054] The communication devices may communicate using various transmission technologies including, but not limited to, code division multiple access (CDMA), global systems for mobile communications (GSM), universal mobile telecommunications system (UMTS), time divisional multiple access (TDMA), frequency division multiple access (FDMA), transmission control protocol-internet protocol (TCP-IP), short messaging service (SMS), multimedia messaging service (MMS), email, instant messaging service (IMS), Bluetooth, IEEE 802.11 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.

[0055] 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.

[0056] 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.

[0057] Video codec consists of 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.

[0058] 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). Many coding schemes enable an encoder to indicate a quantization parameter (QP) in the bitstream, and respectively a decoder can decode quantization parameter (QP) from the bitstream. QP may indicate the quantization step size or the quantization points used in the coefficient quantization. QP may be indicated for various scopes, such as sequence, picture, or block.

[0059] 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. Inter coding may refer to coding modes where inter prediction is applied.

[0060] 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.

[0061] 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.

[0062] 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.

[0063] A motion vector may be considered to represent the displacement of an 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 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 signaling 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 or co-located blocks in temporal reference picture. Motion vectors may have sub-pixel accuracy (e.g. quarter-pixel accuracy), and sample values in fractional-pixel positions may be obtained using interpolation filtering e.g. with a finite impulse response (FIR) filter.

[0064] Figure 4 shows a block diagram of a video encoder suitable for employing embodiments of the invention. Figure 4 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 4 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 4 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. An intra picture encoding method is such that the inter-predictor 306 or its output is omitted and only intra prediction is in use. An inter picture encoding method is such that the inter-predictor 306 is in use and its output is considered in the mode selection 310 and hence inter prediction may be chosen by the encoder. 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.

[0065] 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.

[0066] 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 regional reference frame processing unit 315, 415. The regional reference frame processing unit 315, 415 may regionally resample and/or rearrange the preliminary reconstructed image according to one or more different embodiments described further below to produce a regionally processed reference image. The regionally processed reference image may be passed 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.

[0067] 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.

[0068] It needs to be understood that the regional reference frame processing unit 315, 415 and the filter 316, 416 may, in some embodiments, be located in opposite order in Figure 4. It also needs to be understood that in some embodiments parts of the filtering performed by the filter 316, 416 may be performed prior to regional reference frame processing 315, 415, while the remaining parts may be performed after the regional reference frame processing 315, 415. Likewise, some parts of the regional reference frame processing 315, 415 (e.g. resampling) may be performed prior to the filter 316, 416, while the remaining parts of the reference frame processing 315, 415 (e.g. rearranging) may be performed after the filter 316, 416.

[0069] 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.

[0070] 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.

[0071] 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.

[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). Subsequent versions of H.265/HEVC have included scalable, multiview, fidelity range, three-dimensional, and screen content coding extensions, which may be abbreviated SHVC, MV-HEVC, REXT, 3D-HEVC, and SCC respectively.

[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 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 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 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.

[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 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 decoded 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.
- 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 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.

[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 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 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 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 may be constrained to form 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.

[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 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.

[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. Deblocking may be used to smooth out discontinuities at block boundaries and it may be comprise an averaging filtering with adaptive filter tap weights, e.g. based on quantization parameter and/or coding mode, such as intra or inter mode.

[0098] The SAO used in HEVC is described here as an example, while it needs to be algorithm that SAO could be implemented with other coding standards or systems similarly, either entirely (using both band offset and edge offset modes) or in parts (using either mode). In the SAO algorithm, samples in a CTU are classified according to a set of rules and each classified set of samples are enhanced by adding offset values. The offset values are signalled in the bitstream. There are two types of offsets: 1) Band offset 2) Edge offset. For a CTU, either no SAO or band offset or edge offset is employed. Choice of whether no SAO or band

or edge offset to be used may be decided by the encoder with e.g. rate distortion optimization (RDO) and signaled to the decoder. In the band offset, the whole range of sample values is divided into a certain number of bands, e.g. 32 equal-width bands. For example, for 8-bit samples, width of a band is 8 (=256/32). Out of the total number of bands, a certain number of bands, e.g. four, are selected and different offsets are signalled for each of the selected bands. The selection decision is made by the encoder and may be signalled as follows: The index of the first band is signalled and then it is inferred that the following four bands are the chosen ones. The band offset may be useful in correcting errors in smooth regions. In the edge offset type, the edge offset (EO) type may be chosen out of four possible types (or edge classifications) where each type is associated with a direction: 1) vertical, 2) horizontal, 3) 135 degrees diagonal, and 4) 45 degrees diagonal. The choice of the direction is given by the encoder and signalled to the decoder. Each type defines the location of two neighbour samples for a given sample based on the angle. Then each sample in the CTU is classified into one of five categories based on comparison of the sample value against the values of the two neighbour samples. The five categories are described as follows: i) Current sample value is smaller than the two neighbour samples. ii) Current sample value is smaller than one of the neighbors and equal to the other neighbor. iii) Current sample value is greater than one of the neighbors and equal to the other neighbor. iv) Current sample value is greater than two neighbor samples. v) None of the previous. These five categories are not required to be signalled to the decoder because the classification is based on only reconstructed samples, which may be available and identical in both the encoder and decoder. After each sample in an edge offset type CTU is classified as one of the five categories, an offset value for each of the first four categories is determined and signalled to the decoder. The offset for each category is added to the sample values associated with the corresponding category. Edge offsets may be effective in correcting ringing artifacts. The SAO parameters may be signalled as interleaved in CTU data.

[0099] The adaptive loop filter (ALF) is another method to enhance quality of the reconstructed samples. This may be achieved by filtering the sample values in the loop. The encoder may determine which region(s) of the pictures are to be filtered and the filter coefficients based on e.g. RDO and this information is signalled to the decoder.

[0100] 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 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 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 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.

[0101] 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.

[0102] 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.

[0103] 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:

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

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: TemporalId = 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.

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.

[0124] 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.

[0125] 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 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.

[0126] 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.

[0127] 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.

[0128] 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.

[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 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 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 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 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 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.

[0133] 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.

[0134] 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.

[0135] 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.

[0136] 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.

[0137] 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.

[0138] 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.

[0139] In HEVC, a reference picture set (RPS) syntax structure and decoding process are 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), 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 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 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 subset of the reference picture set. RefPicSetLtCurr and RefPicSetLtFoll are collectively referred to as the long-term subset of the reference picture set.

[0140] 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.

[0141] 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.

[0142] 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.

[0143] 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.

[0144] 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 is 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.

[0145] 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.

[0146] 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 bit stream. 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.

[0147] 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.

[0148] 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.
- 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.
- 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 layer may have UHDTV (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 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 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 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.

[0149] 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.

[0150] 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.

[0151] 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.

[0152] 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 autostrereoscopic 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.

[0153] 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 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.

[0154] 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 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.

[0155] 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 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 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.

[0156] 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 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 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.

[0157] 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 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. 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.

[0158] 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 contain constituent frames of a stereo pair.

[0159] 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 downsampling should be performed accordingly.

[0160] A coding tool or mode called intra block copy (IBC) is similar to inter prediction but uses the current picture being encoded or decoded as a reference picture. Obviously, only the blocks coded or decoded before the current block being coded or decoded can be used as references for the prediction. The screen content coding (SCC) extension of HEVC includes IBC.

[0161] The motion vector prediction of H.265/HEVC is described below as an example of a system or method where embodiments may be applied.

[0162] 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. 5a and 5b. X stands for the current prediction unit. A_0, A_1, B_0, B_1, B_2 in Figure 5a are spatial candidates while C_0, C_1 in Figure 5b are temporal candidates. The block comprising or corresponding to the candidate C_0 or C_1 in Figure 5b, whichever is the source for the temporal candidate, may be referred to as the collocated block.

[0163] 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.

[0164] 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.

[0165] 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).

[0166] 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	LTRP	"unavailable"
LTRP	STRP	"unavailable"
LTRP	LTRP	"available" (but not scaled)

[0167] 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 5b) 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 5b) 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.

[0168] 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;

– 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

– 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 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);

– 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 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.

[0169] 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.

[0170] 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 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.

[0171] Different spatial granularity or units may be applied to represent and/or store a motion field. For example, a regular grid of spatial units may be used. For example, a picture may be divided into rectangular blocks of certain size (with the possible exception of blocks at the edges of the picture, such as on the right edge and the bottom edge). For example, the size of the spatial unit may be equal to the smallest size for which a distinct motion can be indicated by the encoder in the bitstream, such as a 4x4 block in luma sample units. For example, a so-called compressed motion field may be used, where the spatial unit may be

equal to a pre-defined or indicated size, such as a 16x16 block in luma sample units, which size may be greater than the smallest size for indicating distinct motion. For example, an HEVC encoder and/or decoder may be implemented in a manner that a motion data storage reduction (MDSR) or motion field compression is performed for each decoded motion field (prior to using the motion field for any prediction between pictures). In an HEVC implementation, MDSR may reduce the granularity of motion data to 16x16 blocks in luma sample units by keeping the motion applicable to the top-left sample of the 16x16 block in the compressed motion field. The encoder may encode indication(s) related to the spatial unit of the compressed motion field as one or more syntax elements and/or syntax element values for example in a sequence-level syntax structure, such as a video parameter set or a sequence parameter set. In some (de)coding methods and/or devices, a motion field may be represented and/or stored according to the block partitioning of the motion prediction (e.g. according to prediction units of the HEVC standard). In some (de)coding methods and/or devices, a combination of a regular grid and block partitioning may be applied so that motion associated with partitions greater than a pre-defined or indicated spatial unit size is represented and/or stored associated with those partitions, whereas motion associated with partitions smaller than or unaligned with a pre-defined or indicated spatial unit size or grid is represented and/or stored for the pre-defined or indicated units.

[0172] The scalable video coding extension (SHVC)of HEVC provides a mechanism for offering spatial, bit-depth, color gamut, and quality scalability while exploiting the inter-layer redundancy. The multiview extension (MV-HEVC) of HEVC enables coding of multiview video data suitable e.g. for stereoscopic displays.

[0173] SHVC and MV-HEVC enable two types of inter-layer prediction, namely sample and motion prediction. In the inter-layer sample prediction, the inter-layer reference (ILR) picture is used to obtain the sample values of a prediction block. In MV-HEVC, the decoded base-layer picture acts, without modifications, as an ILR picture. In spatial and color gamut scalability of SHVC, inter-layer processing, such as resampling, is applied to the decoded base-layer picture to obtain an ILR picture. In the resampling process of SHVC, the base-layer picture may be cropped, upsampled and/or padded to obtain an ILR picture. The relative position of the upsampled base-layer picture to the enhancement layer picture is indicated through so-called reference layer location offsets. This feature enables region-of-interest (ROI) scalability, in which only subset of the picture area of the base layer is enhanced in an enhancement layer picture.

[0174] Inter-layer motion prediction exploits the temporal motion vector prediction (TMVP) mechanism of HEVC. The ILR picture is assigned as the collocated picture for TMVP, and hence is the source of TMVP candidates in the motion vector prediction process. In spatial scalability of SHVC, motion field mapping (MFM) is used to obtain the motion information for the ILR picture from that of the base-layer picture. In MFM, the prediction dependency in base-layer pictures may be considered to be duplicated to generate the reference picture list(s) for ILR pictures. For each block of a remapped motion field with a particular block grid (e.g. a 16x16 grid in HEVC), a collocated sample location in the source picture for inter-layer prediction may be derived. The reference sample location may for example be derived for the center-most sample of the block. In derivation of the reference sample location, the resampling ratio, the reference region, and the resampled reference region can be taken into account – see further below for the definition of the resampling ratio, the reference region, and the resampled reference region. Moreover, the reference sample location may be rounded or truncated to be aligned with the block grid of the motion field of the reference region or the source picture for inter-layer prediction. The motion vector corresponding to the reference sample location is obtained from the motion field of the reference region or the source picture for inter-layer prediction. This motion vector is re-scaled according to the resampling ratio and then included in the remapped motion field. The remapped motion field can be subsequently used as a source for TMVP or alike. In contrast, MFM is not applied in MV-HEVC for base-view pictures to be referenced during the inter-layer motion prediction process.

[0175] The spatial correspondence of a reference-layer picture and an enhancement-layer picture may be inferred or may be indicated with one or more types of so-called reference layer location offsets. In HEVC, reference layer location offsets may be included in the PPS by the encoder and decoded from the PPS by the decoder. Reference layer location offsets may be used for but are not limited to achieving region-of-interest (ROI) scalability. Reference layer location offsets may be indicated between two layers or pictures of two layers even if the layers do not have an inter-layer prediction relation between each other. Reference layer location offsets may comprise one or more of scaled reference layer offsets, reference region offsets, and resampling phase sets. Scaled reference layer offsets may be considered to specify the horizontal and vertical offsets between the sample in the current picture that is collocated with the top-left luma sample of the reference region in a decoded picture in a reference layer and the horizontal and vertical offsets between the sample in the current picture that is collocated with the bottom-right luma sample of the reference region in a

decoded picture in a reference layer. Another way is to consider scaled reference layer offsets to specify the positions of the corner samples of the upsampled reference region (or more generally, the resampled reference region) relative to the respective corner samples of the enhancement layer picture. The scaled reference layer offsets can be considered to specify the spatial correspondence of the current layer picture (for which the reference layer location offsets are indicated) relative to the scaled reference region of the scaled reference layer picture. The scaled reference layer offset values may be signed and are generally allowed to be equal to 0. When scaled reference layer offsets are negative, the picture for which the reference layer location offsets are indicated corresponds to a cropped area of the reference layer picture. Reference region offsets may be considered to specify the horizontal and vertical offsets between the top-left luma sample of the reference region in the decoded picture in a reference layer and the top-left luma sample of the same decoded picture as well as the horizontal and vertical offsets between the bottom-right luma sample of the reference region in the decoded picture in a reference layer and the bottom-right luma sample of the same decoded picture. The reference region offsets can be considered to specify the spatial correspondence of the reference region in the reference layer picture relative to the decoded reference layer picture. The reference region offset values may be signed and are generally allowed to be equal to 0. When reference region offsets are negative, the reference layer picture corresponds to a cropped area of the picture for which the reference layer location offsets are indicated. A resampling phase set may be considered to specify the phase offsets used in resampling process of a source picture for inter-layer prediction. Different phase offsets may be provided for luma and chroma components.

[0176] Scalability may be enabled in two basic ways. Either by introducing new coding modes for performing prediction of pixel values or syntax from lower layers of the scalable representation or by placing the lower layer pictures to a reference picture buffer (e.g. a decoded picture buffer, DPB) of the higher layer. The first approach may be more flexible and thus may provide better coding efficiency in most cases. However, the second, reference frame based scalability, approach may be implemented efficiently with minimal changes to single layer codecs while still achieving majority of the coding efficiency gains available. Essentially a reference frame based scalability codec may be implemented by utilizing the same hardware or software implementation for all the layers, just taking care of the DPB management by external means.

