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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 September 2017 (28.09.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/162911 A1

(51) International Patent Classification:
H04N 19/55 (2014.01) *H04N 19/597* (2014.01)
H04N 19/53 (2014.01) *H04N 19/33* (2014.01)
H04N 19/56 (2014.01)

(21) International Application Number:
PCT/FI2017/050177

(22) International Filing Date:
17 March 2017 (17.03.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
20165257 24 March 2016 (24.03.2016) FI

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

(72) Inventor: **HANNUKSELA, Miska**; Rusthollinrinne 2, 33610 Tampere (FI).

(74) Agents: **NOKIA TECHNOLOGIES OY** et al.; Ari Aarnio, IPR Department, Karakaari 7, 02610 Espoo (FI).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU,

[Continued on next page]

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

(57) Abstract: There is provided a method comprising determining a reference region of a picture, determining a target region of the picture, obtaining information of a motion field of the reference region,and including an indication into a bit stream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture. There is also provided a method comprising decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture, decoding the reference region of the picture,and decoding the target region of the picture.

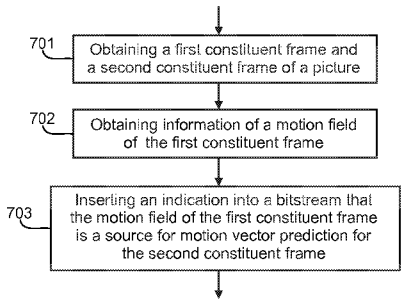


Fig. 7a

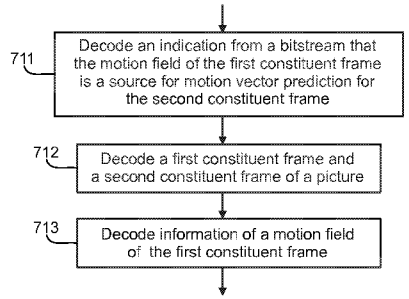


Fig. 7b

WO 2017/162911 A1

LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

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

TECHNICAL FIELD

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

BACKGROUND

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

A video coding system may comprise an encoded 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.

SUMMARY

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.

A method according to a first aspect comprises:
decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

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

decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

A computer readable storage medium according to a third aspect comprises code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:
decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

An apparatus according to a fourth aspect comprises:
means for decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
means for decoding the reference region of the picture; and
means for decoding the target region of the picture.

A method according to a fifth aspect comprises:
obtaining a reference region of a first picture;
obtaining a target region of the first picture;
obtaining information of a motion field of the reference region; and
including an indication into a bitstream that the motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

An apparatus according to a sixth aspect comprises at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
determining a reference region of a picture;
determining a target region of the picture;
obtaining information of a motion field of the reference region; and
including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

A computer readable storage medium according to a seventh aspect comprises code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:
determining a reference region of a picture;
determining a target region of the picture;

obtaining information of a motion field of the reference region; and
including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

An apparatus according to an eighth aspect comprises:
means for determining a reference region of a picture;
means for determining a target region of the picture;
means for obtaining information of a motion field of the reference region; and
means for including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

A method according to a ninth aspect comprises:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
decoding the target region using at least one of sample prediction and motion prediction from the reference region.

An apparatus according to a tenth aspect comprises at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
decoding the target region using at least one of sample prediction and motion prediction from the reference region.

A computer readable storage medium according to an eleventh aspect comprises code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;

decoding the target region using at least one of sample prediction and motion prediction from the reference region.

An apparatus according to a twelfth aspect comprises:
means for decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
means for decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
means for decoding the target region using at least one of sample prediction and motion prediction from the reference region.

A method according to a thirteenth aspect comprises:
encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
encoding the target region using at least one of sample prediction and motion prediction from the reference region.

An apparatus according to a fourteenth aspect comprises at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
encoding the target region using at least one of sample prediction and motion prediction from the reference region.

A computer readable storage medium according to a fifteenth aspect comprises code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:
encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;

encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
encoding the target region using at least one of sample prediction and motion prediction from the reference region.

An apparatus according to a sixteenth aspect comprises:
means for encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
means for encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
means for encoding the target region using at least one of sample prediction and motion prediction from the reference region.

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

BRIEF DESCRIPTION OF THE DRAWINGS

For better understanding of the present invention, reference will now be made by way of example to the accompanying drawings in which:
Figure 1 shows schematically an electronic device employing embodiments of the invention;
Figure 2 shows schematically a user equipment suitable for employing embodiments of the invention;
Figure 3 further shows schematically electronic devices employing embodiments of the invention connected using wireless and wired network connections;
Figure 4 shows schematically an encoder suitable for implementing embodiments of the invention;
Figure 5a shows spatial candidate sources of the candidate motion vector predictor, in accordance with an embodiment;
Figure 5b shows temporal candidate sources of the candidate motion vector predictor, in accordance with an embodiment;
Figure 6a illustrates an example, where a motion field of a first constituent frame of a picture is used as a reference for motion vector prediction for a second constituent frame of the picture, in accordance with an embodiment;
Figure 6b illustrates the relation of reference region offsets and scaled reference layer offsets, in accordance with an embodiment;

Figure 7a shows a flow chart of enabling inter-view motion prediction between the frame-packed constituent frames, in accordance with an embodiment;

Figure 7b shows a flow chart of generating a bitstream according to an embodiment of the invention;

Figure 8 shows a schematic diagram of a decoder suitable for implementing embodiments of the invention;

Figure 9 shows a schematic diagram of an example multimedia communication system within which various embodiments may be implemented;

Figures 10a to 10c shows an example an example of downsampling the top and bottom parts of an equirectangular panorama picture and arranging the downsampled top and bottom parts as well as the middle part into a containing picture; and

Figure 11 illustrates an example embodiment where a reference region is indicated to reside in a first picture of a first layer and a target region is indicated to reside in a second picture of the first layer, the first picture is of the representation format described with Figures 10a to 10c, and the second picture is of the representation format described with Fig. 10a. Figure 11 also illustrates the reference region and the target region when predicting the top part of the second picture from the top part of the first picture, in accordance with an embodiment.

DETAILED DESCRIPTION OF SOME EXAMPLE EMBODIMENTS

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.

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.

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.

The apparatus 50 may comprise a housing 30 for incorporating and protecting the device. The apparatus 50 further may comprise a display 32 in the form of a liquid crystal display. In other embodiments of the invention the display may be any suitable display technology suitable to display an image or video. The apparatus 50 may further comprise a keypad 34. In other embodiments of the invention any suitable data or user interface mechanism may be employed. For example the user interface may be implemented as a virtual keyboard or data entry system as part of a touch-sensitive display.

The apparatus may comprise a microphone 36 or any suitable audio input which may be a digital or analogue signal input. The apparatus 50 may further comprise an audio output device which in embodiments of the invention may be any one of: an earpiece 38, speaker, or an analogue audio or digital audio output connection. The apparatus 50 may also comprise a battery 40 (or in other embodiments of the invention the device may be powered by any suitable mobile energy device such as solar cell, fuel cell or clockwork generator). The 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.

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.

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.

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).

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

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.

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.

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.

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. 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.

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.

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

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.

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.

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.

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.

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).

Version 1 of the High Efficiency Video Coding (H.265/HEVC a.k.a. HEVC) standard was developed by the Joint Collaborative Team – Video Coding (JCT-VC) of VCEG and MPEG. The

standard was published by both parent standardization organizations, and it is referred to as ITU-T Recommendation H.265 and ISO/IEC International Standard 23008-2, also known as MPEG-H Part 2 High Efficiency Video Coding (HEVC). Version 2 of H.265/HEVC included scalable, multiview, and fidelity range extensions, which may be abbreviated SHVC, MV-HEVC, and REXT, respectively. Version 2 of H.265/HEVC was published as ITU-T Recommendation H.265 (10/2014) and as Edition 2 of ISO/IEC 23008-2. There are currently ongoing standardization projects to develop further extensions to H.265/HEVC, including three-dimensional and screen content coding extensions, which may be abbreviated 3D-HEVC and SCC, respectively.

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.

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.

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.

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.

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.

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).

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.

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.

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.

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.

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.

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.

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).

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.

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.

In HEVC, the partitioning to tiles forms a regular grid, where heights and widths of tiles differ from each other by one LCU at the maximum. In HEVC, a slice is defined to be an integer number of coding tree units contained in one independent slice segment and all subsequent dependent slice segments (if any) that precede the next independent slice segment (if any) within the same access unit. In HEVC, a slice segment is defined to be an integer number of coding tree units ordered consecutively in the tile scan and contained in a single NAL unit. The division of each picture into slice segments is a partitioning. In HEVC, an independent slice segment is defined to be a slice segment for which the values of the syntax elements of the slice segment header are not inferred from the values for a preceding slice segment, and a dependent slice segment is defined to be a slice segment for which the values of some syntax elements of the slice segment header are inferred from the values for the preceding independent slice segment in decoding order. In HEVC, a slice

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

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.

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.

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.
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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

The syntax element max_tid_il_ref_pics_plus1 in the VPS extension can be used to indicate that non-IRAP pictures are not used a reference for inter-layer prediction and, if not so, which temporal sub-layers are not used as a reference for inter-layer prediction:
max_tid_il_ref_pics_plus1[i][j] equal to 0 specifies that non-IRAP pictures with nuh_layer_id equal to layer_id_in_nuh[i] are not used as source pictures for inter-layer prediction for pictures with nuh_layer_id equal to layer_id_in_nuh[j]. max_tid_il_ref_pics_plus1[i][j] greater than 0 specifies that pictures with nuh_layer_id equal to layer_id_in_nuh[i] and TemporalId greater than max_tid_il_ref_pics_plus1[i][j] − 1 are not used as source pictures for inter-layer prediction for pictures with nuh_layer_id equal to layer_id_in_nuh[j]. When not present, the value of max_tid_il_ref_pics_plus1[i][j] is inferred to be equal to 7.

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.

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

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.

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.

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.

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

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

The order of byte stream NAL units in the byte stream may be required to follow the decoding order of the NAL units contained in the byte stream NAL units. The semantics of syntax elements may be specified as follows. leading_zero_8bits is a byte equal to 0x00. The leading_zero_8bits

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

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

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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

- The post-production workflows might be tailored for a single video signal. Some post-production tools might not be able to handle two separate picture sequences and/or might not be able to keep the separate picture sequences in synchrony with each other.
- The distribution system, such as transmission protocols, might be such that support single coded sequence only and/or might not be able to keep separate coded sequences in synchrony with each other and/or may require more buffering or latency to keep the separate coded sequences in synchrony with each other.
- The decoding of bitstreams with multiview video coding tools may require support of specific coding modes, which might not be available in players. For example, many smartphones support H.265/HEVC Main profile decoding but are not able to handle H.265/HEVC Multiview Main profile decoding even though it only requires high-level additions compared to the Main profile.

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

A specific projection for mapping a panoramic image covering 360-degree field-of-view horizontally and 180-degree field-of-view vertically (hence representing a sphere) to a rectangular two-dimensional image plane is known as equirectangular projection. In this case, the horizontal

coordinate may be considered equivalent to a longitude, and the vertical coordinate may be considered equivalent to a latitude, with no transformation or scaling applied. In some cases panoramic content with 360-degree horizontal field-of-view but with less than 180-degree vertical field-of-view may be considered special cases of equirectangular projection, where the polar areas of the sphere have not been mapped onto the two-dimensional image plane. In some cases panoramic content may have less than 360-degree horizontal field-of-view and up to 180-degree vertical field-of-view, while otherwise have the characteristics of equirectangular projection format.

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

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 is planned to include IBC.

IBC in HEVC SCC is meant for sample prediction only. It is explicitly disallowed in the HEVC SCC specification text to use the current picture as the collocated picture for TMVP.

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

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.

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.

In H.265/HEVC, AMVP and the merge mode may be characterized as follows. In AMVP, the encoder indicates whether uni-prediction or bi-prediction is used and which reference pictures are used as well as encodes a motion vector difference. In the merge mode, only the chosen candidate from the candidate list is encoded into the bitstream indicating the current prediction unit has the same motion information as that of the indicated predictor. Thus, the merge mode creates regions composed of neighbouring prediction blocks sharing identical motion information, which is only signalled once for each region. Another difference between AMVP and the merge mode in H.265/HEVC is that the maximum number of candidates of AMVP is 2 while that of the merge mode is 5.