[0177] A scalable video encoder for quality scalability (also known as Signal-to-Noise or SNR) and/or spatial scalability may be implemented as follows. For a base layer, a

conventional non-scalable video encoder and decoder may be used. The reconstructed/decoded pictures of the base layer are included in the reference picture buffer and/or reference picture lists for an enhancement layer. In case of spatial scalability, the reconstructed/decoded base-layer picture may be upsampled prior to its insertion into the reference picture lists for an enhancement-layer picture. The base layer decoded pictures may be inserted into a reference picture list(s) for coding/decoding of an enhancement layer picture similarly to the decoded reference pictures of the enhancement layer. Consequently, the encoder may choose a base-layer reference picture as an inter prediction reference and indicate its use with a reference picture index in the coded bitstream. The decoder decodes from the bitstream, for example from a reference picture index, that a base-layer picture is used as an inter prediction reference for the enhancement layer. When a decoded base-layer picture is used as the prediction reference for an enhancement layer, it is referred to as an inter-layer reference picture.

[0178] While the previous paragraph described a scalable video codec with two scalability layers with an enhancement layer and a base layer, it needs to be understood that the description can be generalized to any two layers in a scalability hierarchy with more than two layers. In this case, a second enhancement layer may depend on a first enhancement layer in encoding and/or decoding processes, and the first enhancement layer may therefore be regarded as the base layer for the encoding and/or decoding of the second enhancement layer. Furthermore, it needs to be understood that there may be inter-layer reference pictures from more than one layer in a reference picture buffer or reference picture lists of an enhancement layer, and each of these inter-layer reference pictures may be considered to reside in a base layer or a reference layer for the enhancement layer being encoded and/or decoded. Furthermore, it needs to be understood that other types of inter-layer processing than reference-layer picture upsampling may take place instead or additionally. For example, the bit-depth of the samples of the reference-layer picture may be converted to the bit-depth of the enhancement layer and/or the sample values may undergo a mapping from the color space of the reference layer to the color space of the enhancement layer.

[0179] A scalable video coding and/or decoding scheme may use multi-loop coding and/or decoding, which may be characterized as follows. In the encoding/decoding, a base layer picture may be reconstructed/decoded to be used as a motion-compensation reference picture for subsequent pictures, in coding/decoding order, within the same layer or as a reference for inter-layer (or inter-view or inter-component) prediction. The reconstructed/decoded base layer picture may be stored in the DPB. An enhancement layer picture may likewise be

reconstructed/decoded to be used as a motion-compensation reference picture for subsequent pictures, in coding/decoding order, within the same layer or as reference for inter-layer (or inter-view or inter-component) prediction for higher enhancement layers, if any. In addition to reconstructed/decoded sample values, syntax element values of the base/reference layer or variables derived from the syntax element values of the base/reference layer may be used in the inter-layer/inter-component/inter-view prediction.

[0180] Inter-layer prediction may be defined as prediction in a manner that is dependent on data elements (e.g., sample values or motion vectors) of reference pictures from a different layer than the layer of the current picture (being encoded or decoded). Many types of inter-layer prediction exist and may be applied in a scalable video encoder/decoder. The available types of inter-layer prediction may for example depend on the coding profile according to which the bitstream or a particular layer within the bitstream is being encoded or, when decoding, the coding profile that the bitstream or a particular layer within the bitstream is indicated to conform to. Alternatively or additionally, the available types of inter-layer prediction may depend on the types of scalability or the type of an scalable codec or video coding standard amendment (e.g. SHVC, MV-HEVC, or 3D-HEVC) being used.

[0181] The types of prediction may comprise, but are not limited to, one or more of the following: sample prediction and motion prediction. In sample prediction, at least a subset of the reconstructed sample values of a reference picture are used for predicting sample values of the current picture. Sample prediction may be motion-compensated and/or disparity-compensated e.g. through the use of motion vectors. Sample prediction and inter prediction may sometimes be used interchangeably, while inter prediction may also be understood more generally to cover also types of inter-picture prediction in addition to sample prediction. In motion prediction, at least a subset of the motion vectors of a source picture for motion prediction (a.k.a. collocated pictures) are used as a reference for predicting motion vectors of the current picture. Typically, predicting information on which reference pictures are associated with the motion vectors is also included in motion prediction.

[0182] The types of inter-layer prediction may comprise, but are not limited to, one or more of the following: inter-layer sample prediction, inter-layer motion prediction, inter-layer residual prediction. In inter-layer sample prediction, at least a subset of the reconstructed sample values of a source picture for inter-layer prediction are used as a reference for predicting sample values of the current picture. In inter-layer motion prediction, at least a subset of the motion vectors of a source picture for inter-layer prediction are used as a reference for predicting motion vectors of the current picture. Typically, predicting

information on which reference pictures are associated with the motion vectors is also included in inter-layer motion prediction. For example, the reference indices of reference pictures for the motion vectors may be inter-layer predicted and/or the picture order count or any other identification of a reference picture may be inter-layer predicted. In some cases, inter-layer motion prediction may also comprise prediction of block coding mode, header information, block partitioning, and/or other similar parameters. In some cases, coding parameter prediction, such as inter-layer prediction of block partitioning, may be regarded as another type of inter-layer prediction. In inter-layer residual prediction, the prediction error or residual of selected blocks of a source picture for inter-layer prediction is used for predicting the current picture. In multiview-plus-depth coding, such as 3D-HEVC, cross-component inter-layer prediction may be applied, in which a picture of a first type, such as a depth picture, may affect the inter-layer prediction of a picture of a second type, such as a conventional texture picture. For example, disparity-compensated inter-layer sample value and/or motion prediction may be applied, where the disparity may be at least partially derived from a depth picture.

[0183] A direct reference layer may be defined as a layer that may be used for inter-layer prediction of another layer for which the layer is the direct reference layer. A direct predicted layer may be defined as a layer for which another layer is a direct reference layer. An indirect reference layer may be defined as a layer that is not a direct reference layer of a second layer but is a direct reference layer of a third layer that is a direct reference layer or indirect reference layer of a direct reference layer of the second layer for which the layer is the indirect reference layer. An indirect predicted layer may be defined as a layer for which another layer is an indirect reference layer. An independent layer may be defined as a layer that does not have direct reference layers. In other words, an independent layer is not predicted using inter-layer prediction. A non-base layer may be defined as any other layer than the base layer, and the base layer may be defined as the lowest layer in the bitstream. An independent non-base layer may be defined as a layer that is both an independent layer and a non-base layer.

[0184] A source picture for inter-layer prediction may be defined as a decoded picture that either is, or is used in deriving, an inter-layer reference picture that may be used as a reference picture for prediction of the current picture. In multi-layer HEVC extensions, an inter-layer reference picture is included in an inter-layer reference picture set of the current picture. An inter-layer reference picture may be defined as a reference picture that may be used for inter-

layer prediction of the current picture. In the coding and/or decoding process, the inter-layer reference pictures may be treated as long term reference pictures.

[0185] A source picture for inter-layer prediction may be required to be in the same access unit as the current picture. In some cases, e.g. when no resampling, motion field mapping or other inter-layer processing is needed, the source picture for inter-layer prediction and the respective inter-layer reference picture may be identical. In some cases, e.g. when resampling is needed to match the sampling grid of the reference layer to the sampling grid of the layer of the current picture (being encoded or decoded), inter-layer processing is applied to derive an inter-layer reference picture from the source picture for inter-layer prediction. Examples of such inter-layer processing are described in the next paragraphs.

[0186] Inter-layer sample prediction may be comprise resampling of the sample array(s) of the source picture for inter-layer prediction. The encoder and/or the decoder may derive a horizontal scale factor (e.g. stored in variable ScaleFactorX) and a vertical scale factor (e.g. stored in variable ScaleFactorY) for a pair of an enhancement layer and its reference layer for example based on the reference layer location offsets for the pair. If either or both scale factors are not equal to 1, the source picture for inter-layer prediction may be resampled to generate an inter-layer reference picture for predicting the enhancement layer picture. The process and/or the filter used for resampling may be pre-defined for example in a coding standard and/or indicated by the encoder in the bitstream (e.g. as an index among pre-defined resampling processes or filters) and/or decoded by the decoder from the bitstream. A different resampling process may be indicated by the encoder and/or decoded by the decoder and/or inferred by the encoder and/or the decoder depending on the values of the scale factor. For example, when both scale factors are less than 1, a pre-defined downsampling process may be inferred; and when both scale factors are greater than 1, a pre-defined upsampling process may be inferred. Additionally or alternatively, a different resampling process may be indicated by the encoder and/or decoded by the decoder and/or inferred by the encoder and/or the decoder depending on which sample array is processed. For example, a first resampling process may be inferred to be used for luma sample arrays and a second resampling process may be inferred to be used for chroma sample arrays.

[0187] SHVC enables the use of weighted prediction or a color-mapping process based on a 3D lookup table (LUT) for (but not limited to) color gamut scalability. The 3D LUT approach may be described as follows. The sample value range of each color components may be first split into two ranges, forming up to 2x2x2 octants, and then the luma ranges can be further split up to four parts, resulting into up to 8x2x2 octants. Within each octant, a cross

color component linear model is applied to perform color mapping. For each octant, four vertices are encoded into and/or decoded from the bitstream to represent a linear model within the octant. The color-mapping table is encoded into and/or decoded from the bitstream separately for each color component. Color mapping may be considered to involve three steps: First, the octant to which a given reference-layer sample triplet (Y, Cb, Cr) belongs is determined. Second, the sample locations of luma and chroma may be aligned through applying a color component adjustment process. Third, the linear mapping specified for the determined octant is applied. The mapping may have cross-component nature, i.e. an input value of one color component may affect the mapped value of another color component. Additionally, if inter-layer resampling is also required, the input to the resampling process is the picture that has been color-mapped. The color-mapping may (but needs not to) map samples of a first bit-depth to samples of another bit-depth.

[0188] Inter-layer motion prediction may be realized as follows. A temporal motion vector prediction process, such as TMVP of H.265/HEVC, may be used to exploit the redundancy of motion data between different layers. This may be done as follows: when the decoded base-layer picture is upsampled, the motion data of the base-layer picture is also mapped to the resolution of an enhancement layer. If the enhancement layer picture utilizes motion vector prediction from the base layer picture e.g. with a temporal motion vector prediction mechanism such as TMVP of H.265/HEVC, the corresponding motion vector predictor is originated from the mapped base-layer motion field. This way the correlation between the motion data of different layers may be exploited to improve the coding efficiency of a scalable video coder. In SHVC and/or alike, inter-layer motion prediction may be performed by setting the inter-layer reference picture as the collocated reference picture for TMVP derivation.

[0189] Similarly to MVC, in MV-HEVC, inter-view reference pictures can be included in the reference picture list(s) of the current picture being coded or decoded. SHVC uses multi-loop decoding operation (unlike the SVC extension of H.264/AVC). SHVC may be considered to use a reference index based approach, i.e. an inter-layer reference picture can be included in a one or more reference picture lists of the current picture being coded or decoded (as described above).

[0190] For the enhancement layer coding, the concepts and coding tools of HEVC base layer may be used in SHVC, MV-HEVC, and/or alike. However, the additional inter-layer prediction tools, which employ already coded data (including reconstructed picture samples

and motion parameters a.k.a motion information) in reference layer for efficiently coding an enhancement layer, may be integrated to SHVC, MV-HEVC, and/or alike codec.

[0191] A sender, a gateway, a client, or alike may select the transmitted layers and/or sub-layers of a scalable video bitstream. Terms layer extraction, extraction of layers, or layer down-switching may refer to transmitting fewer layers than what is available in the bitstream received by the sender, gateway, client, or alike. Layer up-switching may refer to transmitting additional layer(s) compared to those transmitted prior to the layer up-switching by the sender, gateway, client, or alike, i.e. restarting the transmission of one or more layers whose transmission was ceased earlier in layer down-switching. Similarly to layer down-switching and/or up-switching, the sender, gateway, client, or alike may perform down- and/or up-switching of temporal sub-layers. The sender, gateway client, or alike may also perform both layer and sub-layer down-switching and/or up-switching. Layer and sub-layer down-switching and/or up-switching may be carried out in the same access unit or alike (i.e. virtually simultaneously) or may be carried out in different access units or alike (i.e. virtually at distinct times).

[0192] 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 only a part of the video is displayed at a single point of time in typical displaying arrangements. For example, VR video may be viewed on a head-mounted display (HMD) that may be capable of displaying e.g. about 100-degree field of view. The spatial subset of the VR video content to be displayed may be selected based on the orientation of the HMD. In another example, a typical flat-panel viewing environment is assumed, wherein e.g. up to 40-degree field-of-view may be displayed. When displaying wide-FOV content (e.g. fisheye) on such a display, it may be preferred to display a spatial subset rather than the entire picture.

[0193] 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 multiple lenses and sensors. The acquisition results in a set of digital image/video signals. The cameras/lenses typically 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 VR frame. The breakdown of image stitching, projection, and mapping process is illustrated with Figure 6a and described as follows. Input images are stitched and projected 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 three-dimensional structure consisting of

one or more surface(s) on which the captured VR image/video content is 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. 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 for example an equirectangular panorama and a cube map representation format.

[0194] Region-wise mapping may be applied to map projected frame onto one or more packed VR frames. In some cases, 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 resampling) to the regions, and placing the transformed regions in spatially non-overlapping areas, a.k.a. constituent frame partitions, within the packed VR frame. If the region-wise mapping is not applied, the packed VR frame is identical to the projected frame. Otherwise, regions of the projected frame are mapped onto a packed VR frame by indicating the location, shape, and size of each region in the packed VR frame. The term mapping may be defined as a process by which a projected frame is mapped to a packed VR frame. The term packed VR frame may be defined as a frame that results from a mapping of a projected frame. In practice, the input images may be converted to a packed VR frame in one process without intermediate steps.

[0195] 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 can be mapped to a bounding cylinder that can be cut vertically to form a 2D picture (this type of projection is known as equirectangular projection). The process of forming a monoscopic equirectangular panorama picture is illustrated in Figure 6b. A set of input images, such as fisheye images of a camera array or a camera device with multiple lenses and sensors, is stitched onto a spherical image. The spherical image is further projected onto a cylinder (without the top and bottom faces). The cylinder is unfolded to form a two-dimensional projected frame. In practice one or more of the presented steps may be merged; for example, the input images may be directly projected onto a cylinder without an intermediate projection onto a sphere. The projection structure for equirectangular panorama may be considered to be a cylinder that comprises a single surface.

[0196] In general, 360-degree content can be mapped onto different types of solid 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

[0197] 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 panoramic 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 panoramic projection format.

[0198] A panorama, such as an equirectangular panorama, can be stereoscopic. In a stereoscopic panorama format, one panorama picture may represent the left view and the other parorama picture (of the same time instant or access unit) may represent the right view. When a stereoscopic panorama is displayed on a stereoscopic display arrangement, such as a virtual reality headset, the left-view panorama may be displayed in appropriate viewing angle and field of view to the left eye, and the right-view panorama may be similarly displayed to the right eye. In a stereoscopic panorama, the stereoscopic viewing may be assumed to happen towards the equator (i.e. vertically the center-most pixel row) of the panorama, causing that greater the absolute inclination of the viewing angle, the worse the correctness of the stereoscopic three-dimensional presentation.

[0199] A family of pseudo-cylindrical projections attempts to minimize the distortion of the polar regions of the cylindrical projections, such as the equirectangular projection, by bending the meridians toward the center of the map as a function of longitude while maintaining the cylindrical characteristic of parallel parallels. Pseudo-cylindrical projections result into non-rectangular contiguous 2D images representing the projected sphere. However, it is possible to present pseudo-cylindrical projections in interrupted forms that are made by joining several regions with appropriate central meridians and false easting and clipping boundaries. Pseudo-cylindrical projections may be categorized based upon the shape of the meridians to sinusoidal, elliptical, parabolic, hyperbolic, rectilinear and miscellaneous pseudo-cylindrical projections. An additional characterization is based upon whether the meridians come to a point at the pole or are terminated along a straight line (in which case the projection represents less than 180 degrees vertically).

[0200] Figure 7a shows an example of a 360-degree video/image projection onto a cube, i.e. a monoscopic cubemap projection. While example are presented below for monoscopic

cubemaps, it is noted that a cubemap can be stereoscopic, which can be reached e.g. by re-projecting each view of a stereoscopic panorama to the cubemap format. The cubemap may be generated, for example, by first rendering the spherical scene six times from a viewpoint, with the views defined by a 90 degree view frustum representing each cube face. When the cube is unfolded, it can be represented as a 2D image. The cube sides may be packed into the same frame or each cube side may be treated individually (in encoding, for example). The cube faces may be arranged on a frame in a plurality of ways, and it is noted that the embodiments described below and the underlying technical problems apply generally when a frame consists of two or more constituent frame partitions belonging to different planes or surfaces. Figures 7b and 7c illustrate two examples, a 4x3 mapping in Figure 7b and a 3x2 mapping in Figure 7c, to unfold a cube onto a frame. The examples of Figures 7b and 7c demonstrate clearly at least two problems underlying the known methods for packing the cube faces onto a frame:

[0201] 1: Suboptimal number of intra prediction references

[0202] Regardless of which packing of cube faces onto the frame is used, some of the boundary block rows or columns of cube faces lack a reference row or column of samples from an adjacent cube face for intra prediction.

[0203] In Figure 7b (4x3 cubemap), the reference samples for intra prediction are missing in the following cases:

- The left cube face has no reference samples for predicting from above. However, the left boundary sample row of the top cube face, which precedes in (de)coding order the left cube face, could be used as the reference sample row for predicting the top block row of the left cube face.
- The right cube face has no reference samples for predicting from above. However, the right boundary sample row of the top cube face could be used as the reference sample row for predicting the top block row of the right cube face.
- The back cube face has no reference samples for predicting from above. However, the horizontally mirrored top sample row of the top cube face could be used as the reference sample row for predicting the top block row of the back cube face.
- The bottom cube face has no reference samples from predicting from left. However, the bottom sample row of the left cube face could be used as the reference sample column for predicting the left-most block column of the bottom cube face.

[0204] In Figure 7c (3x2 cubemap), the reference samples for intra prediction are missing in the following cases:

- The top cube face lacks a correct prediction reference from above.
- The top cube face has no reference samples for predicting from left. However, the top sample row of the left cube face could be used as the reference sample column for predicting the left-most block column of the top cube face.
- The bottom cube face lacks a correct prediction reference in the left side. However, the bottom sample row of the left cube face could be used as the reference sample column for predicting the left-most block column of the bottom cube face.
- The back cube face lacks a correct prediction reference in the left side. However, the right-most sample column of the right cube face could be used as the reference sample column for predicting the left-most block column of the back cube face.
- The back cube face lacks a correct prediction reference for predicting from above. However, the horizontally mirrored top sample row of the top cube face could be used as the reference sample row for predicting the top block row of the back cube face.

[0205] As noticed above, there are discontinuities over some cube face boundaries. When in-loop deblocking filtering is performed across such cube face boundaries, undesirable "leaking" of sample information from one cube face to another happens when these cube faces are actually not adjacent. It is difficult, maybe impossible, to avoid the introduction of such error, since the deblocking filtering is performed after reconstructing the block based on the reconstructed prediction error. In HEVC, a possible way to avoid such error propagation is to align the tile grid with the cube side boundaries and turn off deblocking filtering across tile boundaries. However, tile boundaries also have the impact that no intra prediction takes place from another tile; hence, this approach turns off all intra prediction from one cube face to another even if the cube sides are adjacent in the 3D domain.

[0206] 2: Motion vectors over constituent frame partition boundaries

[0207] When constituent frame partitions, such as cube faces, are packed next to each other motion vectors of blocks close to the boundary between constituent frame partitions are often sub-optimal. Motion vectors pointing outside the boundaries of the constituent frame partition (to another constituent frame partition) or causing sub-pixel interpolation using samples outside the boundaries of the constituent frame partition (within another constituent frame partition) are sub-optimally handled. Such motion vectors would cause prediction blocks in which the content originates from two (or more) constituent frame partitions and are hence typically discontinuous. Alternatively, such motion vectors may be avoided by the encoder, but this has a negative impact on rate-distortion performance when compared to the conventional handling of motion vectors over picture boundaries by causing padding with

boundary samples or, equivalently, saturation of the used sample locations to be within picture boundaries.

[0208] Now in order to at least alleviate the above disadvantages as well as to bring about other benefits as described below, a method for projecting reference samples across cube face boundaries is presented hereinafter.

[0209] In the method, which is disclosed in Figure 8a as a flow diagram in accordance with an embodiment, a first region of the plurality of regions of a first picture is encoded (850), wherein said first region is a projected representation of a first surface and said encoding comprises reconstructing a first reconstructed region corresponding to the first region. At least a first block of the first picture is encoded with a coding mode causing at least a part of the first reconstructed region to be projected (852) onto a second surface and further to a reconstructed first block. Said encoding comprises reconstructing (854) the reconstructed first block. At least a part of the reconstructed first block forms a projected reference signal. A second region of the plurality of regions of the first picture is encoded (856), where said second region is a projected representation of the second surface and said encoding comprises using the projected reference signal as a reference for prediction.

[0210] Correspondingly, a decoding method, which is disclosed in Figure 8 as a flow diagram in accordance with an embodiment, comprises decoding (860) a first coded region of a plurality of regions of a first coded picture into the first reconstructed region, wherein said first reconstructed region is a projected representation of a first surface. At least a first coded block of the first coded picture is decoded, the first coded block having a coding mode causing at least a part of the first reconstructed region to be projected (862) onto a second surface and further to a reconstructed first block. Said decoding comprises reconstructing (864) the reconstructed first block. At least a part of the reconstructed first block form a projected reference signal. A second coded region of the plurality of regions of the first picture is decoded (866) into a second reconstructed region, where said second reconstructed region is a projected representation of the second surface and said decoding comprises using the projected reference signal as a reference for prediction.

[0211] According to an embodiment, said coding mode is indicative of one or more of the following:

- the first reconstructed region or the at least a part of the first reconstructed region;
- the projection and/or transformation to be applied to the at least a part of the first reconstructed region;
- the first surface;

- the second surface.

[0212] According to an embodiment, said coding mode is applied in reconstruction or in decoding in conventional order of processing blocks, such as in raster scan order of LCUs within tiles or within a picture, if tiles are not in use, as in HEVC. In an embodiment, said coding mode is applied in reconstruction or in decoding after reconstructing a picture otherwise. In an embodiment, an encoder specifies with a first coding mode or a parameter of the coding mode that the (first) coding mode is applied in reconstruction or in decoding in conventional order of processing blocks and specifies with a second coding mode or a parameter of the coding mode that the (second) coding mode is applied after reconstructing the picture otherwise. In an embodiment, the encoder and/or the decoder infer e.g. based on the information that the coding mode is indicative of whether the coding mode is applied in conventional order of processing blocks or after reconstructing the picture otherwise. For example, if the coding mode is indicative that the first reconstructed region is used to reconstruct the reconstructed first block but the first reconstructed region has not been formed (i.e. follows in decoding order the first block), it may be inferred that the reconstructed first block is formed after reconstructing the picture otherwise. In another example, if the coding mode is indicative that the first reconstructed region is used to reconstruct the reconstructed first block and the first reconstructed region has been formed (i.e. precedes in decoding order the first block), it may be inferred that the reconstructed first block is formed in conventional order of processing blocks.

[0213] According to an embodiment, a projected frame is obtained (e.g. as a result of stitching). A first region of the projected frame and a second region of the projected frame are mapped onto a packed VR frame, wherein said first region is a projected representation of a first surface and said second region is a projected representation of a second surface. The mapping is performed in a manner that at least a first block of a packed VR frame is spatially adjacent to the second region, the first block being neither a part of the first region nor the second region. For example, a first cube face may be mapped on a first vertical location in the packed VR frame, and a second cube face may be likewise mapped on the first vertical location in the packed VR frame, and the first and second cube faces may be mapped to horizontal locations next to each other, but separated by a column of blocks comprising the first block. The performed mapping and/or the location of the at least first block may be indicated to an encoder. In response to receiving such indication, the encoder may choose a coding mode for the at least first block that causes at least a part of the first reconstructed region to be projected onto the second surface and further to a reconstructed first block.

[0214] According to an embodiment, a packed VR frame is obtained, e.g. as a result of decoding, and information of the mapping applied in the packed VR frame is also obtained. The information indicates that the packed VR frame comprises a first region, a second region, and at least a first block that is spatially adjacent to the second region, wherein said first region is a projected representation of a first surface and said second region is a projected representation of a second surface. The first region and the second region are extracted from the packed VR frame. Geometric transformation, such as rotation, mirroring, and/or resampling, may be applied to the first region and the second region, e.g. based on the information of the mapping. The (potentially transformed) first region and the second region are located into a projected frame. The projected frame may have a representation format that is one of a pre-defined set of representation formats of the projected frame, including for example an equirectangular panorama and a cube map representation format.

[0215] The effects of the above methods for improving intra prediction may be illustrated with reference to an exemplified 3x2 monoscopic cubemap in Figure 9, where the reconstructed picture has three additional block columns 900, 902 904 and one additional block row 906. In general, the number of additional block rows and columns may be different than in this example. For example, the left-most additional block column 900 may be absent. Additionally or alternatively, the additional block row 906 may be absent.

[0216] For example, for the additional block column 902, the upper part of the block column marked with "Left projected to the Front plane (L2F)" is generated by projecting (a part of) the left cube face on to the plane of the front cube face in a manner that the block column and the front cube face form a continuous image plane. Each block in the block column marked with "Left projected to the Front plane (L2F)" is formed prior to the using samples of the block as a reference for intra prediction of the front face. The lower part of the block column marked with "Left projected to the Bottom plane (L2Bo)" is generated by projecting (a part of) the left cube face on to the plane of the bottom cube face in a manner that the block column and the bottom cube face form a continuous image plane.

[0217] The parts of the other additional block columns and rows, i.e. "Front projected to the Right plane (F2R)", "Top projected to the Back plane (T2Ba)", "Front projected to the Top plane (F2T)", "Left projected to the Top plane (L2T)", "Front projected to the Bottom plane (F2Bo)" and "Left projected to the Back plane (L2Ba)"are formed similarly as indicated in Figure 9. It is noted that in the example of Figure 9, the top face and the back face are rotated in order to enable generating the reference samples for them, when the picture is (de)coded in raster scan order of block rows.

[0218] Thanks to the presented method, the availability of reference samples (for intra prediction) across cube face boundaries is optimized and hence compression efficiency improved. The above-illustrated arrangement also makes motion vectors crossing left and top boundaries of cube faces efficient in terms of prediction.

[0219] According to an embodiment, the method further comprises surrounding each cube face from all sides by additional block rows and columns that are associated with the cube face. Thereby, inter prediction can be further improved. An example of such arrangement is illustrated in Figure 10.

[0220] The additional block rows and columns (L2F, F2R, L2T, F2Bo, L2Ba, F2T, L2Bo, and T2Ba) shown in Figure 10 may be reconstructed as explained above or after reconstructing the cube faces. These block rows and columns may be used as a reference for intra prediction. The remaining additional block rows and columns are reconstructed similarly by projecting a cube face on the plane of another cube face after reconstructing that another cube face. For example, all the remaining additional block rows and columns may be generated after reconstructing or decoding the cube faces. The arrangement of Figure 10 enables motion vectors causing references to samples outside the cube face in a manner that the resulting prediction blocks are continuous. Hence, the rate-distortion performance is improved.

[0221] According to an embodiment, a constituent frame partition may be considered to comprise a region and one or more blocks that are adjacent to the region, wherein the region is a projected representation of a surface. A reconstructed or decoded constituent frame partition may be considered to comprise a reconstructed region and one or more reconstructed blocks, where the one or more reconstructed blocks are a projected representation of the surface, adjacent to the region on the surface. For example, in Figure 9, the block column 902 and the front face may be considered to form a constituent frame partition.

[0222] According to an embodiment, the plurality of regions in the coded picture corresponds to only a part of the panorama image. For example, one or more regions in the coded picture corresponds may be coded from a tile or a tile set of the panorama image.

[0223] According to an embodiment, the first reconstructed region and the second reconstructed region are different planes or surfaces of the same projection type. For example, the packed VR frame may consist of two or more constituent frame partitions belonging to different planes or surfaces of a polyhedron or other solid geometrical shape used as the projection structure.

[0224] According to an embodiment, the first reconstructed region and the second reconstructed region are of different projection type.

[0225] According to an embodiment, the first reconstructed region is of cylindrical or equirectangular projection type. When the first reconstructed region is an equirectangular panorama, its vertical field of view is less than 180 degrees. According to an embodiment, the second reconstructed region is a top or bottom face of the cylinder or vertically truncated equirectangular panorama, which in 3D corresponds to a truncated sphere.

[0226] The generation of a first region comprising the vertically center part of an equirectangular panorama and a second region comprising the top or the bottom face for the panorama is illustrated in Figures 11a and 11b.

[0227] First, the height of the top, middle, and bottom parts of the panorama are selected, as shown in Figure 11a. The top and bottom part of the panorama are projected on top and bottom planes, i.e. planes that are parallel with the equator plane of the cylinder or panorama. The top and bottom planes may be rectilinear. The top and bottom faces may be formed as the intersection of the top/bottom plane and the cylinder or truncated sphere, as shown in Figure 11b.

[0228] In another embodiment, the second region and/or the second reconstructed region is a top or bottom face of the cylinder or truncated sphere, but aligned with a block grid used in coding (e.g. CTU grid of HEVC).

[0229] Figure 12a illustrates the embodiment for forming the second region as block-aligned. The effective picture area of the top or bottom face of the cylinder or truncated sphere is circular as illustrated in the left side of Figure 12a. Sharp non-block-aligned edges are typically costly to code in terms of rate-distortion performance and may cause visible artefacts within the effective picture area. Hence, in an embodiment, the second region is formed as a block-aligned (e.g. CTU-aligned in HEVC) bounding area covering the top or bottom face of the cylinder or truncated sphere. This is illustrated in the right side of Figure 12a.

[0230] According to an embodiment, at least a part of the first reconstructed region is projected onto the surface of the top or bottom plane. Said part may comprise, for example, the middle part of the cylinder or equirectangular panorama. With this projection, the areas that are not occupied by the top or bottom face are filled in a bounding box or the constituent frame partition. The projected reference signal is illustrated in Figure 12b as the shaded corners of the area.

[0231] When used for intra prediction, and possibly for inter prediction, the first reconstructed region may be projected on the bounding box prior to (de)coding the top or bottom face and may be treated as available samples for intra prediction.

[0232] When used for inter prediction solely, the first reconstructed region may be projected on the bounding box subsequent to (de)coding the top or bottom face but before it is used as a reference for inter prediction.

[0233] A known problem relates to the equirectangular panorama format is that it stretches the nadir and zenith areas. The number of pixels towards the nadir or zenith is proportionally greater compared to that in the equator. This results in unnecessarily large number of pixels being encoded and decoded in the areas close to the nadir and zenith, which in turn increases encoding and decoding complexity and may result into a decreased rate-distortion performance.