The advanced motion vector prediction may operate for example as follows, while other similar realizations of advanced motion vector prediction are also possible for example with different candidate position sets and candidate locations with candidate position sets. Two spatial motion vector predictors (MVPs) may be derived and a temporal motion vector predictor (TMVP) may be derived. They may be selected among the positions: three spatial motion vector predictor candidate positions located above the current prediction block (B0, B1, B2) and two on the left (A0, A1). The first motion vector predictor that is available (e.g. resides in the same slice, is inter-coded, etc.) in a pre-defined order of each candidate position set, (B0, B1, B2) or (A0, A1), may be selected to represent that prediction direction (up or left) in the motion vector competition. A reference index for the temporal motion vector predictor may be indicated by the encoder in the slice header (e.g. as a collocated_ref_idx syntax element). The first motion vector predictor that is available (e.g. is inter-coded) in a pre-defined order of potential temporal candidate locations, e.g. in the order (C_0, C_1), may be selected as a source for a temporal motion vector predictor. The motion vector obtained from the first available candidate location in the co-located picture may be scaled according to the proportions of the picture order count differences of the reference picture

of the temporal motion vector predictor, the co-located picture, and the current picture. Moreover, a redundancy check may be performed among the candidates to remove identical candidates, which can lead to the inclusion of a zero motion vector in the candidate list. The motion vector predictor may be indicated in the bitstream for example by indicating the direction of the spatial motion vector predictor (up or left) or the selection of the temporal motion vector predictor candidate. The co-located picture may also be referred to as the collocated picture, the source for motion vector prediction, or the source picture for motion vector prediction.

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

In the merging/merge mode/process/mechanism, where all the motion information of a block/PU is predicted and used without any modification/correction. The aforementioned motion information for a PU may comprise one or more of the following: 1) The information whether 'the PU is uni-predicted using only reference picture list0' or 'the PU is uni-predicted using only reference picture list1' or 'the PU is bi-predicted using both reference picture list0 and list1'; 2) Motion vector value corresponding to the reference picture list0, which may comprise a horizontal and vertical motion vector component; 3) Reference picture index in the reference picture list0 and/or an identifier of a reference picture pointed to by the Motion vector corresponding to reference picture list 0, where the identifier of a reference picture may be for example a picture order count value, a layer identifier value (for inter-layer prediction), or a pair of a picture order count value and a layer identifier value; 4) Information of the reference picture marking of the reference picture, e.g. information whether the reference picture was marked as "used for short-term reference" or "used for long-term reference"; 5) - 7) The same as 2) - 4), respectively, but for reference picture list1.

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

improve prediction quality. This type of prediction mode is typically named as an inter-merge mode.

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.

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).

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)

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.

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.

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.

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.

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.

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. For MV-HEVC, the input multiview video sequences for encoding are typically captured by a number of cameras arranged in a row. The camera projection centers are typically collinear and equally distant from each neighbor and cameras typically point to the same direction. SHVC and MV-HEVC share the same high-level syntax and most parts of their decoding process are also identical, which makes it appealing to support both SHVC and MV-HEVC with the same codec implementation.

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.

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.

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.

Figure 6b illustrates the relation of reference region offsets 658 and scaled reference layer offsets 660. In this example, the reference layer locations offsets 658 are indicated for an enhancement layer (EL) picture 650 in relation to a base layer (BL) picture 652. In Figure 6b, the resampled reference region is indicated with the dashed rectangle 654 within the EL picture 650. A horizontal resampling ratio may be defined to be equal to W_{EL} / W_{BL}. A vertical resampling ratio may be defined to be equal to H_{EL} / H_{BL}.

Some scalable video coding schemes may require IRAP pictures to be aligned across layers in a manner that either all pictures in an access unit are IRAP pictures or no picture in an access unit is an IRAP picture. Other scalable video coding schemes, such as the multi-layer extensions of HEVC, may allow IRAP pictures that are not aligned, i.e. that one or more pictures in an access unit are IRAP pictures, while one or more other pictures in an access unit are not IRAP pictures. Scalable bitstreams with IRAP pictures or similar that are not aligned across layers may be used for example for providing more frequent IRAP pictures in the base layer, where they may have a smaller coded size due to e.g. a smaller spatial resolution. A process or mechanism for layer-wise start-up of the decoding may be included in a video decoding scheme. Decoders may hence start decoding of a bitstream when a base layer contains an IRAP picture and step-wise start decoding other layers when they contain IRAP pictures. In other words, in a layer-wise start-up of the decoding mechanism or process, decoders progressively increase the number of decoded layers (where layers may represent an enhancement in spatial resolution, quality level, views, additional components such as depth, or a combination) as subsequent pictures from additional enhancement layers are decoded in the decoding process. The progressive increase of the number of decoded layers may be perceived for example as a progressive improvement of picture quality (in case of quality and spatial scalability).

A layer-wise start-up mechanism may generate unavailable pictures for the reference pictures of the first picture in decoding order in a particular enhancement layer. Alternatively, a decoder may omit the decoding of pictures preceding, in decoding order, the IRAP picture from which the decoding of a layer can be started. These pictures that may be omitted may be specifically labeled by the encoder or another entity within the bitstream. For example, one or more specific NAL unit types may be used for them. These pictures, regardless of whether they are specifically marked with a NAL unit type or inferred e.g. by the decoder, may be referred to as cross-layer random

access skip (CL-RAS) pictures. The decoder may omit the output of the generated unavailable pictures and the decoded CL-RAS pictures.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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).

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.

A coding standard or system may refer to a term operation point or alike, which may indicate the scalable layers and/or sub-layers under which the decoding operates and/or may be associated with

a sub-bitstream that includes the scalable layers and/or sub-layers being decoded. In HEVC, an operation point is defined as bitstream created from another bitstream by operation of the sub-bitstream extraction process with the another bitstream, a target highest TemporalId, and a target layer identifier list as inputs.

The VPS of HEVC specifies layer sets and HRD parameters for these layer sets. A layer set may be used as the target layer identifier list in the sub-bitstream extraction process. In HEVC, a layer set may be defined as set of layers represented within a bitstream created from another bitstream by operation of the sub-bitstream extraction process with the another bitstream, the target highest TemporalId equal to 6, and the target layer identifier list equal to the layer identifier list associated with the layer set as inputs.

An output layer may be defined as a layer whose decoded pictures are output by the decoding process. The output layers may depend on which subset of the multi-layer bitstream is decoded. The pictures output by the decoding process may be further processed, e.g. a color space conversion from the YUV color space to RGB may be performed, and they may be displayed. However, further processing and/or displaying may be considered to be processes external of the decoder and/or the decoding process and might not take place.

In multi-layer video bitstreams, an operation point definition may include a consideration a target output layer set. For example, an operation point may be defined as a bitstream that is created from another bitstream by operation of the sub-bitstream extraction process with the another bitstream, a target highest temporal sub-layer (e.g. a target highest TemporalId), and a target layer identifier list as inputs, and that is associated with a set of output layers. Alternatively, another term, such as an output operation point, may be used when referring to an operation point and the associated set of output layers. For example, in MV-HEVC/SHVC, an output operation point may be defined as a bitstream that is created from an input bitstream by operation of the sub-bitstream extraction process with the input bitstream, a target highest TemporalId, and a target layer identifier list as inputs, and that is associated with a set of output layers.

An output layer set (OLS) may be defined as a set of layers consisting of the layers of one of the specified layer sets, where one or more layers in the set of layers are indicated to be output layers. An output layer may be defined as a layer of an output layer set that is output when the decoder and/or the HRD operates using the output layer set as the target output layer set. In MV-HEVC/SHVC, the variable TargetOlsIdx may specify which output layer set is the target output layer set by setting TargetOlsIdx equal to the index of the output layer set that is the target output layer set. A target output layer set may be defined as the output layer set for which the index is equal to TargetOlsIdx. TargetOlsIdxmay be set for example by the HRD and/or may be set by

external means, for example by a player or alike through an interface provided by the decoder. In MV-HEVC/SHVC, an output layer may be defined as a layer of an output layer set that is output when TargetOlsIdx is equal to the index of the output layer set.

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).

While a constant set of output layers suits well use cases and bitstreams where the highest layer stays unchanged in each access unit, they may not support use cases where the highest layer changes from one access unit to another. It has therefore been proposed that encoders can specify the use of alternative output layers within the bitstream and in response to the specified use of alternative output layers decoders output a decoded picture from an alternative output layer in the absence of a picture in an output layer within the same access unit. Several possibilities exist how to indicate alternative output layers. For example, as specified in HEVC, the alternative output layer set mechanism may be constrained to be used only for output layer sets containing only one output layer, and an output-layer-set-wise flag (alt_output_layer_flag[olsIdx] in HEVC) may be used for specifying that any direct or indirect reference layer of the output layer may serve as an alternative output layer for the output layer of the output layer set. When more than one alternative output layer is enabled to be used, it may be specified that the first direct or indirect inter-layer reference picture present in the access unit in descending layer identifier order down to the indicated minimum alternative output layer is output.

A first reason for frame-packed coding of multiview content having an inferior compression efficiency compared to multiview coding, such as MV-HEVC, is the lack of capability to perform inter-view sample prediction from a first constituent frame of a first view to a second constituent frame of a second view, the first and second constituent frames being frame-packed within the same frame. A second reason for frame-packed coding of multiview content having an inferior compression efficiency compared to multiview coding, such as MV-HEVC, is the lack of

capability to perform inter-view motion prediction from a first constituent frame of a first view to a second constituent frame of a second view, the first and second constituent frames being frame-packed within the same frame.

Below, many embodiments are described with reference to the term constituent frame. It needs to be understood that embodiments are not limited to be applied to constituent frames, but are generally applicable to any portions of a picture. For example, embodiments may be applicable to tiles or tile sets. In another example, embodiments are applicable to units that are smaller than a constituent frame as long as such a unit occupies a spatially contiguous area within a picture being coded or decoded. In an example embodiment, a constituent frame is partitioned into two or more spatially disjoint constituent frame partitions in a pre-processing step prior to encoding or as a part of encoding. A same picture may include constituent frame partitions from more than one constituent frame. Embodiments may be applied to e.g. for inter-view motion prediction between constituent frame partitions of different constituent frames frame-packed within the same picture.

In accordance with an embodiment, inter-view sample prediction between the frame-packed constituent frames may be enabled e.g. as follows.

In an embodiment, a first constituent frame representing a first view and a second constituent frame representing a second view of the same multiview (e.g. stereoscopic) content are frame-packed. Intra-block-copy prediction from the first constituent frame to the second constituent frame is enabled in encoding, but intra-block-copy prediction from the second constituent frame to the first constituent frame is disabled in encoding. Hence, inter-view sample prediction is essentially achieved by applying intra-block-copy prediction between constituent frames of different views. Compression efficiency is hence improved compared to coding of frame-packed multiview content without intra-block-copy prediction.

In an embodiment, it is indicated (e.g. by an encoder) that intra-block-copy prediction is applied only within a second constituent frame in a constrained manner such that no sample values of blocks succeeding, in a block scan order, a block collocated with the any first block in the first constituent frame, is used for intra-block-copy prediction within the second constituent frame of the current picture. The collocated block in the first constituent frame may be determined to be the block with the coordinates that are identical to those of the current block (being encoded or decoded) within the second constituent frame. Alternatively, a vertical and/or horizontal offset e.g. in units of coding tree units may be indicated by an encoder in a bitstream and/or decoded from a bitstream by a decoder, and the collocated block in the first constituent frame may be determined to be the block with sum of the vertical and/or horizontal coordinates of the current block (being encoded or decoded) in the second constituent frame and the vertical and/or horizontal offset,

respectively. It is noted that the collocated block defined in this embodiment for sample prediction from the first constituent frame to the second constituent frame using intra block copy is not necessarily the same as the collocated block potentially used for motion prediction from the first constituent frame to the second constituent frame, as described in other embodiments.

In an embodiment, the encoder is made aware of, e.g. through an application programming interface (API) or alike, of the maximum vertical and/or horizontal disparity between the constituent frames or the encoder analyzes the source content to discover the maximum vertical and/or horizontal disparity between the constituent frames or the encoder makes an assumption of the maximum vertical and/or horizontal disparity between the constituent frames. The encoder limits the search range of inter-block-copy prediction according to the maximum vertical and/or horizontal disparity. For example, the maximum vertical disparity may be assumed to be 0, and the search range is hence limited to be on the same row as the current block.

In accordance with an embodiment, inter-view motion prediction between the frame-packed constituent frames may be enabled e.g. as follows.

In the method, which is disclosed in Figure 7a as a flow diagram in accordance with an embodiment, a first constituent frame 601 of a first picture and a second constituent frame 602 of the first picture are obtained 701. The first constituent frame may be encoded. Inter prediction may be applied in the encoding of the first constituent frame. Also information of a motion field of the first constituent frame 601 may be obtained 702, for example as an outcome of encoding the first constituent frame. An indication that the motion field of the first constituent frame 601 is a source for motion vector prediction for the second constituent frame 602 may be formed and included in a bitstream 703. The first constituent frame of the first picture may be indicated e.g. by an encoder to be a source for motion vector prediction (a.k.a. collocated picture), e.g. HEVC TMVP or alike, for the second constituent frame 602 of the first picture. The second constituent frame may be encoded. At least a part of the second constituent frame may use motion prediction for which the source is the motion field of the first constituent frame. Respectively, in an embodiment an indication is decoded, e.g. by a decoder, that a first constituent frame of a first picture is indicated, e.g. by an encoder, to be a source for motion vector prediction, e.g. HEVC TMVP or alike, for a second constituent frame of the first picture.

Correspondingly, in a decoding method, which is disclosed in Figure 7b as a flow diagram in accordance with an embodiment, the indication that the motion field of the first constituent frame 601 is a source for motion vector prediction for the second constituent frame 602 may be decoded 711 and used when performing motion prediction of the second constituent frame 602 of the first picture. Also the first constituent frame 601 of the first picture and the second constituent frame

602 of the first picture may be decoded 712. Further, information of a motion field of the first constituent frame 601 may be decoded 713, e.g. as a part of decoding the first constituent frame. The decoding of at least a part of the second frame may comprise motion prediction for which the source is the motion field of the first constituent frame.

In an embodiment, a reference region of the first constituent frame of the first picture is indicated and/or a target region of the second constituent frame of the first picture is indicated. The reference region and/or the target region indicate the spatial correspondence of the motion vectors and may be used to compensate for the global or average disparity between the views, i.e. to achieve disparity-compensated inter-view motion prediction in frame-packed multiview video coding. A reference region indicates the region of the motion field of the first constituent frame that corresponds to the target region of the second constituent frame. In some cases a reference region may be allowed to be partially outside the first constituent frame, in which case the motion parameters outside the first constituent frame may be considered unavailable for motion prediction. Figure 6a illustrates the embodiment, where the motion field of constituent frame #1 601 of a first picture is used as a reference for motion vector prediction for constituent frame #2 602 of the first picture. The dotted rectangle 603 illustrates an example of a reference region for motion vector prediction for the constituent frame #1 601 of the first picture.

In different embodiments, the reference region may be considered to be similar or identical to the reference region in scalable video coding, as described earlier. In different embodiments, the target region may be considered to be similar or identical to the resampled reference region in scalable video coding, as described earlier e.g. with reference to Figure 6b.

While many embodiments are described with reference to the reference region and the target region residing in the same picture and hence enabling inter-view sample and/or motion prediction for frame-packed multiview content, it needs to be understood that embodiments may be applied alternatively or additionally in a manner that the reference region is indicated to reside in a first picture of a first layer and the target region is indicated to reside in a second picture of the first layer. In other words, embodiments may be used for spatially scalable sample prediction and/or spatially scalable motion prediction within a layer, from a reference region in a reference picture to a target region of the current picture being encoded or decoded. Spatially scalable sample prediction may be used for example when a first constituent frame or a first constituent frame partition in a reference picture is used as a source for sample prediction of but has a different width and/or height and/or location within the containing picture from a respective constituent frame or constituent frame partition in the current picture. Spatially scalable motion prediction may be used for example to compensate global motion, such as zooming in or out, camera dollying (i.e. camera transition), and/or camera turning.

An example embodiment for the case where the reference region is indicated to reside in a first picture of a first layer and the target region is indicated to reside in a second picture of the first layer is illustrated without loss of generality in Fig. 10a to 10c and in Fig. 11. In the example embodiment, an equirectangular panorama picture 1020 is vertically partitioned into three parts, the top part 1021, middle part 1022, and bottom part 1023, as illustrated in Fig. 10a, where the top part 1021 and bottom part 1023 may for example cover approximately or exactly 45 degrees vertical field of view, and the middle part 1022 may cover approximately or exactly 90 degrees vertical field of view. The top part 1021 and bottom part 1023 may be horizontally downsampled e.g. to half of the original width, as illustrated in Fig. 10b. The downsampled top part 1024 and the downsampled bottom part 1025 as well as the original middle part 1022 may be packed into a single frame 1026 for example as illustrated in Fig. 10c, although it needs to be understood that other packing arrangements can also be used, such as locating the middle part 1022 on the left of the downsampled top part 1024 and the downsampled bottom part 1025, and arranging the downsampled top part 1024 above the downsampled bottom part 1025. Each of the downsampled top part 1024, the downsampled bottom part 1025, and the middle part 1022 may be handled similarly to a constituent frame in different embodiments. A first picture 1030 is encoded with an input picture being frame-packed similarly or identically to what is illustrated in Fig. 10c. Likewise, a coded first picture 1030 is decoded into a decoded first picture being frame-packed similarly or identically to what is illustrated in Fig. 10c. A second picture 1040 is encoded and/or decoded with inter prediction from the first picture 1030. A reference region 1034 is selected in encoding and/or decoding to be the top part 1031, the bottom part 1033, and the middle part 1032 of the first picture 1030 for a target region 1044 that is the top part 1041, the bottom part 1043, and the middle part 1042 of the second picture 1040, respectively. Fig. 11 illustrates the reference region 1034 and the target region 1044 when predicting the top part 1041 of the second picture from the top part 1031 of the first picture. The reference region 1034 that is the top part 1031 of the first picture is upsampled horizontally to match the width of the target region 1044 that is the top part 1041 of the second picture, and the upsampled top part (i.e., the top part 1031 that is upsampled to the resolution of the top part 1041) is used as a reference for sample prediction 1050 of the target region 1044, as described in more details in other embodiments. Similarly, the reference region that is the bottom part 1033 of the first picture is upsampled horizontally to match the width of the target region that is the bottom part 1043 of the second picture, and the upsampled bottom part is used as a reference for sample prediction of the target region.

In an embodiment, the motion field of the reference region is remapped to match the spatial extents of the target region as a part of encoding and/or decoding. A horizontal remapping ratio and/or a vertical remapping ratio may be derived e.g. identically or similarly to a horizontal resampling ratio and/or a vertical resampling ratio, respectively, as described earlier with reference to Figure 6b. In an embodiment, the motion field of the reference region is resampled as a part of encoding

and/or decoding according to the remapping ratio(s), similarly to motion field mapping of SHVC. Basically, horizontal and vertical motion vector components are re-scaled according to the horizontal and vertical remapping ratios, respectively, and the motion field of the reference region is resampled to match a block grid of the target region. In another embodiment, horizontal and vertical motion vector components of the reference region are re-scaled according to the horizontal and vertical remapping ratios, when used as a source for motion vector prediction within the target region. Moreover, a function or formula to determine which block or location of the motion field of the reference region is accessed (as a part of encoding and/or decoding) for a particular target region block or location takes the remapping ratio into account. For example, if a horizontal or vertical location in the target region (relative to the top-left location of the resampled reference region) is locTar and the remapping ratio is remapRatio, the horizontal or vertical location (respectively) in the reference region (relative to the top-left lection where the reference region and target region match) is locTar / remapRatio.

In an embodiment, the picture comprising a reference region is considered as a source picture for inter-layer prediction in encoding and/or in decoding the current picture comprising a target region. In some embodiments the picture comprising the reference region may be the current picture, while in other embodiments the picture comprising the reference region may be a reference picture on a same layer as the current picture. It may be required that the reference region indicated for the current picture when it is used as a source picture for inter-layer prediction does not overlap with the target region of the current picture to which the inter-layer prediction from the reference region may be performed. Alternatively, it may be required that no source block for inter-layer sample or motion prediction from the reference region in the current picture overlaps with the corresponding target block in the target region, the target block being the block being encoded and/or decoded with reference to the source block(s) for inter-layer sample or motion prediction. In addition, it may be required that the reference region precedes in encoding order and/or in decoding order that target region, or that all source block(s) for inter-layer sample or motion prediction from the reference region precede the corresponding target block of the target region in encoding order and/or in decoding order. When the reference region resides in a reference picture (other than the current picture) of the current picture, the reference picture may be temporarily treated as a long-term reference picture for sample and/or motion prediction of the target region in some embodiments, while in other embodiments the reference picture may be treated according to its present marking as a short-term or long-term reference picture for sample and/or motion prediction of the target region.

In an embodiment, an encoder indicates with one or more indications in the bitstream whether the picture comprising the reference region is considered as a source picture for inter-layer prediction or as a reference picture for inter prediction in encoding and/or in decoding the current picture

comprising the target region. In an embodiment, a decoder decodes from one or more indications in the bitstream whether the picture comprising the reference region is considered as a source picture for inter-layer prediction or as a reference picture for inter prediction in encoding and/or in decoding the current picture comprising the target region.

In an embodiment, the reference region and/or the target region are indicated by an encoder with reference layer location offset syntax (e.g. similar or identical to the reference layer location offset syntax of HEVC) or alike, and likewise decoded by a decoder from a reference layer location offset syntax or alike. In an embodiment, the reference layer location offset syntax or alike comprises a layer identifier for identifying the reference layer. When the reference region resides in the current picture, a layer identifier of the reference layer location offset syntax applicable for the current picture is set, by an encoder, equal to the layer identifier of the current picture, to indicate that the current picture is considered as a source picture for inter-layer prediction in encoding and/or in decoding the current picture. In an embodiment, a decoder concludes from the reference layer location offset syntax applicable for the current picture that has a layer identifier equal to the layer identifier of the current picture that inter-layer prediction is in use when referencing the current picture. In an embodiment, in the absence of such reference layer location offset being applicable to the current picture, the decoder concludes that inter prediction is in use when referencing the current picture in prediction.

In an embodiment, a first constituent frame is determined, e.g. as a part of encoding or as a part of pre-processing prior to encoding, to have different width and/or height than those of a second constituent frame of the same picture. The reference region of the first constituent frame to be used for prediction of the target region of the second constituent frame may therefore have different spatial extents.

In an embodiment, a first constituent frame of a reference picture on a first layer is determined, e.g. as a part of encoding or as a part of pre-processing prior to encoding, to have different width and/or height than those of a second constituent frame of the current picture (being encoded or decoded) on the first layer. The second constituent frame may be predicted from the first constituent frame in encoding and/or decoding. The reference region of the first constituent frame to be used for prediction of the target region of the second constituent frame may therefore have different spatial extents.

In an embodiment applicable to embodiments of the previous two paragraphs, the reference region is resampled (upsampled or downsampled) to be used as a reference for sample prediction of the target region in encoding and/or in decoding, similarly to inter-layer sample prediction of spatial scalable video coding, such as spatial scalability in SHVC. Horizontal and/or vertical resampling

ratios may be derived from the widths of the reference and target regions and the heights of the reference and target regions, respectively. For example, the resampled reference region may be considered a reference picture for the target region, similarly to an inter-layer reference picture for the target region. In an embodiment, the motion field of the reference region is mapped using motion field mapping (MFM) or alike in encoding and/or in decoding. The mapped motion field is used in motion prediction as a source for motion vector prediction of the second constituent frame in encoding and/or in decoding, similarly to inter-layer motion prediction. For example, when the current picture is indicated to be a collocated picture for TMVP or alike, the mapped motion field is used as a source for temporal motion vector predictor(s).

In an embodiment, the reference region and the target region are concluded to have the same spatial extents in decoding, and the reference region and/or its motion field are used as such in sample prediction (e.g. inter-layer sample prediction) and/or in motion prediction (e.g. inter-layer motion prediction), respectively.