[0234] The above-described embodiments that mix equirectangular panorama and rectilinear projections addresses said problem by improving the compression efficiency of the top and bottom faces of cylinders and truncated spheres.

[0235] According to an embodiment, at least a part of the first reconstructed region is projected onto a second surface and additionally resampled to form a projected reference signal. According to an embodiment, the cube faces are of different spatial resolution. Figure 13 shows an example of resampling cube face into different spatial resolutions, where the cube face 1 is resampled to largest spatial resolution, cube faces 2, 3 and 5 to a spatial resolution half of that of cube face 1, and the remaining cube face to even smaller spatial resolutions.

[0236] When additional block rows and columns, i.e. rows and/columns that are outside the region but inside the constituent frame partition, are formed as described above, resampling according to the cube face resolution may be performed.

[0237] It is noted that in a multi-resolution format such as described above, the spatial extents of a region or a constituent frame partition may be small and hence the compression benefit provided by the invention is proportionally greater.

[0238] According to an embodiment, at least a part of the first reconstructed region is projected onto a second surface and additionally a geometric transform is performed to form a projected reference signal. The geometric transform may for example be rotation and/or mirroring.

[0239] Herein, when projecting cube map faces from front-to-left (F2L), and from top-to-left (T2L), there is small square at top right of the front face which should be filled up. This

may be carried out by determining pixel-wise which cube face should be projected (for example, whether to project from left or top face to the surface of the left face). Alternatively, the small square may be filled by stretching or padding, for example by stretching the F2L and T2L pixels.

[0240] According to an embodiment, a first constituent frame partition and a second constituent frame partition are of the same projection type and of the same surface. A part of the first constituent frame partition is resampled and/or geometrically transformed (e.g. rotated or mirrored) and placed on a second constituent frame partition to form a reference signal. The effective picture area of the second constituent frame partition is predicted from the reference signal.

[0241] In other words, in an encoding method, a first constituent frame partition of a plurality of constituent frame partitions of a first picture is encoded, wherein said encoding comprises reconstructing a first reconstructed constituent frame partition corresponding to the first constituent frame partition. At least a part of the first reconstructed constituent frame partition is resampled and/or geometrically transformed onto a second constituent frame partition to form a reference signal. Effective picture area of a second constituent frame partition of the plurality of constituent frame partitions of the first picture is encoded, where said encoding comprises using the reference signal as a reference for prediction.

[0242] According to an embodiment, rather than or in addition to using the reference signal for predicting the effective picture area of the second constituent frame partition of the same picture, the reference signal is used for predicting the effective picture area of the second constituent frame partition in a subsequent picture. The reference signal may be used when a motion vector points outside the effective picture area or when a motion vector causes fractional sample interpolation that uses sample(s) outside the effective picture area as input.

[0243] According to an embodiment, a constituent frame partition is encoded as a motion-constrained tile set.

[0244] According to an embodiment, the constituent frame partitions are partitions of an equirectangular panorama which have undergone a different amount of downsampling prior to encoding. Figure 14a shows an example of an equirectangular panorama divided into three stripes. In this example the top and bottom stripes are horizontally downsampled by a factor of 2 prior to encoding. The resampled top and bottom stripes and the middle stripe may be arranged into constituent frame partitions as illustrated in Figure 14b:

- The resampled top stripe is arranged into a first constituent frame partition, forming its effective picture area. Below this effective picture area (but still within the first

constituent frame partition), there is a block row reserved for reference signal (M2T) that is filled in by the top block row of the reconstructed middle stripe of the same picture, horizontally downsampled by a factor of 2. This reference signal is filled in after reconstructed the top block row of the middle stripe.

- The middle stripe is arrange into a second constituent frame partition, forming its effective picture area. On top of the effective picture area (but still within the second constituent frame partition), there is a block row reserved for reference signal (T2M) that is filled in by the bottom block row of the reconstructed top stripe of the same picture, horizontally upsampled by a factor of 2. This reference signal is filled in after reconstructed the bottom block row of the top stripe. In an embodiment, this reference signal is filled in before reconstructing the middle stripe, while in another embodiment this reference signal is filled in after reconstructing the middle stripe (e.g. after reconstructing all effective picture areas).
- The resampled bottom stripe is mirrored vertically or rotated by 180 degrees and arranged into a third constituent frame partition, forming its effective picture area. Below this effective picture area (but still within the third constituent frame partition), there is a block row reserved for reference signal (M2B) that is filled in by the bottom block row of the reconstructed middle stripe of the same picture, horizontally downsampled by a factor of 2. This reference signal is filled in after reconstructed the bottom block row of the middle stripe.

[0245] It is noted that the above arrangement enables efficient coding of motion vectors over the top boundary of the picture with conventional boundary extension. However, the above arrangement is merely an example embodiment and other embodiments with same technical effects could be similarly realized. For example, additional block column(s) may be located between resampled top and bottom stripes. In another example, an equirectangular panorama may be additionally or alternatively horizontally partitioned and embodiments may be applied to the so-formed constituent frame partitions.

[0246] According to an embodiment, an equirectangular panorama picture is logically partitioned into two constituent frame partitions, i.e. into the left and the right constituent frame partitions as shown in Figure 15a. An additional block row is located above and below the effective picture area. The top block row of the left constituent frame partition is rotated by 180 degrees to form a reference signal above the right constituent frame partition. The top block row of the right constituent frame partition is rotated by 180 degrees to form a reference signal above the left constituent frame partition. The bottom block row of the left constituent

frame partition is rotated by 180 degrees to form a reference signal below the right constituent frame partition.

[0247] According to an embodiment, a similar logical partitioning to the left and right partitions is performed for the top and bottom stripes of an equirectangular panorama picture, as presented above in Figures 14a and 14b. Figure 15b shows an example for arranging the resampled top and bottom stripes and the middle stripe into constituent frame partitions. The additional block row is formed by rotating by 180 degrees the top row of the left part above the right part and vice versa, and likewise rotating by 180 degrees the bottom row of the left part below the right part and vice versa. These additional block rows are marked by "left-right mirroring".

[0248] The above embodiments have been described with reference to regions or constituent frame partitions that are part of the same frame. It needs to be understood that embodiments can be similarly realized when a first layer comprises the first region and a second layer comprises the second region. The projection of projecting at least a part of the first reconstructed region onto a second surface may be performed as an inter-layer prediction process. According to an embodiment, the projected reference signal may form or be a part of an inter-layer reference picture generated from the first region.

[0249] The above embodiments have been described with reference to including the projected reference signal into a reference picture used for prediction. It needs to be understood that the projected reference signal need not be a part of a reference picture in some embodiments. In an embodiment, the projected reference signal forms a prediction signal for inter-layer prediction for a single-loop scalable video codec.

[0250] The above embodiments reduce the encoding bitrate by improving the intra and inter prediction when compared to coding the regions without reprojection.

[0251] Figure 16 shows a block diagram of a video decoder suitable for employing embodiments of the invention. Figure 16 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.

[0252] The video decoder 550 comprises a first decoder section 552 for base view components and a second decoder section 554 for non-base view components. Block 556 illustrates a demultiplexer for delivering information regarding base view components to the first decoder section 552 and for delivering information regarding non-base view components 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 view/non-base view components to obtain the preliminary reconstructed images (I'n). Preliminary reconstructed and filtered base view images may be output 710 from the first decoder section 552 and preliminary reconstructed and filtered base view images may be output 810 from the second decoder section 554.

[0253] Herein, the decoder should 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.

[0254] The entropy decoder 700, 800 perform entropy decoding on the received signal. The entropy decoder 700, 800 thus performs the inverse operation to the entropy encoder of the encoder 330, 430 described above. The entropy decoder 700, 800 outputs the results of the entropy decoding to the prediction error decoder 701, 801 and pixel predictor 704, 804.

[0255] The pixel predictor 704, 804 receives the output of the entropy decoder 700, 800. The output of the entropy decoder 700, 800 may include an indication on the prediction mode used in encoding the current block. A predictor 707, 807 may perform intra or inter prediction as determined by the indication and output a predicted representation of an image block to a first combiner 709, 809. The predicted representation of the image block is used in conjunction with the reconstructed prediction error signal to generate a preliminary reconstructed image. The preliminary reconstructed image may be used in the predictor or may be passed to a regional reference frame processing unit 711, 811. The regionally resampled and/or rearranged reference image may be passed to a filter 708, 808. The filter 708, 808 may apply a filtering which outputs a final reconstructed signal. The final reconstructed signal may be stored in a reference frame memory 706, 806. The reference frame memory 706, 806 may further be connected to the predictor 707, 807 for prediction operations.

[0256] It needs to be understood that the regional reference frame processing unit 711, 811 and the filter 708, 808 may, in some embodiments, be located in opposite order in Figure 13. It also needs to be understood that in some embodiments parts of the filtering performed by the filter 708, 808 may be performed prior to regional reference frame processing 711, 811,

while the remaining parts may be performed after the regional reference frame processing 711, 811. Likewise, some parts of the regional reference frame processing 711, 811 (e.g. resampling) may be performed prior to the filter 708, 808, while the remaining parts of the reference frame processing 711, 811 (e.g. rearranging) may be performed after the filter 708, 808.

[0257] Figure 17 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.

[0258] 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.

[0259] 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.

[0260] 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 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.

[0261] 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 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 stream according to prevailing downlink network conditions.

[0262] The system includes one or more receivers 1560, typically capable of receiving, de-modulating, 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. 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 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 excerption of the recorded stream, is maintained in the recording storage 1570, while any earlier recorded data is discarded from the recording storage 1570.

[0263] 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

[0264] 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.

[0265] 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.

[0266] 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 1580 is multi-tasking and uses computing resources for other purposes than decoding

the 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.

[0267] In the above, some embodiments have been described with reference to resampling and/or rearranging. It needs to be understood that rearranging may comprise relocating, rotating, and/or mirroring even if they are not explicitly mentioned each time. It needs to be understood that the order of operations resampling, relocating, rotating, and mirroring may be pre-defined (e.g. in a coding standard) or may be indicated by an encoder in a bitstream and/or decoded by a decoder from a bitstream. It needs to be understood that more than one operation of the same type (e.g. resampling) may occur in the sequence of operations for the same region.

[0268] In the above, some embodiments have been described with reference to resampling and/or rearranging region(s) for forming reference pictures. It needs to be understood that in addition to or instead of resampling and/or rearranging the sample array(s) of the region(s), the motion fields of the region(s) may be similarly resampled and/or rearranged in various embodiments. The resampled and/or rearranged motion fields may then be used as a source for motion vector prediction, such as TMVP of HEVC or alike. Motion field resampling may be performed similarly as motion field mapping of inter-layer prediction used in spatial scalability, as described earlier. Relocating, rotating, and/or mirroring of motion fields can be performed similarly to the respective operations for sample arrays.

[0269] In the above, some embodiments have been described assuming monoscopic panorama video content. It needs to be understood that embodiments can be applied to stereoscopic content too, where partitions of two constituent frames are packed into the same frame to be coded.

[0270] 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.

[0271] In the above, some embodiments have been described in relation to HEVC and/or terms used in the HEVC specification. It needs to be understood that embodiments similarly

apply to other codecs and coding formats and other terminology with equivalency or similarity to the terms used in the above-described embodiments.

[0272] In the above, where the example embodiments have been described with reference to an encoder, it needs to be understood that the resulting bitstream and the decoder may have corresponding elements in them. Likewise, where the example embodiments have been described with reference to a decoder, it needs to be understood that the encoder may have structure and/or computer program for generating the bitstream to be decoded by the decoder.

[0273] The embodiments of the invention described above describe the codec in terms of separate encoder and decoder apparatus in order to assist the understanding of the processes involved. However, it would be appreciated that the apparatus, structures and operations may be implemented as a single encoder-decoder apparatus/structure/operation. Furthermore, it is possible that the coder and decoder may share some or all common elements.

[0274] In the above, some embodiments have been described with reference to an encoder including indications into a bitstream and/or a decoder decoding indications from a bitstream. It needs to be understood that in addition to or instead of an encoder, another entity may include the indications into the bitstream, such as a sender 1530. It needs to be understood that in addition to or instead of a decoder, another entity may decode the indications from the bitstream, such as a receiver 1550. It needs to be understood that in addition to or instead of a bitstream, the indications may be included into or decoded from a container file and/or a media description, such as the Media Presentation Description (MPD) for Dynamic Adaptive Streaming over HTTP (DASH) or in the Session Description Protocol (SDP) describing an RTP session or stream.

[0275] Although the above examples describe embodiments of the invention operating within a codec within an electronic device, it would be appreciated that the invention as defined in the claims may be implemented as part of any video codec. Thus, for example, embodiments of the invention may be implemented in a video codec which may implement video coding over fixed or wired communication paths.

[0276] Thus, user equipment may comprise a video codec such as those described in embodiments of the invention above. It shall be appreciated that the term user equipment is intended to cover any suitable type of wireless user equipment, such as mobile telephones, portable data processing devices or portable web browsers.

[0277] Furthermore elements of a public land mobile network (PLMN) may also comprise video codecs as described above.

[0278] In general, the various embodiments of the invention may be implemented in hardware or special purpose circuits, software, logic or any combination thereof. For example, some aspects may be implemented in hardware, while other aspects may be implemented in firmware or software which may be executed by a controller, microprocessor or other computing device, although the invention is not limited thereto. While various aspects of the invention may be illustrated and described as block diagrams, flow charts, 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.

[0279] The embodiments of this invention may be implemented by computer software executable by a data processor of the mobile device, such as in the processor entity, or by hardware, or by a combination of software and hardware. Further in this regard it should be noted that any blocks of the logic flow as in the Figures may represent program steps, or interconnected logic circuits, blocks and functions, or a combination of program steps and logic circuits, blocks and functions. The software may be stored on such physical media as memory chips, or memory blocks implemented within the processor, magnetic media such as hard disk or floppy disks, and optical media such as for example DVD and the data variants thereof, CD.

[0280] The memory may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor-based memory devices, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The data processors may be of any type suitable to the local technical environment, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on multi-core processor architecture, as non-limiting examples.

[0281] 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.

[0282] Programs, such as those provided by Synopsys, Inc. of Mountain View, California and Cadence Design, of San Jose, California automatically route conductors and locate components on a semiconductor chip using well established rules of design as well as libraries

of pre-stored design modules. Once the design for a semiconductor circuit has been completed, the resultant design, in a standardized electronic format (e.g., Opus, GDSII, or the like) may be transmitted to a semiconductor fabrication facility or "fab" for fabrication.

[0283] The foregoing description has provided by way of exemplary and non-limiting examples a full and informative description of the exemplary embodiment of this invention. However, various modifications and adaptations may become apparent to those skilled in the relevant arts in view of the foregoing description, when read in conjunction with the accompanying drawings and the appended claims. However, all such and similar modifications of the teachings of this invention will still fall within the scope of the claims.

CLAIMS:

1. A method comprising:

encoding a first region of a first picture comprising a plurality of regions, wherein said first region is a projected representation of a first surface and the encoding comprises reconstructing a first reconstructed region corresponding to said first region;

encoding at least a first block of the first picture with an encoding mode causing at least a part of the first reconstructed region to be projected onto a second surface and further to a reconstructed first block;

reconstructing the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal; and

encoding at least a second region of the plurality of regions of the first picture, wherein said second region is a projected representation of the second surface and said encoding comprises using the projected reference signal as a reference for prediction.