In an embodiment, the motion vector field of the current picture may be accessed from an indicated reference position, which is treated like a collocated block for TMVP or alike. The reference position may have the same horizontal and vertical coordinates relative to the reference region as those of current block being encoded or decoded relative to the target region.

In an embodiment, the motion vectors of the reference region of the first constituent frame may be copied onto the location of the motion vectors of the target region of the second constituent frame within the same motion vector field. The collocated block for TMVP or alike may be selected, for example, similarly or identically to selecting the collocated block from any other picture chosen to be the source picture for TMVP or alike. For example, the collocated block may be the current block being encoded or decoded. When the current block has been encoded or decoded, the final motion vector of the current block may be placed onto the motion vector field, replacing the earlier motion vector that had been copied from the reference region of the first constituent frame.

In an embodiment, it is pre-defined e.g. in a coding standard or indicated by an encoder in a bitstream or decoded by a decoder from a bitstream which order of potential temporal candidate locations is used when the collocated picture is the current picture. The order may e.g. comprise only C_1 as indicated in Fig. 5b. The order of potential temporal candidate locations used when the collocated picture is the current picture may differ from the order used when the collocated picture is not the current picture.

In an embodiment, the copying may take place block by block in the target region as an initial step in encoding or decoding a block in the target region. The encoding or decoding of the target region

may be comprise an initial step before encoding or decoding any blocks of the target region to initialize all blocks or locations or motion vectors of the motion field of the target region to be unavailable (e.g. mark the blocks to be intra coded). Consequently, if the order (C_0, C_1) (as indicated in Fig. 5b) is used in selecting the collocated block for TMVP or alike, the motion vector of C_0 is unavailable for the encoding or decoding of the current block of the target region and the motion vector may be taken from the current block of the target region, which has been copied from the collocated block (within the reference region).

In an embodiment, the copying may take place on block row basis, where a block may be similar or identical to a CTU as defined in HEVC. In other words, the motion vectors of a block row in the reference region are copied to the respective block row in the target region. The (de)coding order of the block row is such that the (de)coding of the block row in the reference region precedes the (de)coding of the respective block row in the target region.

In an embodiment, the copying may take place on tile row basis. In other words, the motion vectors of a tile row in the reference region are copied to the respective tile row in the target region. The (de)coding order of the tile row is such that the (de)coding of the tile row in the reference region precedes the (de)coding of the respective tile row in the target region.

In an embodiment, the copying may take place for the entire reference region, essentially as a single operation. The (de)coding order of constituent frames is such that the (de)coding of the first constituent frame precedes the (de)coding of the second constituent frame.

In an embodiment, the encoder may select or may be configured to use tile partitioning such that a first tile set covers the first constituent frame entirely and includes no data from the second constituent frame, and a second tile set covers the second constituent frame entirely and includes no data from the first constituent frame.

In an embodiment, the encoder encodes the first tile set as a motion-constrained tile set and the second tile set as another motion-constrained tile set. The encoder may indicate that the encoding of the first and second tile sets have been performed in a motion-constrained manner e.g. with an SEI message, such as the temporal motion-constrained tile sets SEI message of HEVC.

In an embodiment, a decoder, a player, or alike decodes an indication e.g. from an SEI message, such as the temporal motion-constrained tile sets SEI message of HEVC, that the encoding of the first and second tile sets have been performed in a motion-constrained manner. In an embodiment, the decoder, the player, or alike determines on the basis of the decoded indication to decode the

first tile set with a first decoder instance (e.g. a first decoding process) and the second tile set with a second decoder instance (e.g. a second decoding process).

In an embodiment, the encoder encodes the first tile set as a motion-constrained tile set. In an embodiment, the encoder encodes the second tile set as a motion-constrained tile set except when referencing to the current picture in intra-block-copy prediction or alike (as described in another embodiment). When referencing to the current picture in intra-block-copy prediction or alike in blocks of the second tile set, samples outside the first tile set are not used in sample prediction (when sample prediction is in use) and/or motion information outside the first tile set is not used in motion prediction. In an embodiment, the encoder indicates in a bitstream, e.g. in an SEI message, that the motion-constrained tile sets have been used in encoding as described in this paragraph.

In an embodiment, a decoder, a player, or alike decodes an indication e.g. from an SEI message, such as the temporal motion-constrained tile sets SEI message of HEVC, that the encoding of the first and second tile sets have been performed in a motion-constrained manner except when referencing to the current picture in intra-block-copy prediction or alike (as described in another embodiment). In an embodiment, the decoder, the player, or alike determines on the basis of the decoded indication to decode the first tile set with a first decoder instance (e.g. a first decoding process) and the second tile set with a second decoder instance (e.g. a second decoding process). The first decoder instance is configured to decode the first tile set prior to the second decoder instance decoding the second tile set of the same picture. However, the first decoder instance may decode the first tile set of a succeeding picture (following a current picture) in decoding order, while the second decoder instance decodes the second tile set of the current picture.

In an embodiment, the encoder may select tile structuring in a manner that motion vectors of the first constituent frame accessed in the decoding of the second constituent frame are available.

HEVC TMVP is such that the default location in the collocated picture to pick the motion vector is bottom-right of the location of the current PU (being encoded or decoded). Only if no motion vector is available in the default TMVP candidate location, e.g. when the corresponding block is intra-coded, the (spatially collocating) location of the current PU is considered. The default location for the TMVP candidate may be considered to cause a kind of a shift from the actual disparity between the constituent frames of different views towards the bottom-right corner.

In an embodiment, the encoder or the decoder may pre-compensate the choice of the TMVP default location in the copying of the motion field of the first constituent frame or the reference region to the motion field of the second constituent frame or the target region. For example, the encoder or

the decoder may pre-compensate by 8 luma samples both horizontally and vertically, i.e. select a reference region or such that is moved by 8 luma samples horizontally and vertically relative to the inferred or indicated first constituent frame or reference region.

In an embodiment, the encoder may pre-compensate the choice of the TMVP default location in the reference region selection. For example, the encoder may pre-compensate by 8 luma samples both horizontally and vertically, i.e. "move" the reference region selected on the basis of the global disparity or such by 8 luma samples horizontally and vertically.

In an embodiment, it is concluded, e.g. in a decoder, that a first constituent frame can be decoded in parallel with a second constituent frame . For example, it can be concluded that there is no sample prediction between the first and second constituent frames (within the same picture) or that parallel decoding is enabled, e.g. based on the in-picture motion-constrained tile set or alike. It may further be concluded, e.g. based on the reference region and the target region, what is the offset e.g. in terms of CTUs between the decoding of the constituent frames. In some cases, TMVP or alike may require access to motion vectors of blocks on the right and below the collocated block, which can be taken into account in the choosing an offset e.g. in terms of CTUs between decoding the first constituent frame and the second constituent frame.

In an embodiment, an encoder may choose slice boundaries in a manner, e.g. a slice per tile row, that motion vector prediction from a tile row below the current tile row is disabled, when motion prediction from the first constituent frame to the second constituent frame of the same picture is performed in encoding and/or in decoding. An encoder may indicate such a constraint e.g. in VUI.

In an embodiment, an encoder may select the motion vector predictor in a manner that the TMVP candidate (or alike) from a tile row below the current tile row is not used. An encoder may indicate such a constraint e.g. in VUI. A decoder may decode the indication and conclude that an offset between the decoding of the first constituent frame and the second constituent frame can be less than an entire tile row, such as one CTU or alike. The offset may be e.g. the number of blocks that the decoding of the reference region needs to be ahead of the decoding the target region in order to have all information of the reference region available for prediction of the target region.

In the following, some examples of signalling options are disclosed.

In an embodiment, signalling similar to reference layer location offsets may be used to indicate the reference region and the target region. The granularity of the offsets may be selected according to the smallest block size in the motion vector field. The signalling may be specific to a tile set, a constituent frame, or a slice (enclosing a tile set or a constituent frame, or parts thereof). The

reference region may be indicated similarly or identically to the reference region offsets of the reference layer location offsets, as described earlier. The target region may be indicated similarly or identically to the scaled reference layer location offsets of the reference layer location offsets, as described earlier.

In an embodiment, the reference layer location offset syntax or alike is used for the indication of the reference and target regions for different embodiments as well as indicating spatial correspondence of pictures in different layers. One or more of the following is used in or in connection with the reference layer location offset syntax or alike:

- The layer on which the target region resides is indicated or inferred (e.g. to be the layer of the current picture being encoded or decoded). The layer on which the reference region resides is indicated e.g. by its layer identifier value, such as nuh_layer_id in HEVC. If the layer of the target region is the same as the layer of the reference region and that layer is the same as the layer of the current picture being encoded or decoded, it may be concluded (e.g. as part of decoding) that another embodiment applies or other embodiment(s) apply for using sample prediction and/or motion prediction from the first constituent frame of the current picture being encoded or decoded to the second constituent frame of the current picture being encoded or decoded.
- An indicated first picture within the decoded picture buffer (DPB) or within the reference picture set (RPS) applied to the current picture is indicated to comprise the reference region. The indicated first picture in the DPB or the RPS may be indicated e.g. through its picture order count difference relative to the current picture or through its index within the DPB or the RPS. If the indicated first picture in the DPB or the RPS is the current picture being encoded or decoded, it may be concluded (e.g. as part of decoding) that another embodiment applies or other embodiment(s) apply for using sample prediction and/or motion prediction from the first constituent frame of the current picture being encoded or decoded to the second constituent frame of the current picture being encoded or decoded. If the indicated first picture in the DPB or the RPS is different from the current picture being encoded or decoded (but on the same layer than the current picture), it may be concluded (e.g. as part of decoding) that another embodiment applies or other embodiment(s) apply for using sample prediction and/or motion prediction from the first constituent frame of the indicated picture to the second constituent frame of the current picture being encoded or decoded.
- An indicated first picture within a reference picture list (e.g. reference picture list 0 of HEVC) is indicated to comprise the reference region. The indicated first picture in the reference picture list may be indicated e.g. through its index within the reference picture list. If the indicated first picture in the reference picture list is the current picture being encoded or decoded, it may be concluded (e.g. as part of decoding) that another

embodiment applies or other embodiment(s) apply for using sample prediction and/or motion prediction from the first constituent frame of the current picture being encoded or decoded to the second constituent frame of the current picture being encoded or decoded. If the indicated first picture in the reference picture list is different from the current picture being encoded or decoded (but on the same layer than the current picture), it may be concluded (e.g. as part of decoding) that another embodiment applies or other embodiment(s) apply for using sample prediction and/or motion prediction from the first constituent frame of the indicated first picture to the second constituent frame of the current picture being encoded or decoded.

In an embodiment, when an encoder uses or is prepared to use one or more embodiments, it indicates in a sequence-level syntax structure, such as SPS or VPS, that features of the one or more embodiments are turned on in encoding or decoding or may be turned on in encoding or decoding. In an embodiment, when an encoder indicates in a sequence-level syntax structure that features of the one or more embodiments may be turned on in encoding or decoding, the encoder indicates e.g. in a picture-level syntax structure, such as PPS, or in a slice-level syntax structure, such as slice header or slice segment header, that features of the one or more embodiments are turned on in encoding or decoding.

In an embodiment, an encoder indicates reference layer location offsets or alike in a syntax structure applicable or corresponding to the target region. For example, the encoder may enclose the target region into a tile set and encode the tile set in one or more slices. The encoder may choose not to include any other tiles in the one or more slices than those enclosing the target region or the second constituent frame; or the encoder may choose to cover the second constituent frame with the tile set, exclude data of any other constituent frames from the tile set, and encode the tile set in one or more slices excluding any tiles from other tile sets. In this example, the encoder includes reference layer location offsets or alike in the slice header(s) or slice segment header(s) of the one or more slices. Correspondingly, in an embodiment, the decoder decodes reference layer location offsets or alike from a syntax structure applicable or corresponding to the target region. For example, the decoder may decode reference layer location offsets or alike from a slice header or a slice segment header. The slice corresponding to the header may for example contain a tile set enclosing the target region or the second constituent frame.

In an embodiment continuing the embodiment in the previous paragraph, the signalling of reference location offsets or alike is accompanied with the signalling identifying a picture containing the reference region, e.g. according to an embodiment further above. When the signalling indicates the current picture being encoded or decoded, it may be concluded (e.g. as part of decoding) that another embodiment applies or other embodiment(s) apply for using sample

prediction and/or motion prediction from the first constituent frame of the current picture being encoded or decoded to the area corresponding to the syntax structure containing the reference layer location offsets or alike, such as the slice in the example where the signalling is included in a slice segment header or a slice header. When the signalling indicates a previous picture in decoding order, such as a source picture for inter-layer prediction, it may be concluded that the reference region is taken from that picture. The reference region may be resampled as explained earlier, based on the sampling ratio(s) derived from the target region and the reference region, for sample prediction. The motion field of the reference region may be remapped as explained earlier, based on the remapping ratio(s) derived from the target region and the reference region, for motion prediction.

In an embodiment, one or more horizontal duplication offsets and/or one or more vertical duplication offsets may be indicated by an encoder and decoded by a decoder. The granularity of the offsets may be selected according to the smallest block size in the motion vector field. When a motion vector is reconstructed in an encoder or decoded in a decoder, in addition to the location of the motion vector (i.e. the location of the current encoded or decoded block), the motion vector is copied in the motion field also the location indicated by the duplication offset(s). In an embodiment, the signalling is picture-based; if the location indicated by the duplication offset(s) is outside of the picture boundaries, the copying is not performed. In an embodiment, the signalling may be specific to a tile set, a constituent frame, or a slice (enclosing a tile set or a constituent frame, or parts thereof).

Texture-plus-depth frame packing

In an embodiment, a texture frame of a first view may be packed into the same picture with a depth frame of the first view (e.g. by a pre-processing unit prior to encoding). In an embodiment, the packing may be done in a manner that enables the prediction of texture motion vectors from the depth motion vectors; in another embodiment, the packing may be done in a manner that enables the prediction of depth motion vectors from texture motion vectors.

Disparity-based motion field mapping in texture-plus-depth frame packing

In an embodiment, a first texture frame of a first view, a second texture frame of a second view, and a depth frame may be packed into the same picture (e.g. by a pre-processing unit prior to encoding). The depth frame may e.g. represent the second view. The depth frame may e.g. contain disparity values between the first the first texture frame and the second texture frame or such disparity values may be generated on the basis of the depth frame.

In an embodiment, horizontal and/or vertical offset(s) to derive a collocated block in the first view for a first block being encoded or decoded in the second view may be derived on the basis of disparity value(s) associated with the first block. The collocated block may be used as source to obtain a TMVP candidate or alike in the motion vector prediction for the first block.

In an embodiment, horizontal and/or vertical offset(s) to derive a collocated block in the first view for a first block being encoded or decoded in the second view may be derived on the basis of disparity value(s) associated with the first block. The motion vector of the collocated block may be stored as a motion vector for the first block in the motion vector field of the current picture. The motion vector of the first block may be used as source to obtain a TMVP candidate or alike in the motion vector prediction for the first block. When the motion vector prediction for the first block has been performed and the differential motion vector, if any, summed up with the motion vector predictor to obtain a final motion vector for the first block, the final motion vector may be stored for the first block in the motion vector field of the current picture, replacing the previous motion vector for the first block.

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

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.

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.

Figure 9 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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

In the above, some embodiments have been described with reference to the term constituent frame. It needs to be understood that a constituent frame may represent an input frame or an input field, where 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.

In the above, some embodiments have been described with reference to a first constituent frame of a first view and a second constituent frame of a second view that is different from the first view. It needs to be understood that embodiments are not limited to two constituent frames originating from different views, but generally the constituent frames may represent the same view. For example, the first constituent frame and the second constituent frame may represent the same original frame but have different picture quality and/or spatial resolution.

In the above, some embodiments have been described with reference to a first constituent frame and a second constituent frame. It needs to be understood that embodiments are not limited to two constituent frames per picture but also apply generally to more than two constituent frames per picture.

In the above, some embodiments have been described with reference to the term motion vector. It needs to be understood that the embodiments also apply with reference to the more general term motion parameters instead of the term motion vector. Motion parameters may for example comprise a motion vector and a reference index, as described above.

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.

In the above, some embodiments have been described in relation to SHVC and/or MV-HEVC and/or the terminology used in SHVC and/or MV-HEVC. 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. It also needs to be understood that some embodiments similarly apply to other single-layer codecs and coding formats and other terminology with equivalency or similarity to the terms used in the above-described embodiments.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

In the following some examples will be provided.

In accordance with a first example there is provided a method comprising:
decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

In accordance with an embodiment of the method the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

In accordance with an embodiment the method comprises at least one of:
decoding information of the reference region of the first picture;
decoding information of the target region of the first picture.

In accordance with an embodiment the method further comprises:
using at least one of the reference region and the target region as an indication of the spatial correspondence of the motion vectors.

In accordance with an embodiment the method further comprises:
concluding that the decoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the method said concluding comprises decoding an indication from the bitstream that the decoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the method said indication is the information of the reference region of the first picture or the information of the target region of the first picture.

In accordance with an embodiment the method further comprises at least one of:
resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the method further comprises:
obtaining an indication of a reference position; and
accessing the motion vector field from the reference position.
In accordance with an embodiment the method further comprises:
copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

In accordance with an embodiment the method further comprises at least one of:
copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;
copying the motion vectors on block row basis;
copying the motion vectors on tile row basis;

copying the motion vectors of the entire reference region.

In accordance with an embodiment of the method a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the method comprises:
decoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;
decoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

In accordance with a second example there is provided an apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

In accordance with an embodiment of the apparatus the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
decoding information of the reference region of the first picture;
decoding information of the target region of the first picture.
In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
using at least one of the reference region and the target region as an indication of the spatial correspondence of the motion vectors.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
concluding that the decoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the apparatus said concluding comprises decoding an indication from the bitstream that the decoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the apparatus said indication is the information of the reference region of the first picture or the information of the target region of the first picture.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
obtaining an indication of a reference position; and
accessing the motion vector field from the reference position.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;
copying the motion vectors on block row basis;
copying the motion vectors on tile row basis;
copying the motion vectors of the entire reference region.

In accordance with an embodiment of the apparatus a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the apparatus comprises code causing the apparatus to perform at least:
decoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;

decoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

In accordance with a third example there is provided 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:
decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
decoding the reference region of the picture; and
decoding the target region of the picture.

In accordance with a fourth example there is provided an apparatus comprising:
means for decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;
means for decoding the reference region of the picture; and
means for decoding the target region of the picture.

In accordance with a fifth example there is provided a method comprising:
determining a reference region of a picture;
determining a target region of the picture;
obtaining information of a motion field of the reference region; and
including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

In accordance with an embodiment of the method the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

In accordance with an embodiment the method further comprises at least one of:
encoding information of the reference region of the first picture;
encoding information of the target region of the first picture.

In accordance with an embodiment the method further comprises:
using at least one of the reference region and the target region as an indication of the spatial correspondence of the motion vectors.

In accordance with an embodiment the method further comprises:
using the encoded reference region as a source for inter-layer prediction of the target region.

In accordance with an embodiment the method further comprises:
encoding an indication into the bitstream that the encoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the method said indication is the information of the reference region of the first picture or information of the target region of the first picture.

In accordance with an embodiment the method further comprises at least one of:
resampling the encoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the method further comprises:
obtaining the motion vector field from a reference position; and
forming an indication of the reference position.
In accordance with an embodiment the method further comprises:
copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

In accordance with an embodiment the method further comprises at least one of:
copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;
copying the motion vectors on block row basis;
copying the motion vectors on tile row basis;
copying the motion vectors of the entire reference region.

In accordance with an embodiment of the method a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the method comprises:
encoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;
encoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that

the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

In accordance with a sixth example there is provided an apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
determining a reference region of a picture;
determining a target region of the picture;
obtaining information of a motion field of the reference region; and
including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

In accordance with an embodiment of the apparatus the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
encoding information of the reference region of the first picture;
encoding information of the target region of the first picture.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
using at least one of the reference region and the target region as an indication of the spatial correspondence of the motion vectors.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
using the encoded reference region as a source for inter-layer prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
encoding an indication into the bitstream that the encoded reference region is used as a source for inter-layer prediction of the target region.

In accordance with an embodiment of the apparatus said indication is the information of the reference region of the first picture or information of the target region of the first picture.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
resampling the encoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
obtaining the motion vector field from a reference position; and
forming an indication of the reference position.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least one of:
copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;
copying the motion vectors on block row basis;
copying the motion vectors on tile row basis;
copying the motion vectors of the entire reference region.

In accordance with an embodiment of the apparatus a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the apparatus comprises code causing the apparatus to perform at least:
encoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;
encoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

In accordance with a seventh example there is provided 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:
determining a reference region of a picture;
determining a target region of the picture;
obtaining information of a motion field of the reference region; and
including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

In accordance with a eighth example there is provided an apparatus comprising:
means for determining a reference region of a picture;
means for determining a target region of the picture;
means for obtaining information of a motion field of the reference region; and
means for including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

In accordance with a ninth example there is provided a method comprising:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
decoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with an embodiment the sample prediction further comprises:
resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the method further comprises:
concluding that first extents match to the second extents; and
using the reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the motion prediction further comprises:
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the method further comprises:
concluding that first extents match to the second extents; and
using the reference region as a reference for motion prediction of the target region.

In accordance with a tenth example there is provided an apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
decoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with an embodiment the sample prediction further comprises code causing the apparatus to perform at least:
resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
concluding that first extents match to the second extents; and
using the reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the motion prediction further comprises code causing the apparatus to perform at least:
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
concluding that first extents match to the second extents; and
using the reference region as a reference for motion prediction of the target region.

In accordance with an eleventh example there is provided a computer readable storage medium comprising code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:
decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
decoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with a twelfth example there is provided an apparatus comprising:
means for decoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
means for decoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
means for decoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with a thirteenth example there is provided a method comprising:
encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
encoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with an embodiment the sample prediction further comprises:
resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the method further comprises:
concluding that first extents match to the second extents; and
using the reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the motion prediction further comprises:
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the method further comprises:
concluding that first extents match to the second extents; and
using the reference region as a reference for motion prediction of the target region.

In accordance with a fourteenth example there is provided an apparatus comprising at least one processor and at least one memory, said at least one memory stored with code thereon, which when executed by said at least one processor, causes the apparatus to perform at least:
encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;
encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;
encoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with an embodiment the sample prediction further comprises code causing the apparatus to perform at least:
resampling the encoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:
concluding that first extents match to the second extents; and
using the reference region as a reference for sample prediction of the target region.

In accordance with an embodiment the motion prediction further comprises code causing the apparatus to perform at least:
mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

In accordance with an embodiment the apparatus further comprises code causing the apparatus to perform at least:

concluding that first extents match to the second extents; and

using the reference region as a reference for motion prediction of the target region.

In accordance with an eleventh example there is provided a computer readable storage medium comprising code for use by an apparatus, which when executed by a processor, causes the apparatus to perform:

encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;

encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;

encoding the target region using at least one of sample prediction and motion prediction from the reference region.

In accordance with a twelfth example there is provided an apparatus comprising:

means for encoding a first location and first extents of a reference region and a first identification of a first picture comprising the reference region;

means for encoding a second location and second extents of a target region of a second picture, the first picture and the second picture residing in the same layer, and the first picture being either the same as the second picture or a reference picture for the second picture;

means for encoding the target region using at least one of sample prediction and motion prediction from the reference region.

CLAIMS:

1. A method comprising:

decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;

decoding the reference region of the picture; and

decoding the target region of the picture.

2. A method according to claim 1 wherein the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

3. A method according to claim 1 comprising at least one of:

decoding information of the reference region of the first picture;

decoding information of the target region of the first picture.

4. A method according to claim 1 further comprising:

using at least one of the reference region and the target region as an indication of the spatial correspondence of the motion vectors.

5. A method according to claim 1 further comprising:

decoding an indication from the bitstream that the decoded reference region is used as a source for inter-layer prediction of the target region; and

concluding that the decoded reference region is used as the source for inter-layer prediction of the target region.

6. A method according to claim 5 wherein said indication is the information of the reference region of the first picture or the information of the target region of the first picture.

7. A method according to claim 1 further comprising at least one of:

resampling the decoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and

mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

8. A method according to claim 1 comprising:

obtaining an indication of a reference position; and

accessing the motion vector field from the reference position.