2. A method according to claim 1, wherein said coding mode is indicative of one or more of the following:

- the first reconstructed region or the at least a part of the first reconstructed region;
- the projection and/or transformation to be applied to the at least a part of the first reconstructed region;
- the first surface;
- the second surface.

3. A method according to claim 1 or 2, the method further comprising:

specifying, by an encoder, with a first coding mode or a first parameter value of the coding mode, or inferring, by the encoder, that the first coding mode is applied in reconstruction in conventional order of processing blocks; or

specifying, by an encoder, with a second coding mode or a second parameter value of the coding mode, or inferring, by the encoder, that the second coding mode is applied after reconstructing the picture otherwise.

4. A method according to any preceding claim, the method further comprising

obtaining a projected frame; and

mapping a first region of the projected frame and a second region of the projected frame onto the first picture, wherein said first region of the projected frame is a projected representation of a first surface and said second region of the projected frame is a projected representation of a second surface, wherein said first region of the first picture corresponds to said first region of the projected frame and said second region of the first picture corresponds to said second region of the projected frame, and wherein said at least the first block of the first picture is spatially adjacent to the second region of the first picture, the first block being neither a part of the first region of the first picture nor the second region of the first picture.

5. A method according to claim 4, the method further comprising

indicating performed mapping and/or a location of the at least first block to an encoder; and

in response to receiving the indication in the encoder, choosing a coding mode for the at least first block that causes at least a part of the first reconstructed region to be projected onto the second surface and further to a reconstructed first block.

6. A method according to any preceding claim, wherein the plurality of regions in the coded picture corresponds to only a part of the panorama image.

7. A method according to any preceding claim, wherein the first reconstructed region and the second reconstructed region are of different projection type.

8. A method according to claim 7, wherein the first reconstructed region is of cylindrical or equirectangular projection type, and the second reconstructed region is a top or bottom face of the cylinder or vertically truncated equirectangular panorama, and the second region is formed as a block-aligned bounding area covering the top or bottom face of the cylinder or truncated sphere.

9. A method according to any preceding claim, wherein said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises resampling the first reconstructed region.

10. A method according to any preceding claim, wherein said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises performing a geometric transform.

11. 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 the method of any of claims 1 – 10.

12. A computer readable storage medium stored with code thereon for use by an apparatus, which when executed by a processor, causes the apparatus to perform the method of any of claims 1 – 10.

13. A method comprising:

decoding, from a bitstream, a first encoded region of a plurality of regions of a first coded picture into the first reconstructed region, wherein said first reconstructed region is a projected representation of a first surface;

decoding at least a first coded block of the first coded picture , the first coded block having a coding mode causing at least a part of the first reconstructed region to be projected onto a second surface and further to a reconstructed first block;

reconstructing the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal; and

decoding at least a second coded region of the plurality of regions of the first picture into a second reconstructed region, where said second reconstructed region is a projected representation of the second surface and said decoding comprises using the projected reference signal as a reference for prediction.

14. A method according to claim 13, wherein said coding mode is indicative of one or more of the following:

- the first reconstructed region or the at least a part of the first reconstructed region;
- the projection and/or transformation to be applied to the at least a part of the first reconstructed region;
- the first surface;
- the second surface.

15. A method according to claim 13 or 14, the method further comprising:

receiving, in a decoder, information comprising a first coding mode or a first parameter value of the coding mode or information comprising, and

inferring, by the decoder, that the first coding mode is applied in decoding in conventional order of processing blocks; or

receiving, in the decoder, information comprising a second coding mode or a second parameter value of the coding mode; and

inferring, by the decoder, that the second coding mode is applied after reconstructing the picture otherwise.

16. A method according to any of claims 13 - 15, the method further comprising

obtaining a projected frame; and

mapping a first region of the projected frame and a second region of the projected frame onto the first picture, wherein said first region of the projected frame is a projected representation of a first surface and said second region of the projected frame is a projected representation of a second surface, wherein said first region of the first picture corresponds to said first region of the projected frame and said second region of the first picture corresponds to said second region of the projected frame, and wherein at least the first block of the first picture is spatially adjacent to the second region of the first picture, the first block being neither a part of the first region of the first picture nor the second region of the first picture.

17. A method according to claim 16, the method further comprising

indicating performed mapping and/or a location of the at least first block to the decoder; and

in response to receiving the indication in the decoder, choosing a coding mode for the at least first block that causes at least a part of the first reconstructed region to be projected onto the second surface and further to a reconstructed first block.

18. A method according to any of claims 13 - 17, wherein the plurality of regions in the coded picture corresponds to only a part of the panorama image.

19. A method according to any of claims 13 - 18, wherein the first reconstructed region and the second reconstructed region are of different projection type.

20. A method according to claim 21, wherein the first reconstructed region is of cylindrical or equirectangular projection type, and the second reconstructed region is a top or bottom face of the cylinder or vertically truncated equirectangular panorama, and the second region is formed as a block-aligned bounding area covering the top or bottom face of the cylinder or truncated sphere.

21. A method according to any of claims 13 - 20, wherein said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises resampling the first reconstructed region.

22. A method according to any of claims 13 - 21, wherein said projecting at least a part of the first reconstructed region onto a second surface to form the projected reference signal further comprises performing a geometric transform.

23. 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 the method of any of claims 13 – 22.

24. A computer readable storage medium stored with code thereon for use by an apparatus, which when executed by a processor, causes the apparatus to perform the method of any of claims 13 – 22.

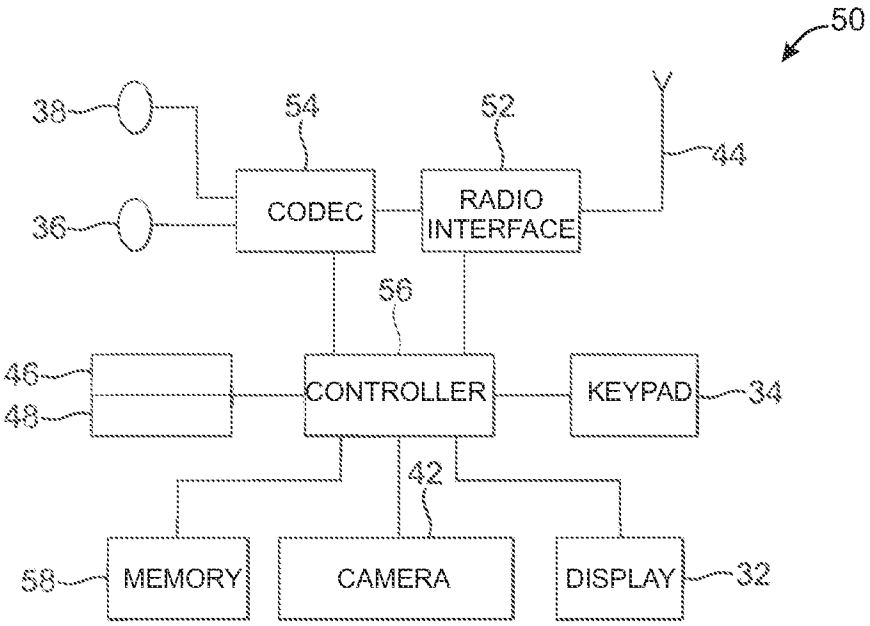

50
38
54
52
44
CODEC
RADIO INTERFACE
36
56
46
48
CONTROLLER
KEYPAD
34
42
58
MEMORY
CAMERA
DISPLAY
32

Fig. 1

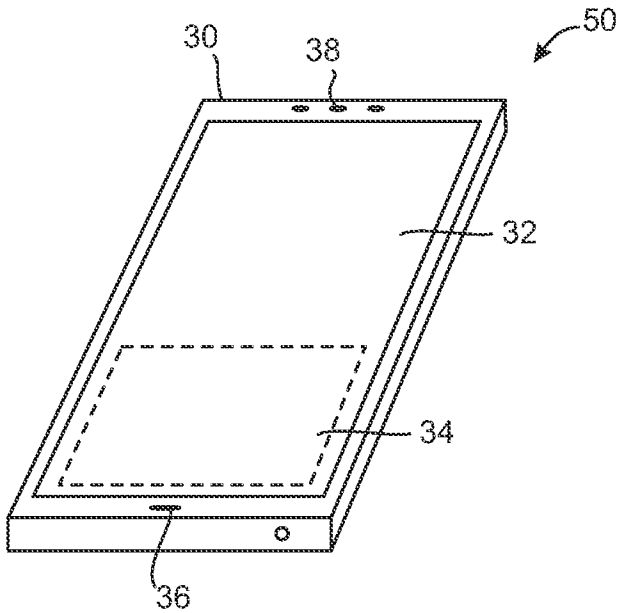

50
30
38
32
34
36

Fig. 2

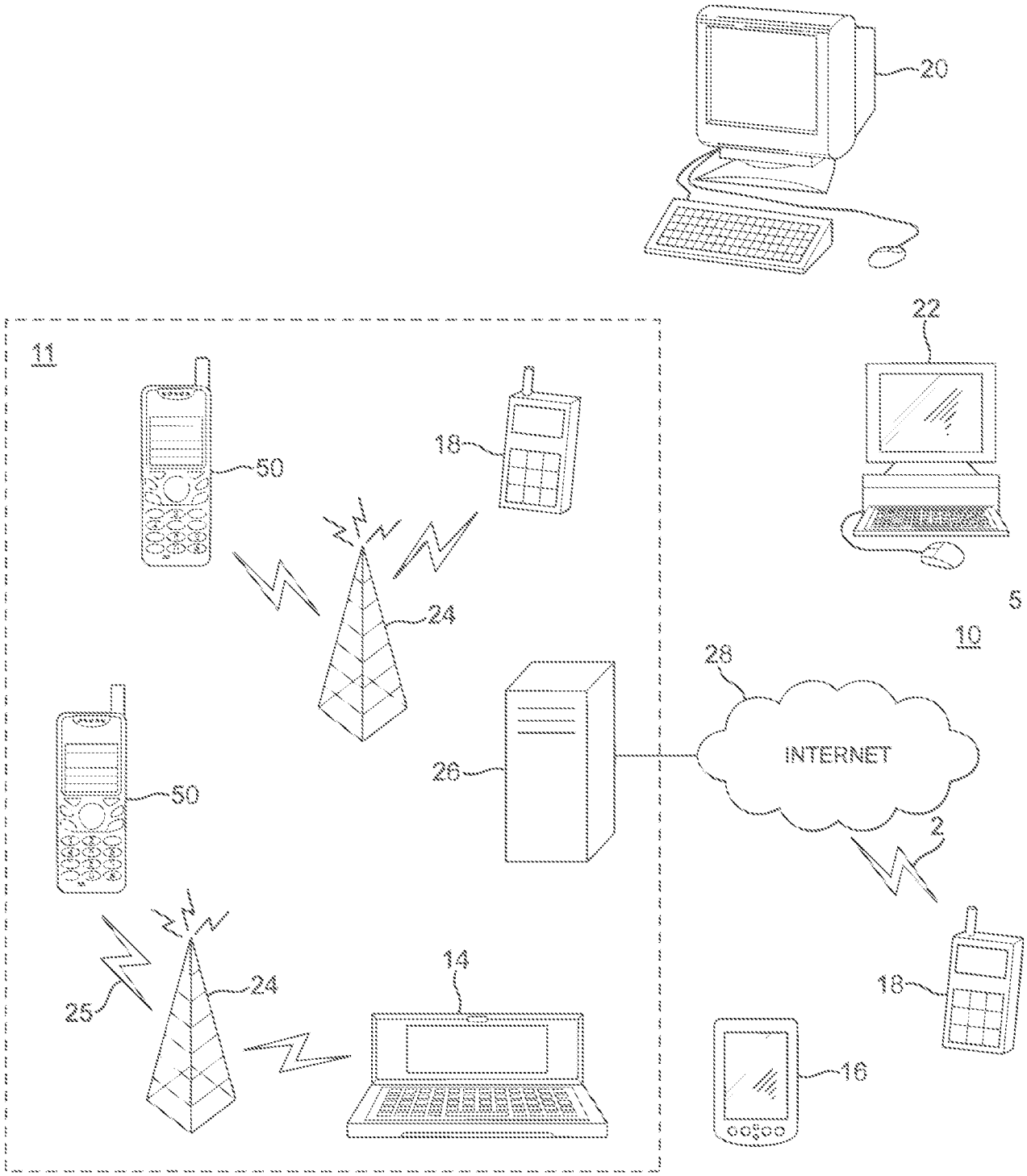


Fig. 3

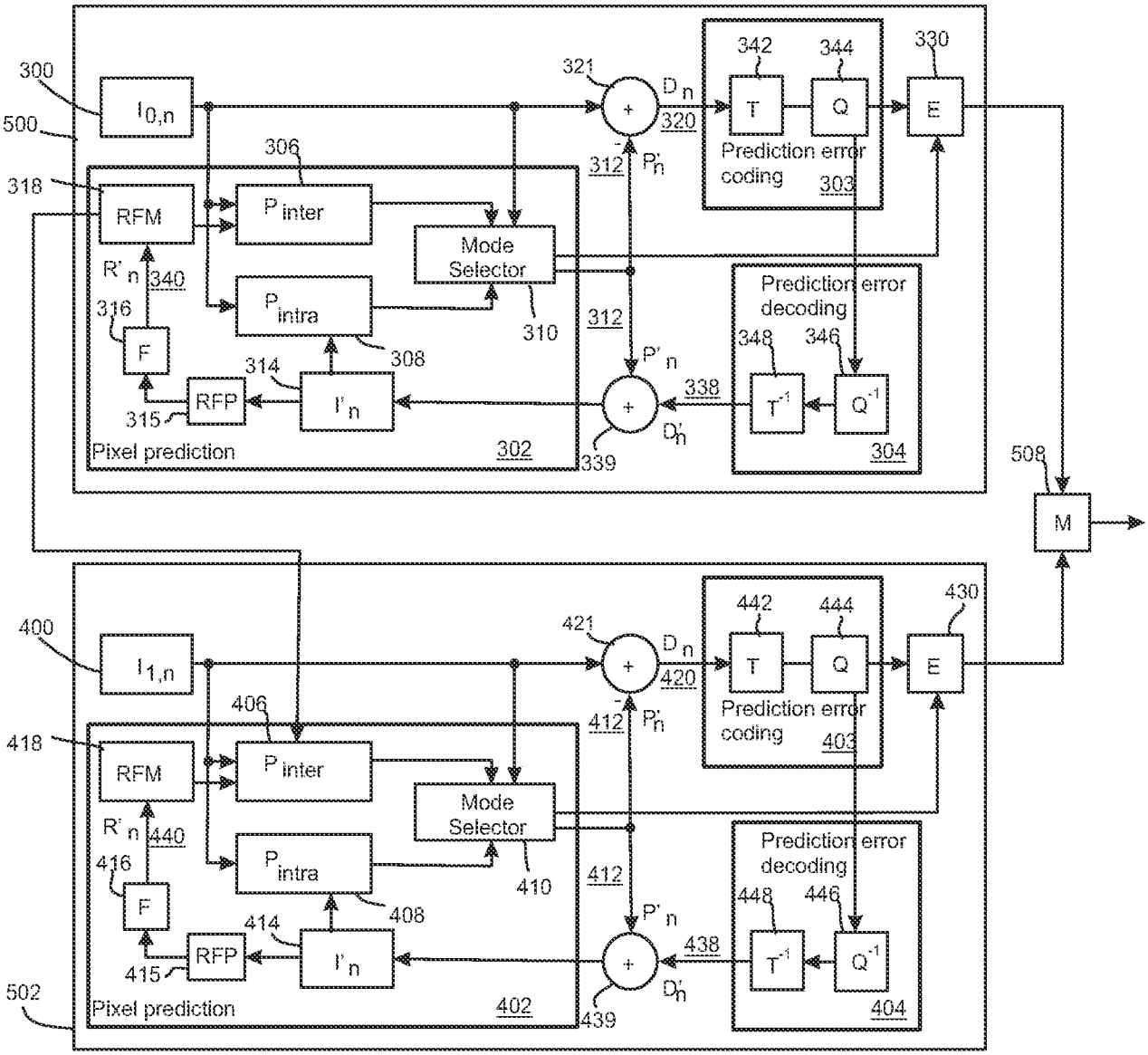

300
500
318
306
342
344
330
321
D n
320
T
Q
E
Prediction error coding
303
312
P'n
RFM
P inter
Mode Selector
R' n
340
P intra
310
312
Prediction error decoding
316
F
308
348
346
314
P' n
338
RFP
I' n
T-1
Q-1
315
D'n
Pixel prediction
302
339
304
508
M
400
442
444
430
421
D n
420
T
Q
E
Prediction error coding
403
412
P'n
406
418
RFM
P inter
Mode Selector
R' n
440
P intra
410
412
Prediction error decoding
416
F
408
448
446
414
P' n
438
RFP
I' n
T-1
Q-1
415
D'n
502
Pixel prediction
402
439
404