9. A method according to claim 1 comprising:

copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

10. A method according to claim 9 comprising at least one of:

copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;

copying the motion vectors on block row basis;

copying the motion vectors on tile row basis;

copying the motion vectors of the entire reference region.

11. A method according to claim 1 wherein a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the method comprises:

decoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;

decoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

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

decoding information that a motion vector field of a reference region of a picture is a source for motion vector prediction for a target region of the picture;

decoding the reference region of the picture; and

decoding the target region of the picture.

13. An apparatus according to claim 12, wherein the apparatus is further configured to perform the method as claimed in any of the claims 2 to 11.

14. A method comprising:

determining a reference region of a picture;

determining a target region of the picture;

obtaining information of a motion field of the reference region; and

including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

15. A method according to claim 14 wherein the reference region comprises a first constituent frame and the target region comprises a second constituent frame.

16. A method according to claim 14 further comprising:

using the encoded reference region as a source for inter-layer prediction of the target region.

17. A method according to claim 14 further comprising at least one of:

resampling the encoded reference region into a resampled reference region, and using the resampled reference region as a reference for sample prediction of the target region; and

mapping the motion vector field of the reference region into a mapped motion vector field of the reference region, and using the mapped motion vector field of the reference region for motion prediction of the target region.

18. A method according to claim 14 comprising:

obtaining the motion vector field from a reference position; and

forming an indication of the reference position.

19. A method according to claim 14 comprising:

copying motion vectors of the reference region of the first constituent frame onto a location of motion vectors of the target region of the second constituent frame.

20. A method according to claim 19 comprising at least one of:

copying the motion vectors block by block in the target region as an initial step in decoding a block in the target region;

copying the motion vectors on block row basis;

copying the motion vectors on tile row basis;

copying the motion vectors of the entire reference region.

21. A method according to claim 14 wherein a first tile set encloses the reference region and a second tile set encloses the target region, the first tile set and the second tile set being non-overlapping, and the method comprises:

encoding a first indication that motion vectors of the first tile set are constrained so that no sample values outside the first tile set are used in prediction of the first tile set;

encoding a second indication that motion vectors of the second tile set are constrained so that no sample values outside the second tile set are used in prediction of the second tile set, except that the motion vectors of the second tile set may refer to sample values outside the second tile set within the first picture.

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

determining a reference region of a picture;

determining a target region of the picture;

obtaining information of a motion field of the reference region; and

including an indication into a bitstream that a motion vector field of the reference region of the picture is a source for motion vector prediction for the target region of the picture.

23. An apparatus according to claim 22 wherein the apparatus is further configured to perform the method as claimed in any of the claims 15 to 21.