Fig. 4

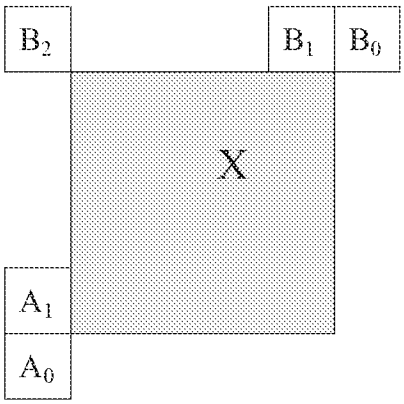


Fig.5a

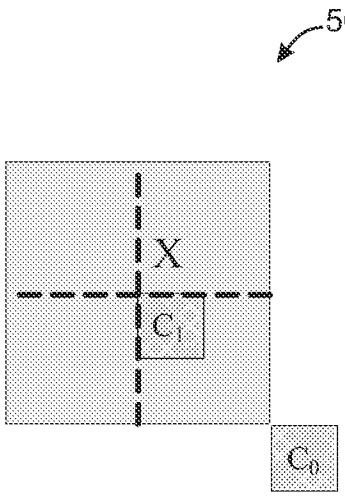


Fig.5b

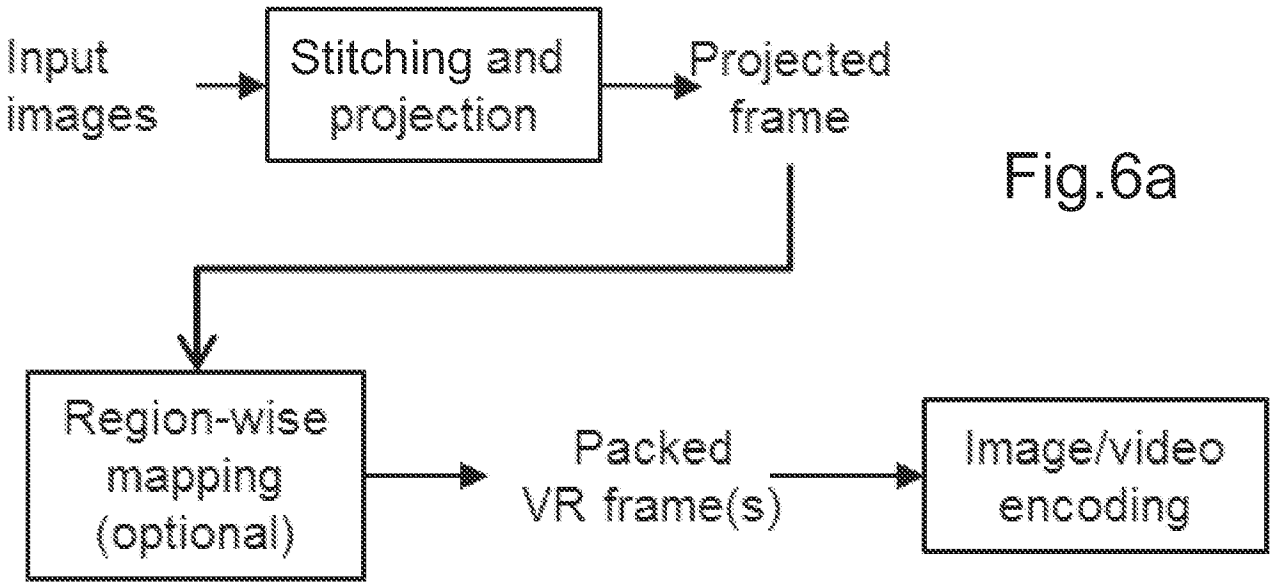


Fig.6a

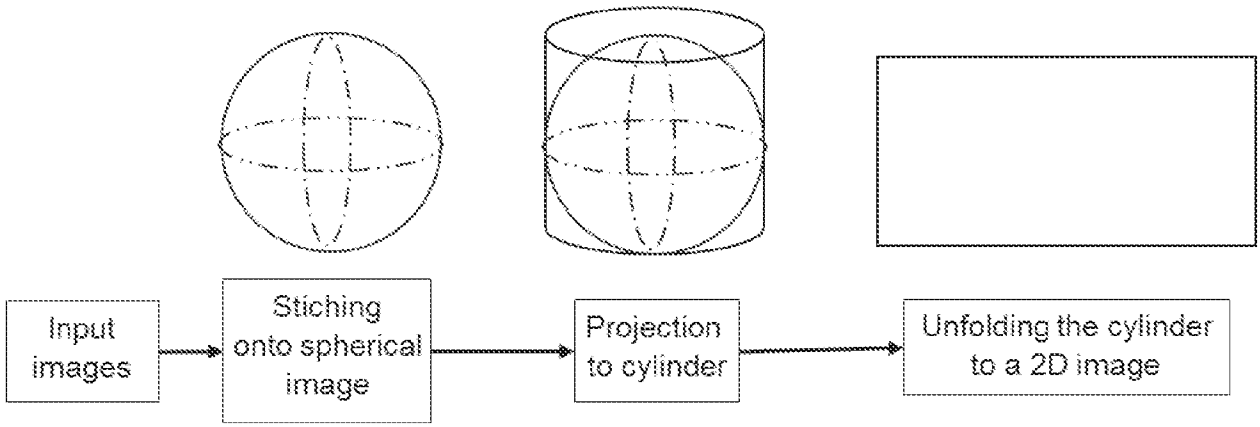


Fig.6b

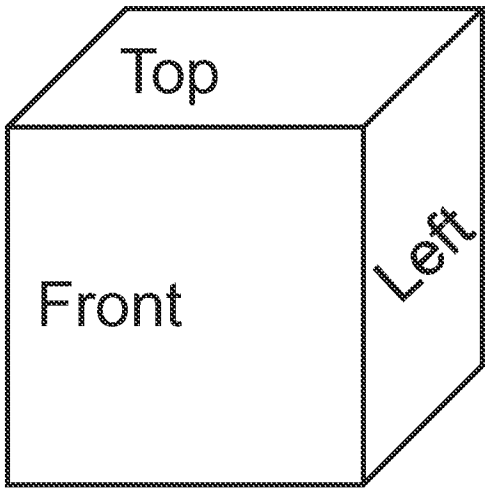


Fig. 7a

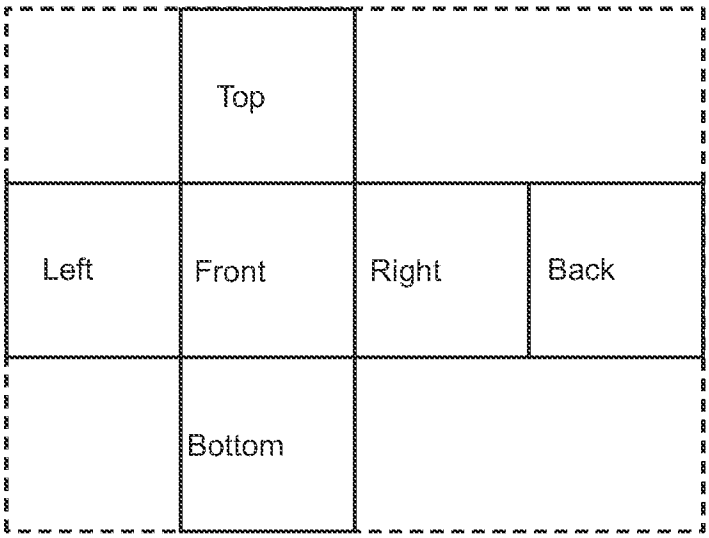


Fig. 7b

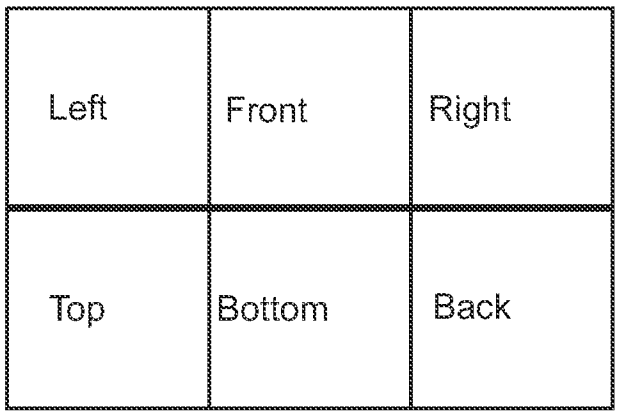


Fig. 7c

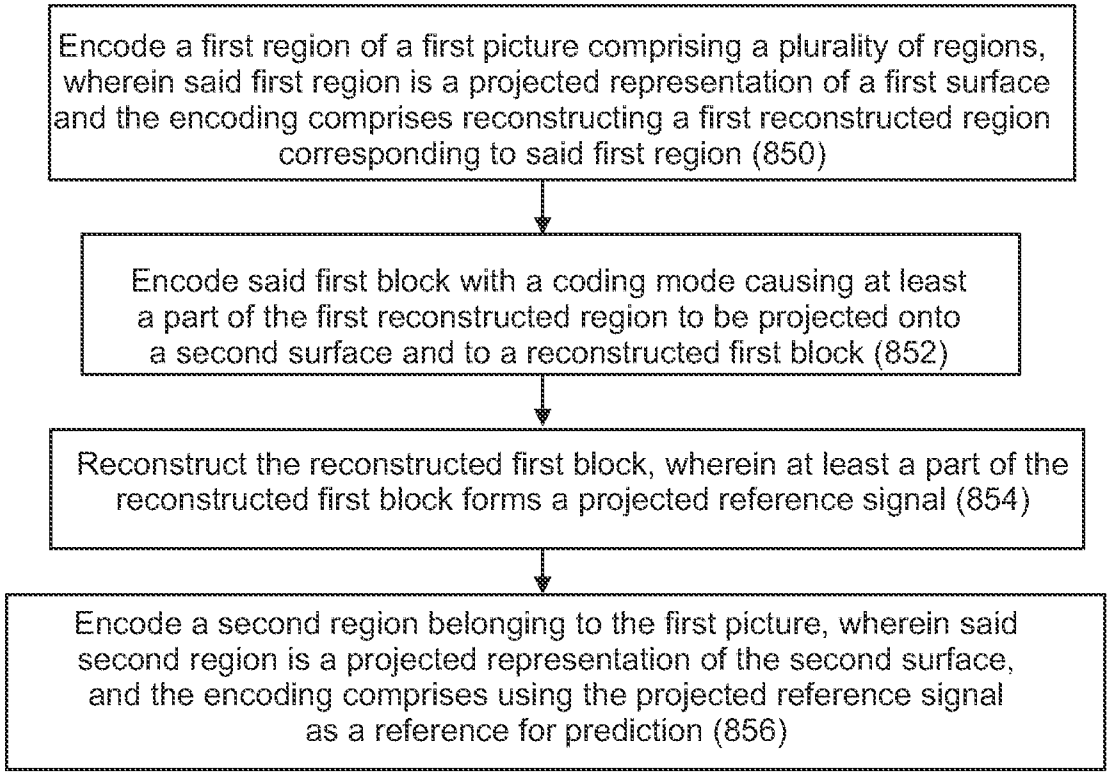

Encode a first region of a first picture comprising a plurality of regions, wherein said first region is a projected representation of a first surface and the encoding comprises reconstructing a first reconstructed region corresponding to said first region (850)
Encode said first block with a coding mode causing at least a part of the first reconstructed region to be projected onto a second surface and to a reconstructed first block (852)
Reconstruct the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal (854)
Encode a second region belonging to the first picture, wherein said second region is a projected representation of the second surface, and the encoding comprises using the projected reference signal as a reference for prediction (856)

Fig. 8a

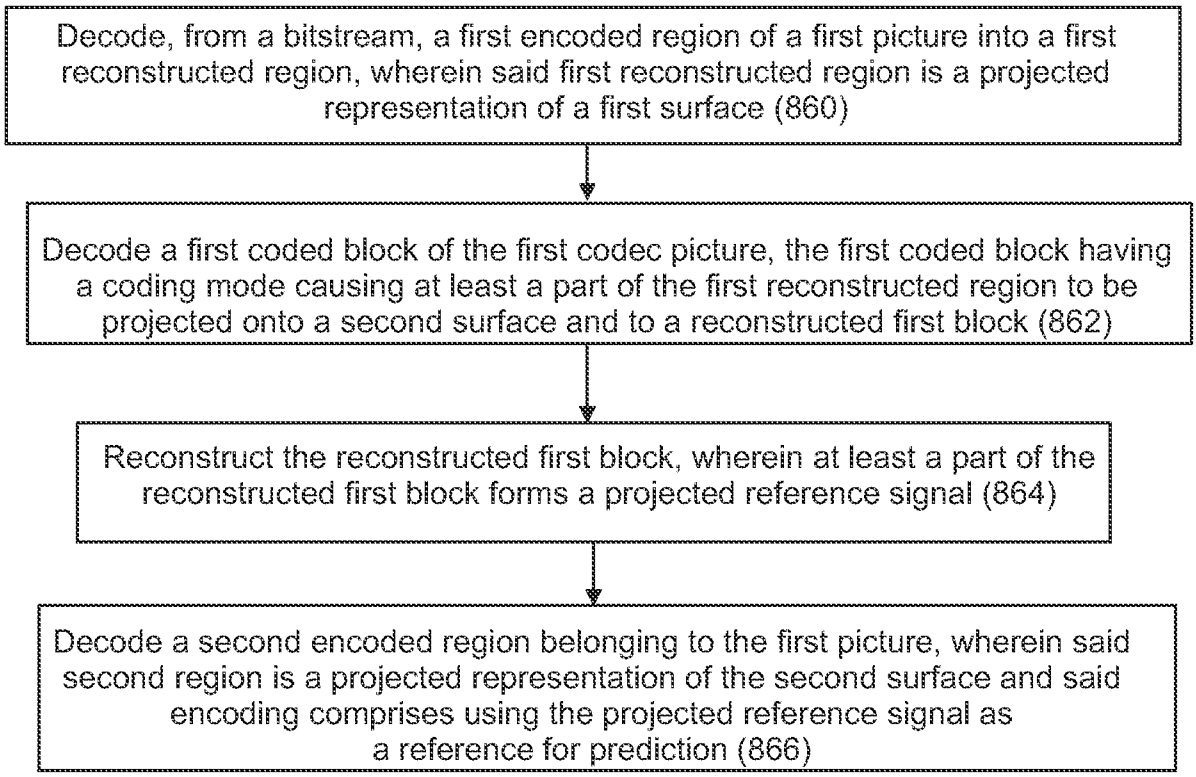

Decode, from a bitstream, a first encoded region of a first picture into a first reconstructed region, wherein said first reconstructed region is a projected representation of a first surface (860)
Decode a first coded block of the first codec picture, the first coded block having a coding mode causing at least a part of the first reconstructed region to be projected onto a second surface and to a reconstructed first block (862)
Reconstruct the reconstructed first block, wherein at least a part of the reconstructed first block forms a projected reference signal (864)
Decode a second encoded region belonging to the first picture, wherein said second region is a projected representation of the second surface and said encoding comprises using the projected reference signal as a reference for prediction (866)