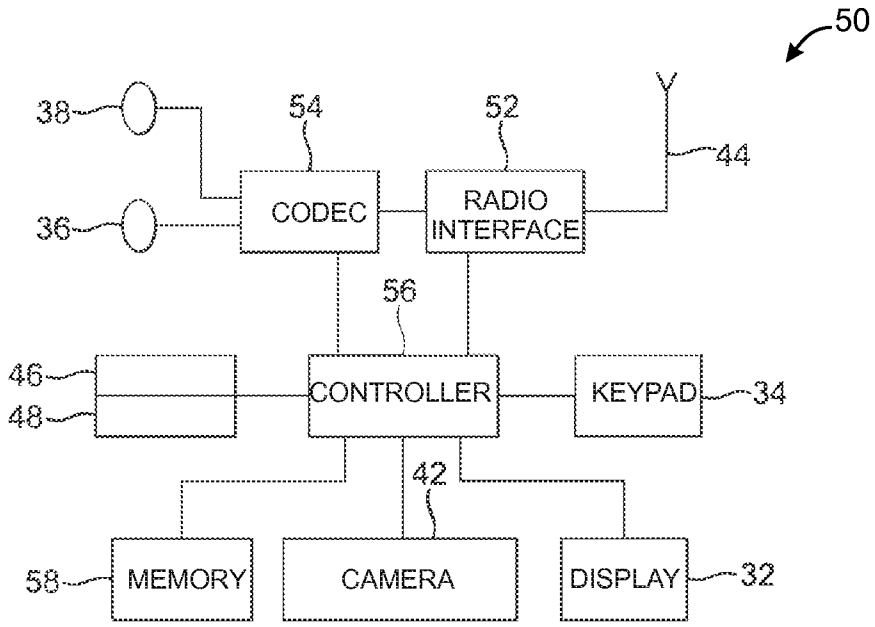

50
38
54
52
44
36
CODEC
RADIO INTERFACE
56
46
48
CONTROLLER
KEYPAD
34
42
58
MEMORY
CAMERA
DISPLAY
32

Fig. 1

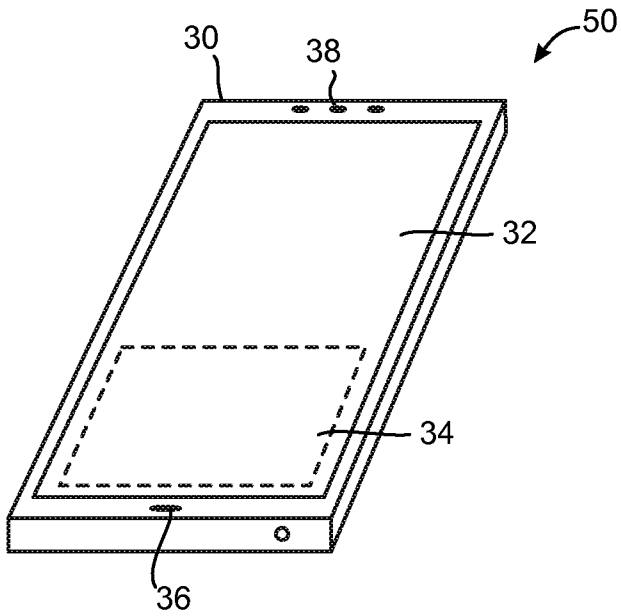

50
30
38
32
34
36

Fig. 2

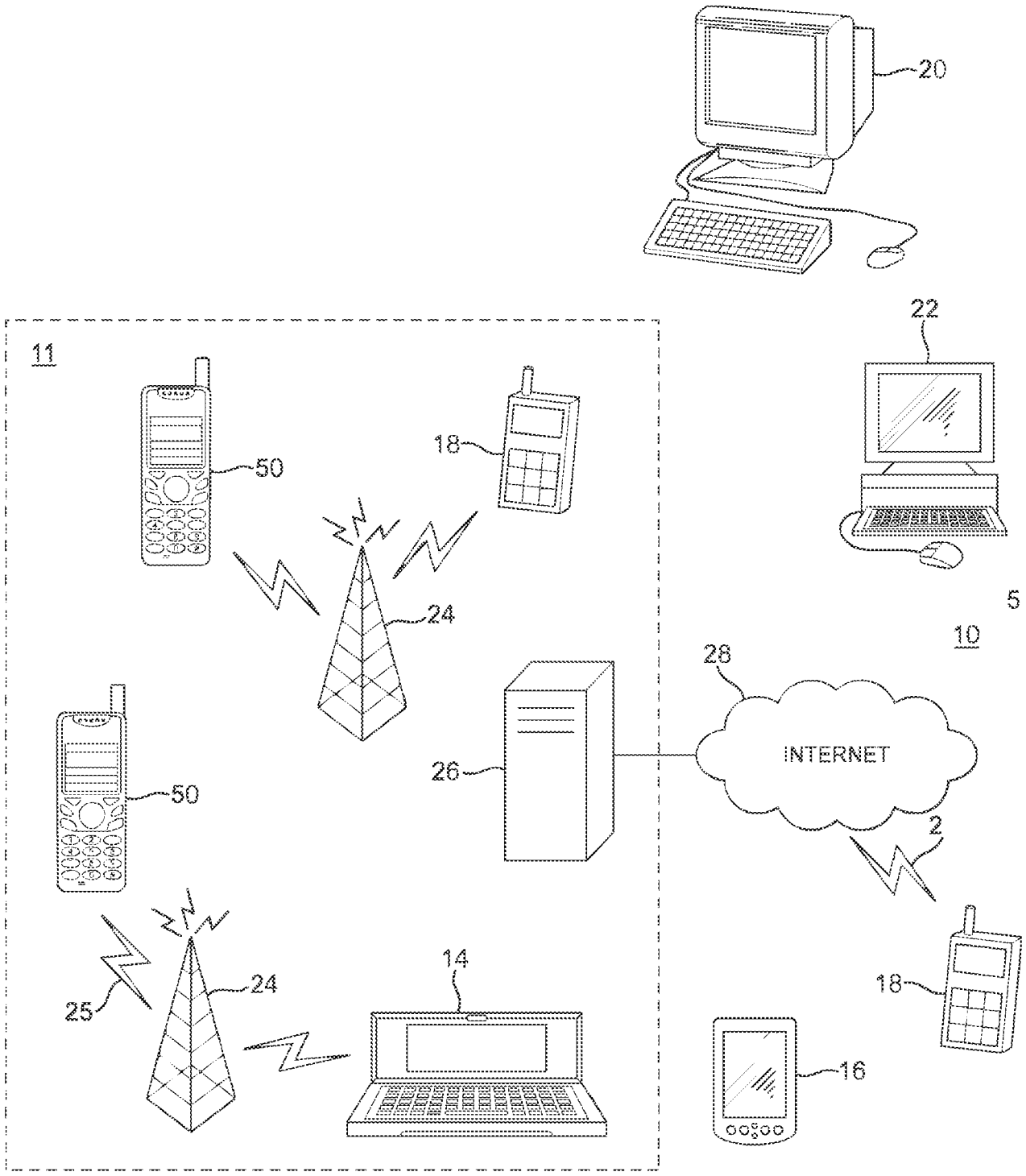


Fig. 3

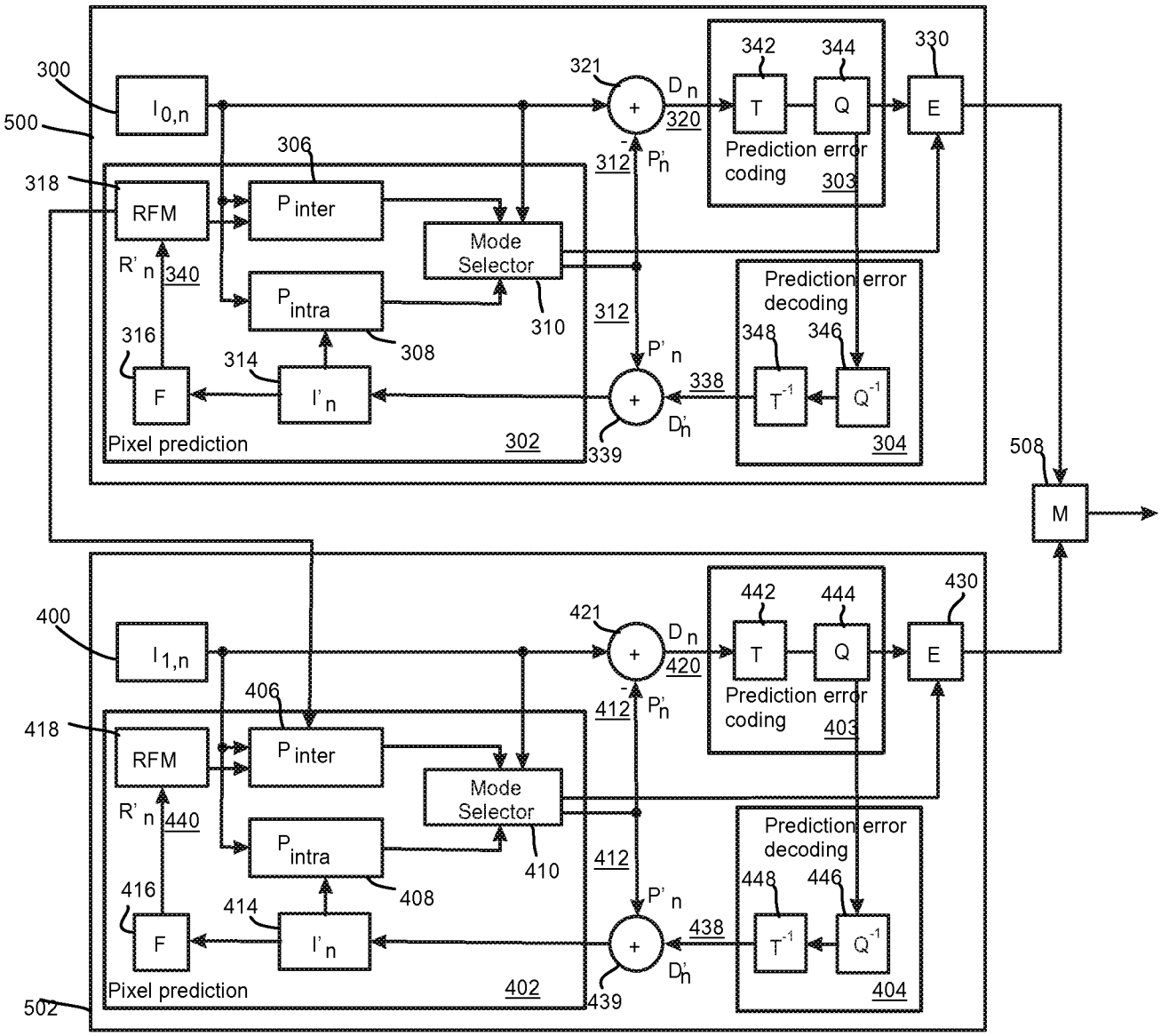

300
500
I 0,n
321
D n
320
342
344
330
T
Q
E
Prediction error coding
303
306
318
RFM
P inter
Mode Selector
R' n 340
P intra
310
312
P' n
316
314
308
F
I' n
338
Prediction error decoding
348
346
T-1
Q-1
304
Pixel prediction
302
339
D'n
508
M
400
I 1,n
421
D n
420
442
444
430
T
Q
E
Prediction error coding
403
406
418
RFM
P inter
Mode Selector
R' n 440
P intra
410
412
P' n
416
414
408
F
I' n
438
Prediction error decoding
448
446
T-1
Q-1
404
Pixel prediction
402
439
D'n
502

Fig. 4

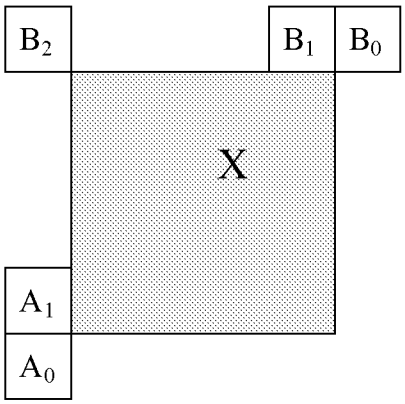


Fig.5a

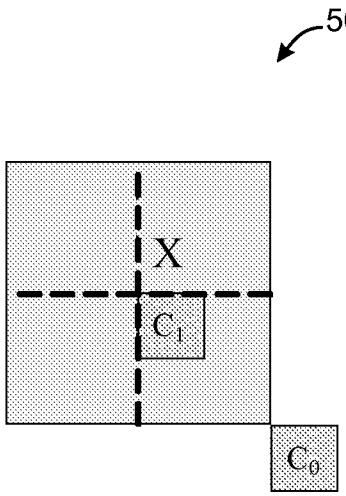


Fig.5b

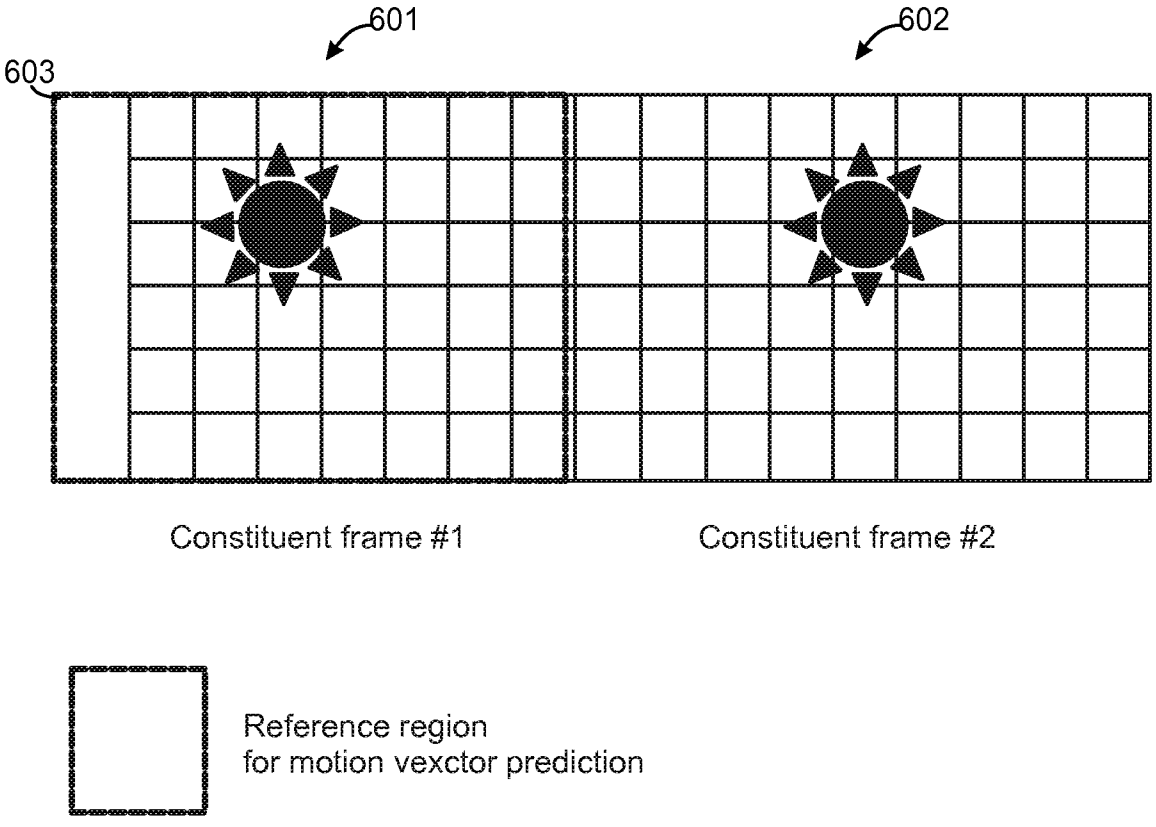


Fig. 6a

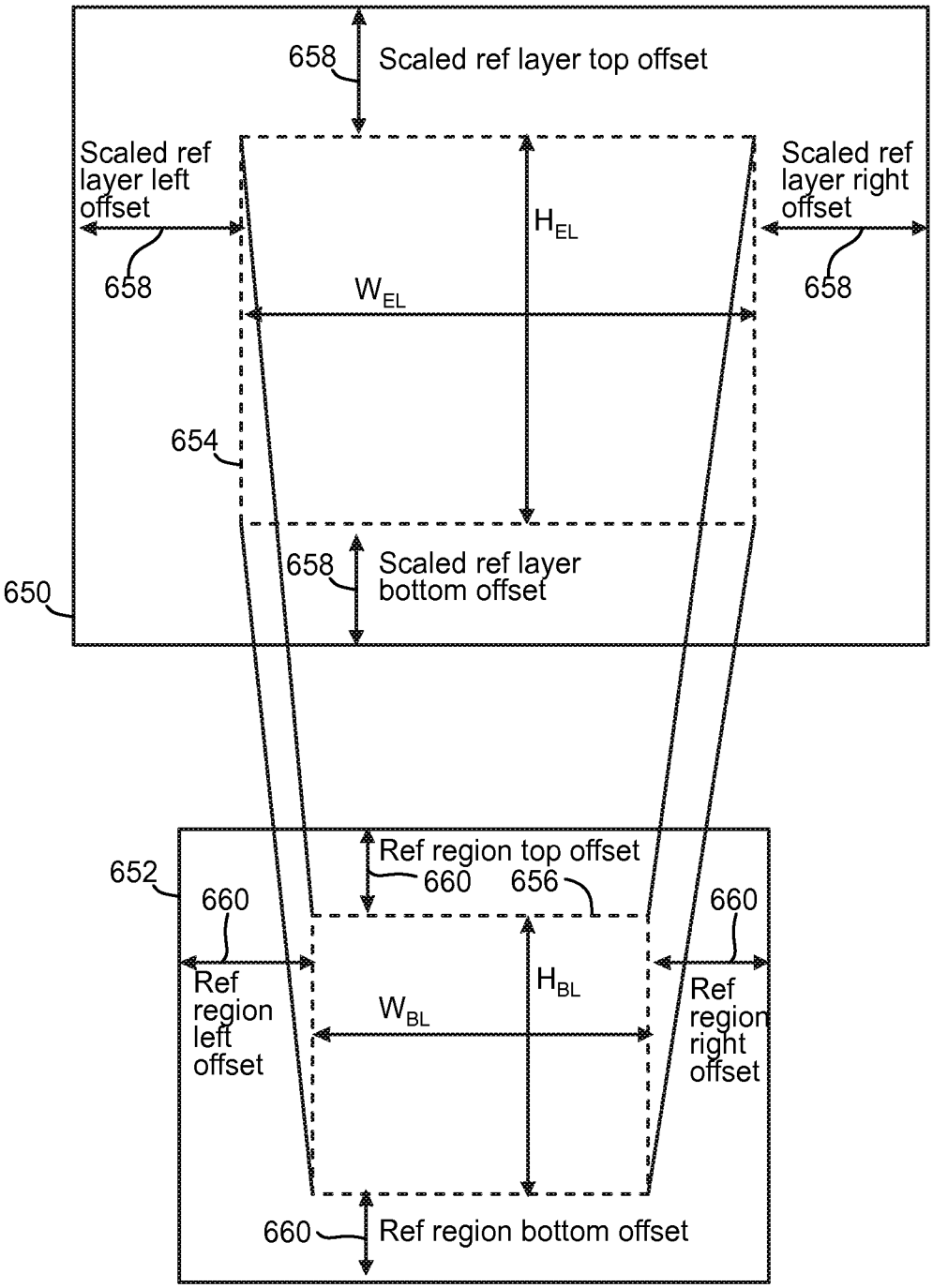

658
Scaled ref layer top offset
Scaled ref layer left offset
658
Scaled ref layer right offset
658
H_{EL}
W_{EL}
654
658
Scaled ref layer bottom offset
650
652
Ref region top offset
660
656
660
660
Ref region left offset
Ref region right offset
H_{BL}
W_{BL}
660
Ref region bottom offset

Fig. 6b

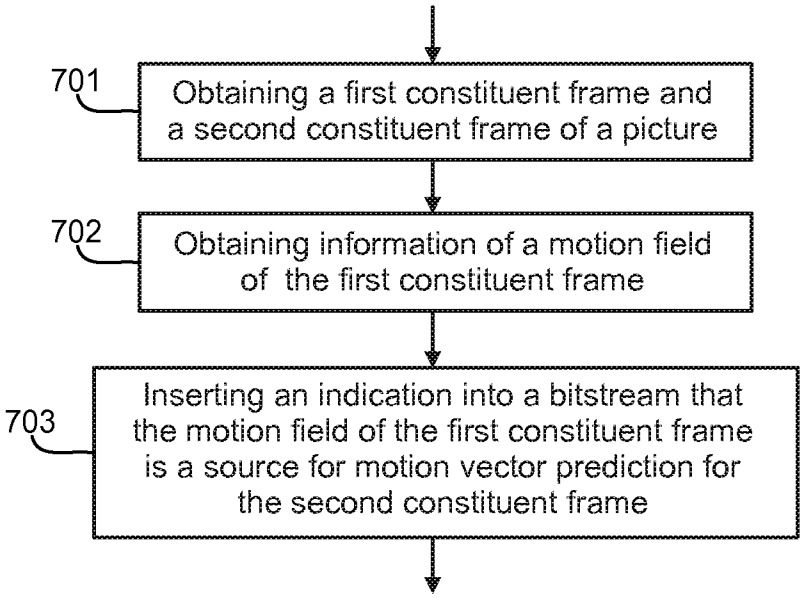

701
Obtaining a first constituent frame and a second constituent frame of a picture
702
Obtaining information of a motion field of the first constituent frame
703
Inserting an indication into a bitstream that the motion field of the first constituent frame is a source for motion vector prediction for the second constituent frame