Fig. 8b

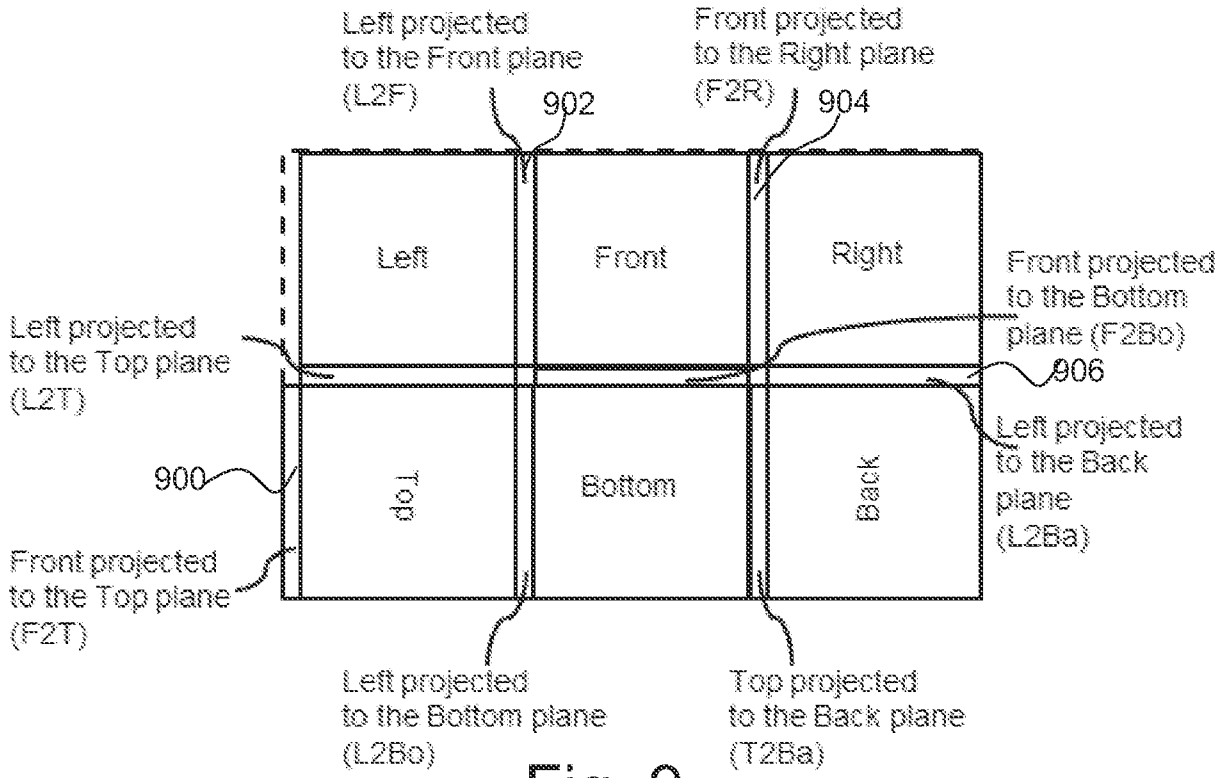

Left projected to the Front plane (L2F)
902
Front projected to the Right plane (F2R)
904
Left
Front
Right
Front projected to the Bottom plane (F2Bo)
906
Left projected to the Top plane (L2T)
900
Top
Bottom
Back
Left projected to the Back plane (L2Ba)
Front projected to the Top plane (F2T)
Left projected to the Bottom plane (L2Bo)
Top projected to the Back plane (T2Ba)

Fig. 9

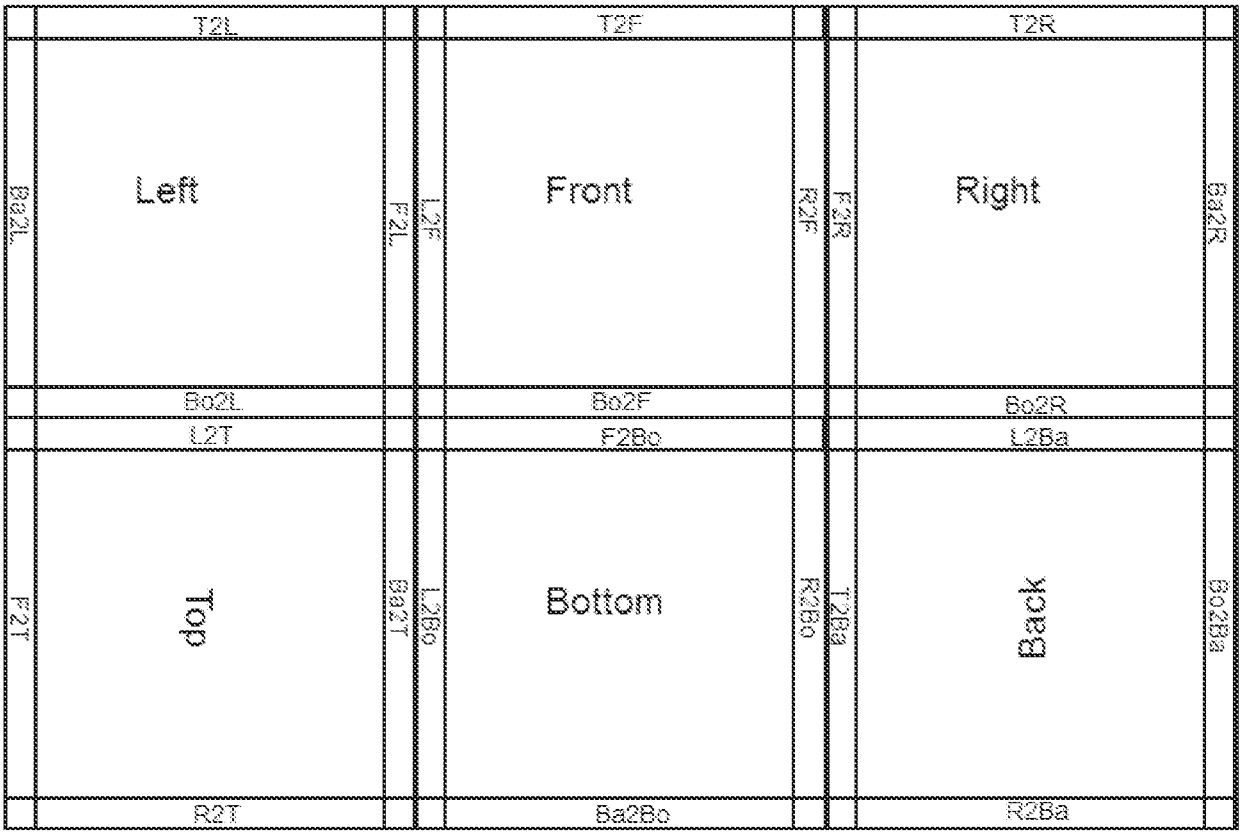

T2L
T2F
T2R
Left
Front
Right
Ba2L
L2F
F2L
R2F
F2R
Ba2R
Bo2L
Bo2F
Bo2R
L2T
F2Bo
L2Ba
Top
Bottom
Back
F2T
L2Bo
Ba2T
R2Bo
T2Ba
Bo2Ba
R2T
Ba2Bo
R2Ba

Fig. 10

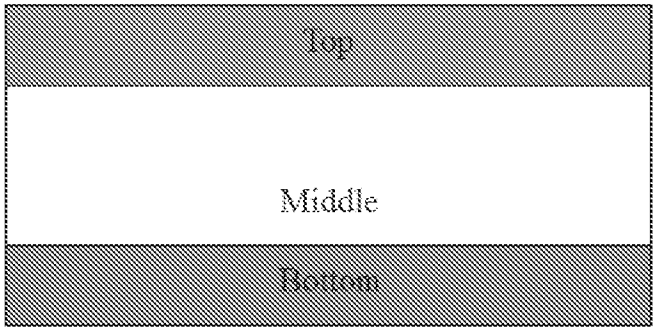


Fig. 11a

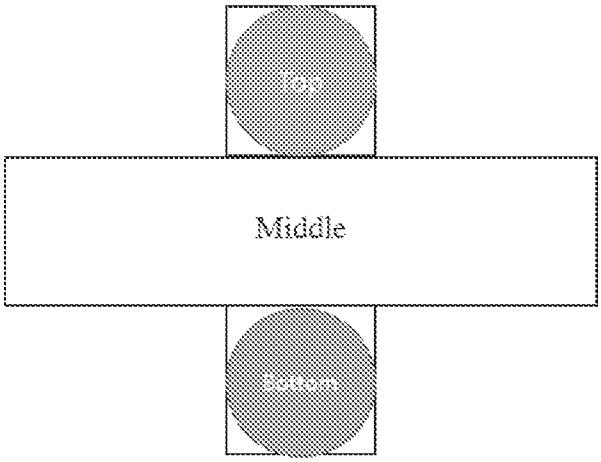


Fig. 11b

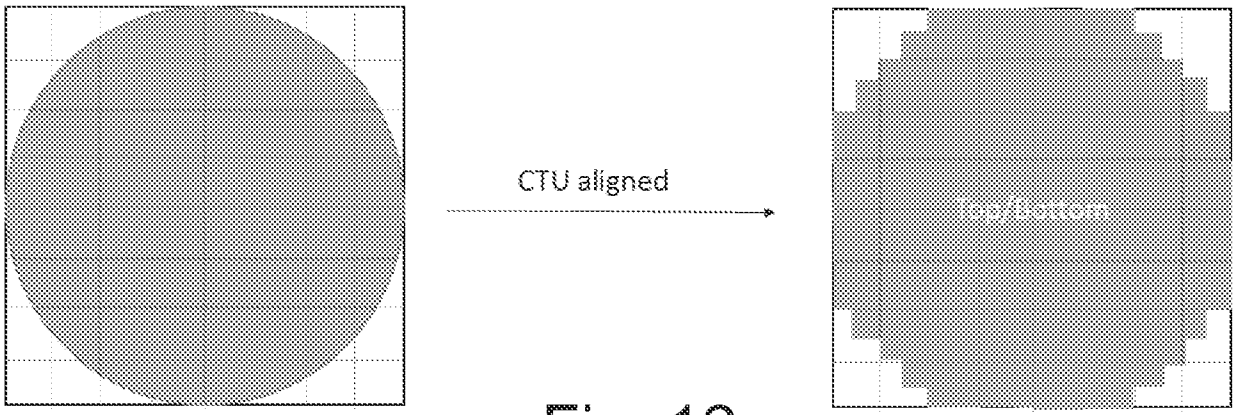


Fig. 12a

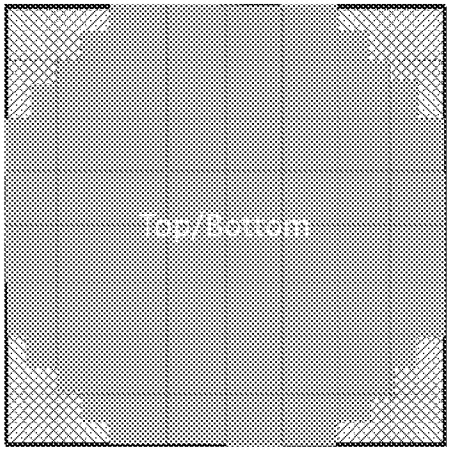


Fig. 12b

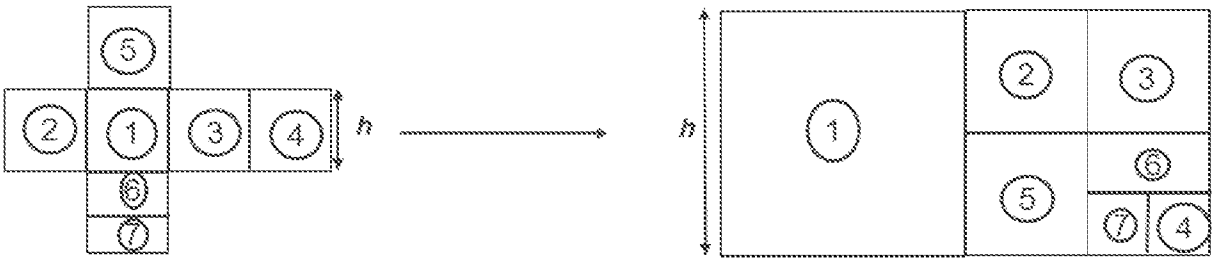


Fig. 13

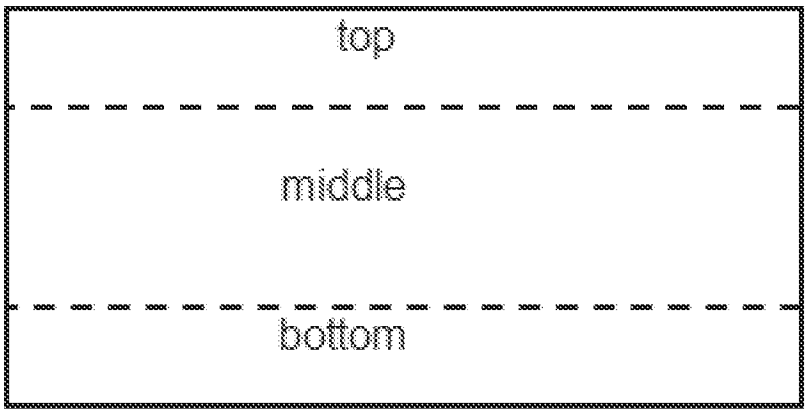


Fig. 14a

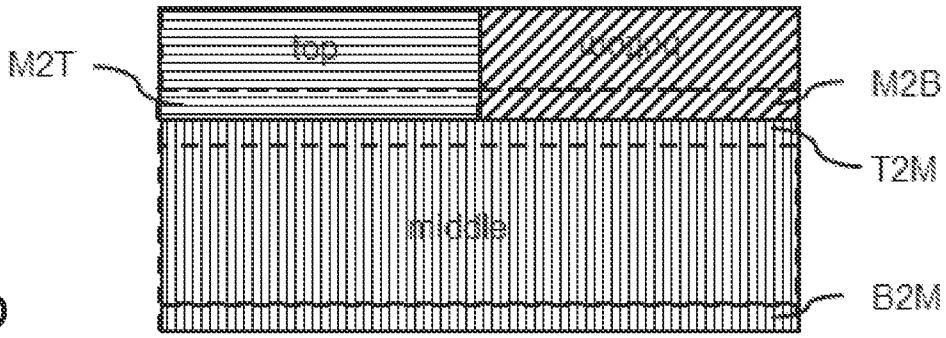
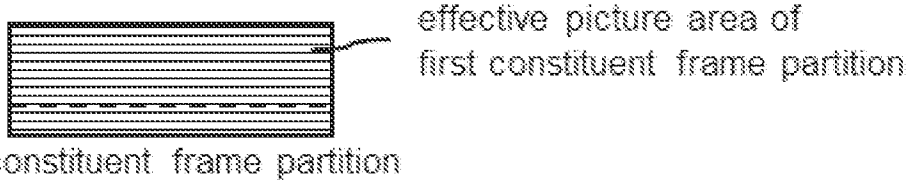
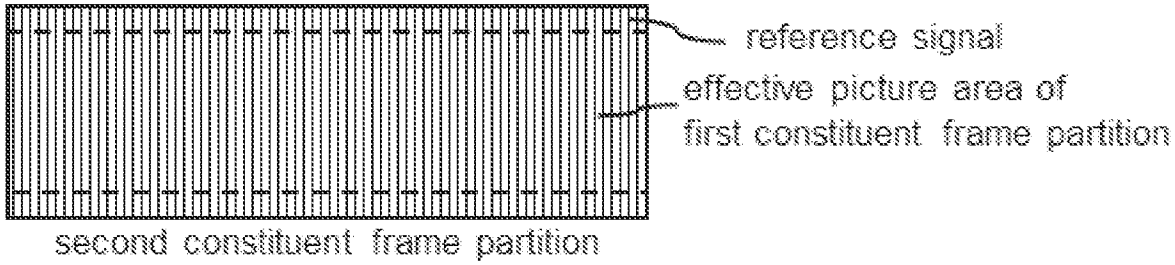


Fig. 14b





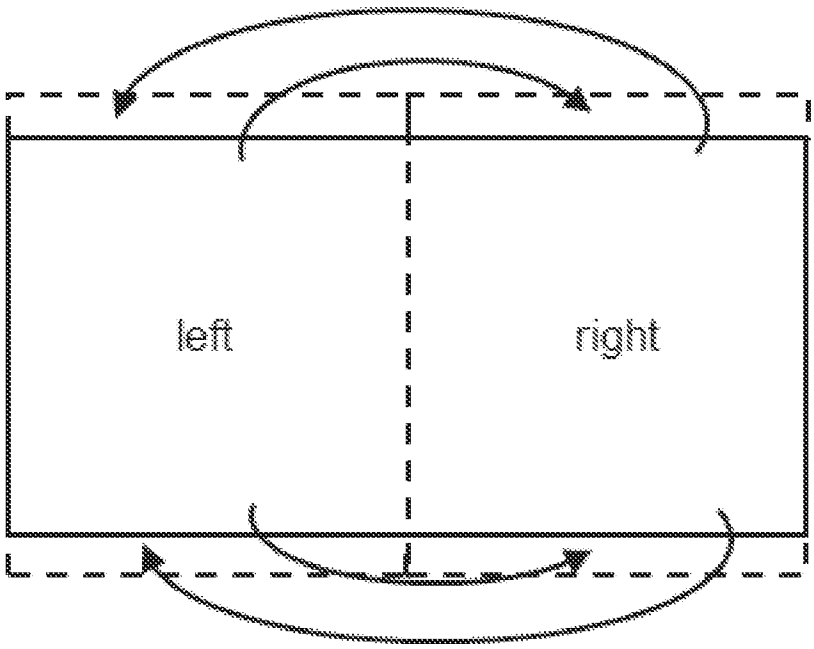


Fig. 15a

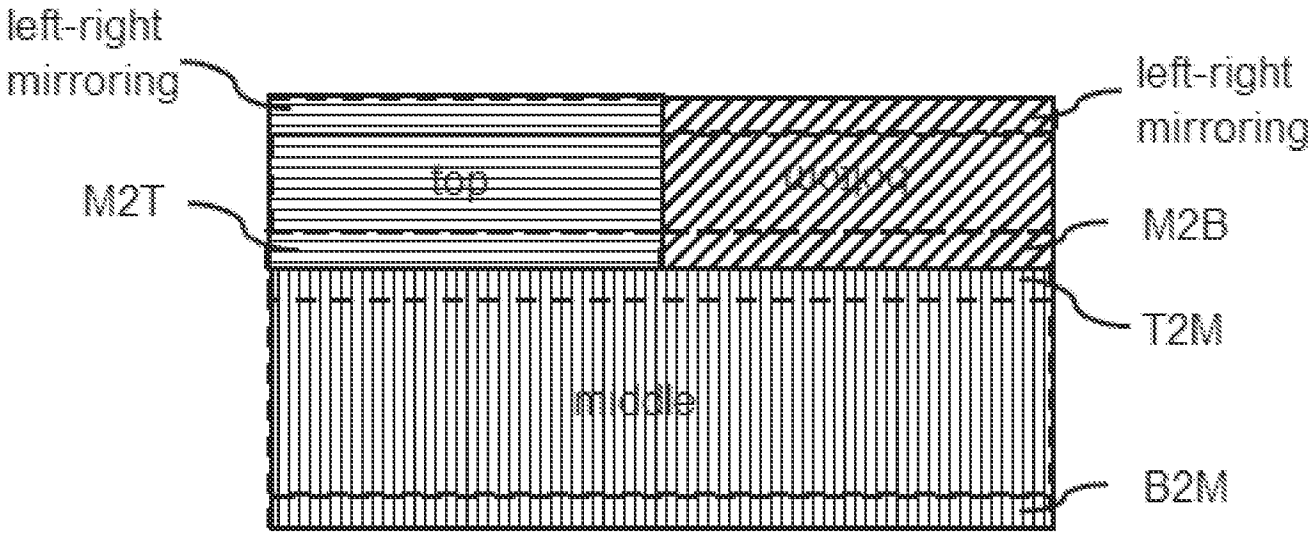


Fig. 15b

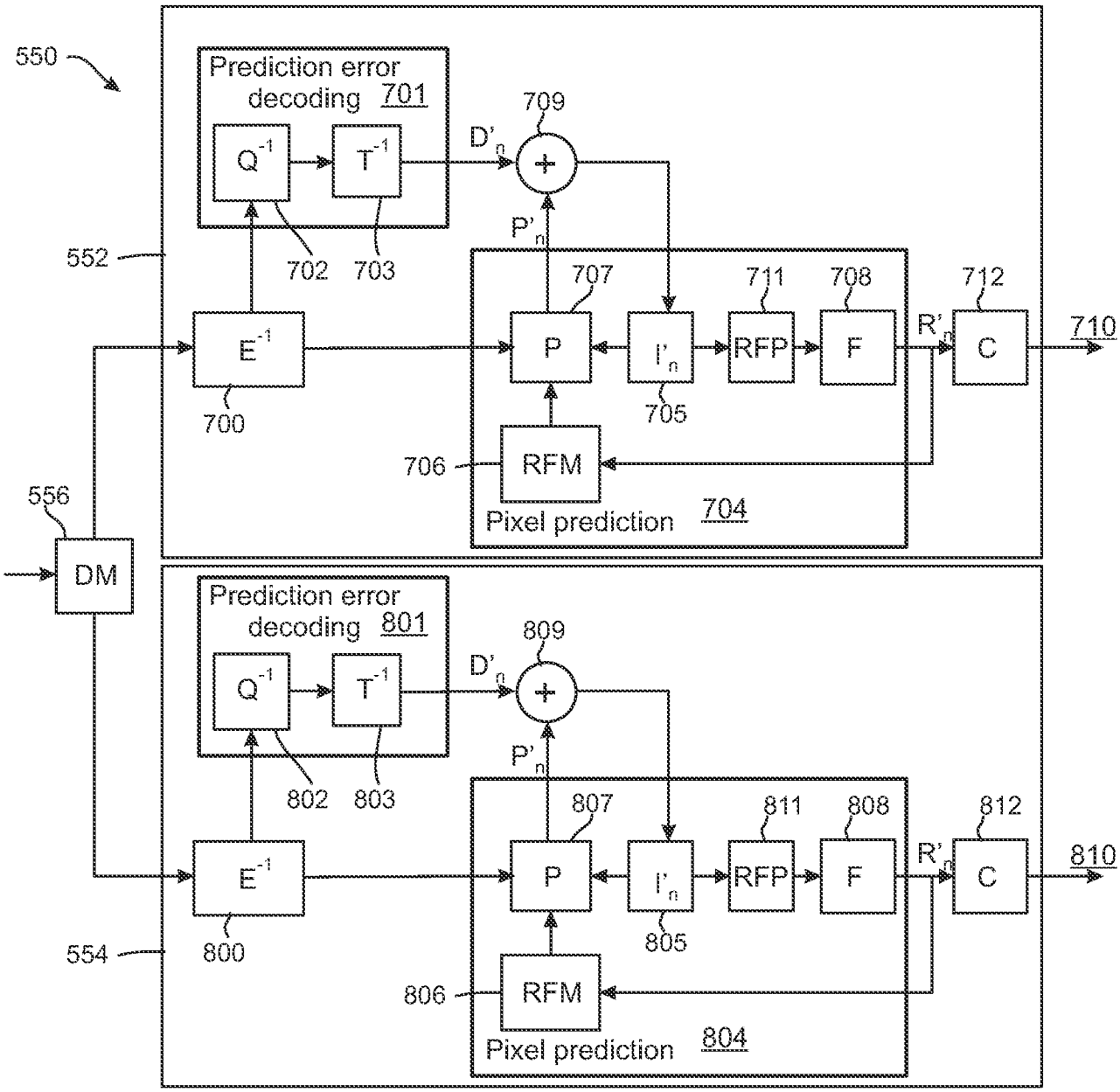

550
Prediction error decoding 701
Q-1
T-1
D'n
709
+
P'n
552
702
703
707
711
708
712
E-1
P
I'n
RFP
F
R'n
C
710
700
705
706
RFM
Pixel prediction 704
556
DM
Prediction error decoding 801
809
802
803
807
811
808
812
810
554
800
805
806
804

Fig. 16

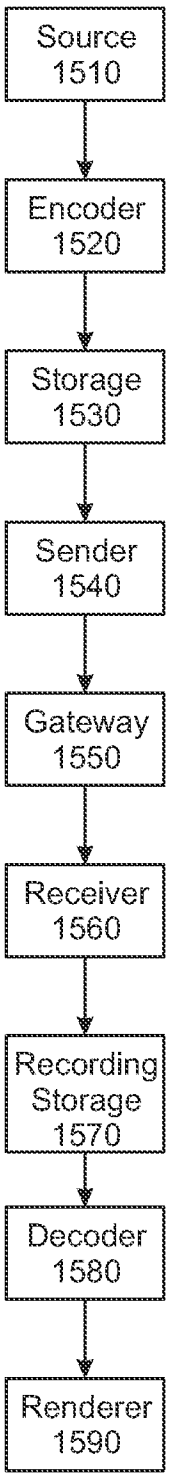

Source
1510
Encoder
1520
Storage
1530
Sender
1540
Gateway
1550
Receiver
1560
Recording
Storage
1570
Decoder
1580
Renderer
1590

Fig. 17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2017/050756

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, XP3GPP, XPAIP, XPESP, XPETSI, XPI3E, XPIEE, XPIETF, XPIOP, XPIPCOM, XPJPEG, XPMISC, XPOAC, XPRD, XPTK, COMPDX, EMBASE, INSPEC, TDB, NPL, JVET document management system, JCT-VC document management system

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HE, Y. et al. AHG8: Geometry padding for 360 video coding, Document: JVET-D0075. In: The 4th meeting of the Joint Video Exploration Team (JVET), Chengdu, CN: [online], 2016-10-15, [retrieved on 2018-02-15]. Retrieved from <http://phenix.it-sudparis.eu/jvet/doc_end_user/current_meeting.php?id_meeting=168&search_id_group=1&search_sub_group=1> abstract; sections 1, 2.2; Figs. 10-12	1-24
A	HANHART, P. et al. AHG8: Reference samples derivation using geometry padding for intra coding, Document: JVET-D0092. In: The 4th meeting of the Joint Video Exploration Team (JVET), Chengdu, CN: [online], 2016-10-15, [retrieved on 2018-02-27]. Retrieved from <http://phenix.it-sudparis.eu/jvet/doc_end_user/current_meeting.php?id_meeting=168&search_id_group=1&search_sub_group=1> abstract	1-24

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

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 February 2018 (22.02.2018)	02 March 2018 (02.03.2018)

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/FI2017/050756

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016112704 A1 (GRANGE ADRIAN WILLIAM [US] et al.) 21 April 2016 (21.04.2016) abstract; pars. 0062-0069; Figs. 4A-D, 5	1-24
A	US 2006034374 A1 (PARK GWANG-HOON [KR] et al.) 16 February 2006 (16.02.2006) abstract; pars. 0058-0060, 0064-0065; Figs. 7A-B, 8A-B, 9	1-24
A	CHOI, B. et al. WD on ISO/IEC 23000-20 Omnidirectional Media Application Format. In: The 115th MPEG meeting, Geneva, Switzerland: [online], June 2016, [retrieved on 2018-02-19]. Retrieved from <https://mpeg.chiariglione.org/standards/mpeg-a/omnidirectional-media-application-format/wd-isoiec-23000-20-omnidirectional-media> sections A.3.2.3, A.3.2.6; Figs. A-6 – A-8, A-20	7-8, 19-20
E	WO 2018009746 A1 (VID SCALE INC [US]) 11 January 2018 (11.01.2018)	1-2, 6, 9-14, 18, 21-24

INTERNATIONAL SEARCH REPORT
Information on Patent Family Members

International application No.
PCT/FI2017/050756

Patent document cited in search report	Publication date	Patent family members(s)	Publication date
US 2016112704 A1	21/04/2016	CN 106664403 A	10/05/2017
		DE 112015004764 T5	19/10/2017
		EP 3210379 A1	30/08/2017
		GB 201621542 D0	01/02/2017
		GB 2545999 A	05/07/2017
		US 2016112489 A1	21/04/2016
		US 2016112705 A1	21/04/2016
		US 2016112713 A1	21/04/2016
		WO 2016064862 A1	28/04/2016
US 2006034374 A1	16/02/2006	CN 101002479 A	18/07/2007
		CN 100544444 C	23/09/2009
		EP 1779672 A1	02/05/2007
		JP 2008510358 A	03/04/2008
		JP 4842945 B2	21/12/2011
		KR 20060015225 A	16/02/2006
		KR 100677142 B1	02/02/2007
		WO 2006016783 A1	16/02/2006
WO 2018009746 A1	11/01/2018	None	

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2017/050756

CLASSIFICATION OF SUBJECT MATTER

IPC
H04N 19/563 (2014.01)
H04N 19/593 (2014.01)
H04N 19/597 (2014.01)
G06T 17/30 (2006.01)
G06T 17/10 (2006.01)