Fig. 7a

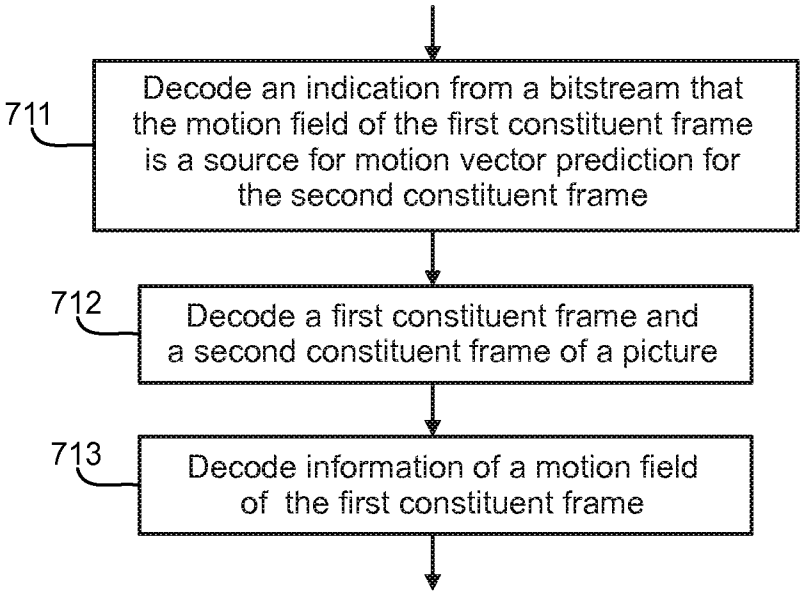

711
Decode an indication from a bitstream that the motion field of the first constituent frame is a source for motion vector prediction for the second constituent frame
712
Decode a first constituent frame and a second constituent frame of a picture
713
Decode information of a motion field of the first constituent frame

Fig. 7b

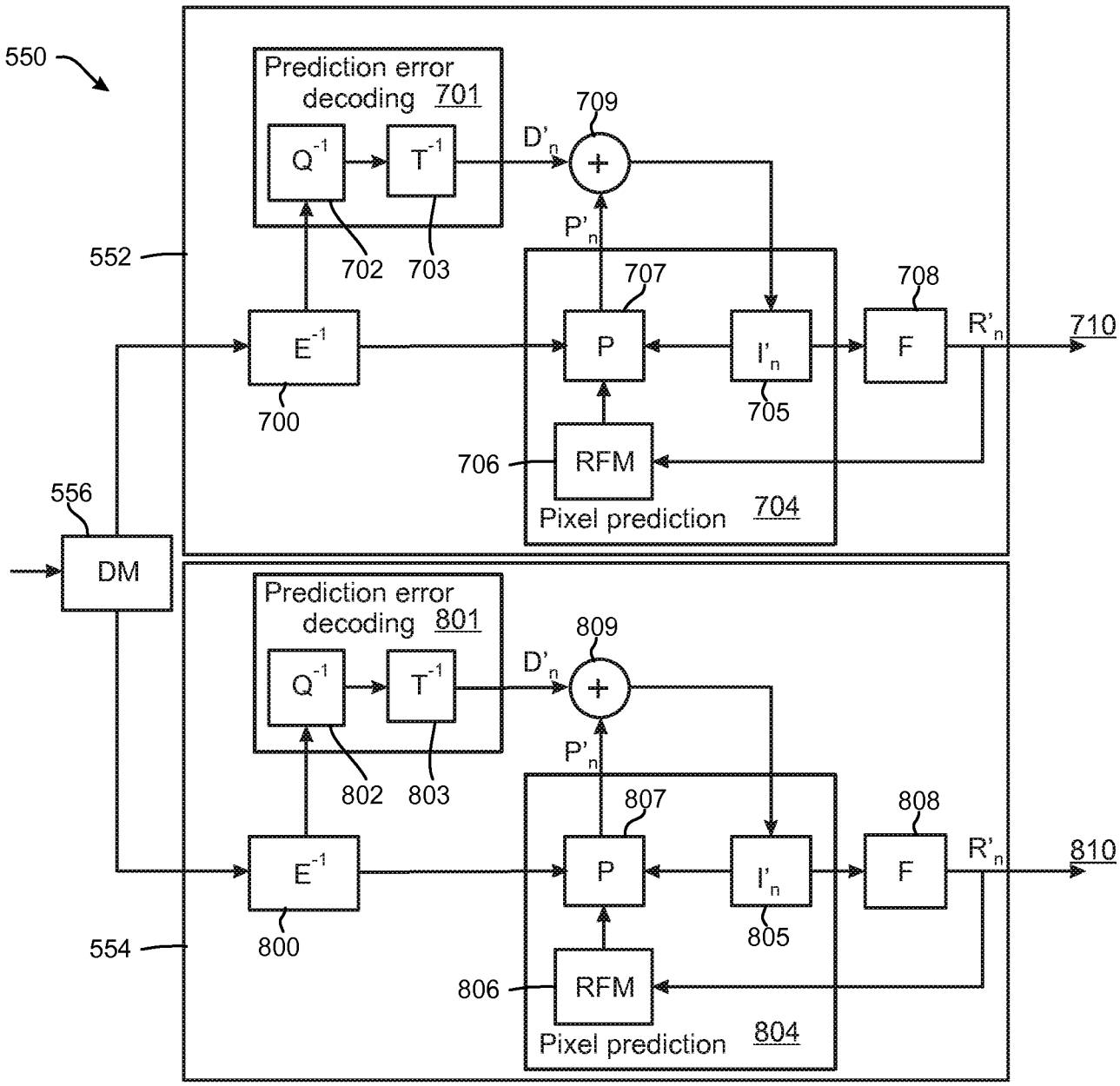

550
Prediction error decoding 701
Q-1
T-1
D'n
709
+
P'n
552
702
703
707
708
E-1
P
I'n
F
R'n
710
700
705
706
RFM
Pixel prediction 704
556
DM
Prediction error decoding 801
809
802
803
807
808
800
805
554
806
804

Fig. 8

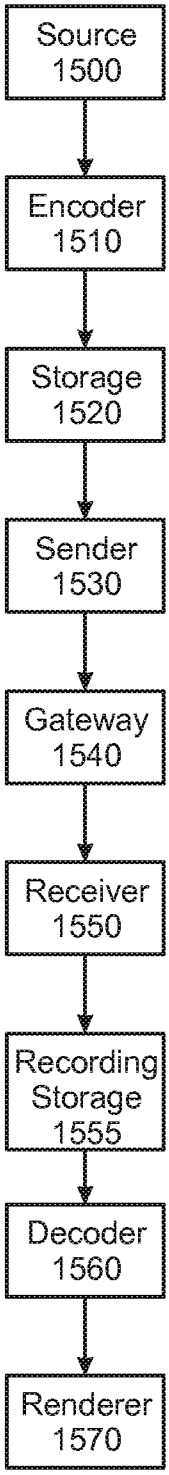

Source
1500
Encoder
1510
Storage
1520
Sender
1530
Gateway
1540
Receiver
1550
Recording
Storage
1555
Decoder
1560
Renderer
1570

Fig. 9

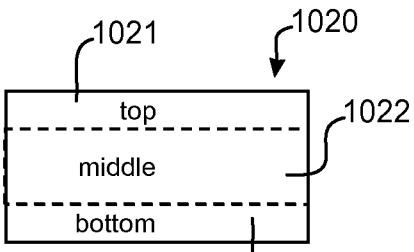


Fig. 10a

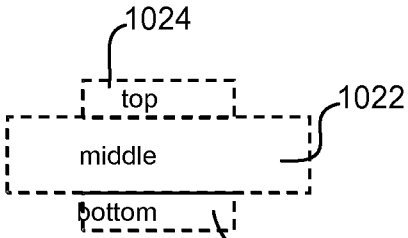


Fig. 10b

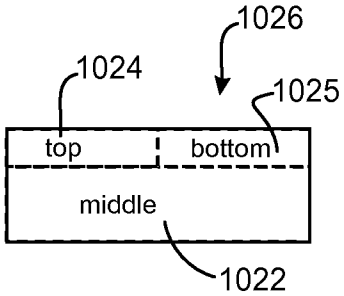


Fig. 10c

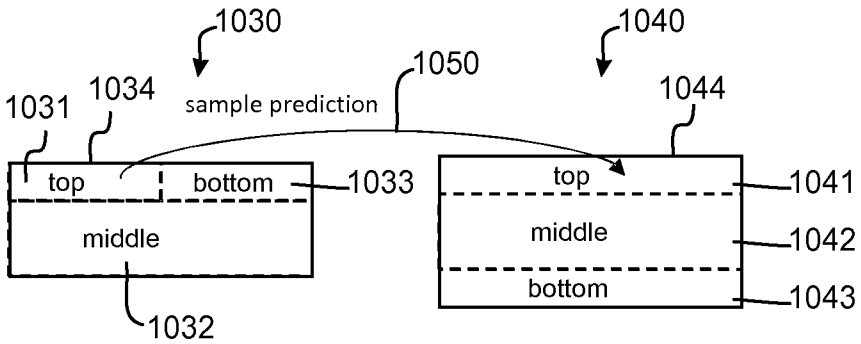


Fig. 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2017/050177

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

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

FI, SE, NO, DK

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2014354771 A1 (WANG JIAO [CA] et al.) 04 December 2014 (04.12.2014) abstract; pars. 0003, 0012, 0015-0016, 0022-0023, 0025, 0033, 0036; Figs. 2-3	1-10, 12-20, 22-23
X	CHEN, Y. ET AL. Test Model 11 of 3D-HEVC and MV-HEVC. Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V), Document: JCT3V-K1003, [online], 2015-05-01, [retrieved on 2017-05-24]. Retrieved from <http://phenix.int-evry.fr/jct3v/doc_end_user/documents/11_Geneva/wg11/JCT3V-K1003-v1.zip> section 2.2.4	1-4, 8-10, 12-15, 18-20, 22-23

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

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 May 2017 (26.05.2017)	05 June 2017 (05.06.2017)

Name and mailing address of the ISA/FI
Finnish Patent and Registration Office
P.O. Box 1160, FI-00101 HELSINKI, Finland
Facsimile No. +358 9 6939 5328

Authorized officer
Ari Peltola
Telephone No. +358 9 6939 500

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2017/050177

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KONIECZNY, J. ET AL. Depth-based inter-view prediction of motion vectors for improved multiview video coding. In: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010 Piscataway, NJ, USA: IEEE, 2010-06-07, pages 1-4, ISBN 978-1-4244-6377-0, <DOI:10.1109/3DTV.2010.5506319> abstract; sections 2-4; Figs. 1-2	1-4, 8-10, 12-15, 18-20, 22-23
X	DONG, J. ET AL. Scalable video coding technology proposal by InterDigital Communications. Joint Collaborative Team on Video Coding (JCT-VC), [online], 2012-10-11, [retrieved on 2017-05-26]. Retrieved from <http://phenix.int-evry.fr/jct/doc_end_user/documents/11_Shanghai/wg11/JCTVC-K0034-v5.zip> slides 2, 12-13, 26	1, 3-8, 12-14, 16-18, 22-23
X	WO 2015082763 A1 (NOKIA TECHNOLOGIES OY [FI]) 11 June 2015 (11.06.2015) abstract; pars. 0130, 0280-0281, 0302	1, 3-8, 12-14, 16-18, 22-23
X	SCHWARZ, H. ET AL. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. In: IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY PISCATAWAY, NJ, US: IEEE SERVICE CENTER, 2007-09-01, Vol. 17, pages 1103-1120, ISSN 1051-8215, <DOI:10.1109/TCSVT.2007.905532> section V.B.1 (“Inter-Layer Prediction”)	1, 3-8, 12-14, 16-18, 22-23
X	US 2014254666 A1 (RAPAKA KRISHNAKANTH [US] et al.) 11 September 2014 (11.09.2014) abstract; pars. 0086-0089, 0160-0162	1, 3-8, 11-14, 16-18, 21-23

INTERNATIONAL SEARCH REPORT
Information on Patent Family Members

International application No.
PCT/FI2017/050177

Patent document cited in search report	Publication date	Patent family members(s)	Publication date
US 2014354771 A1	04/12/2014	None	
WO 2015082763 A1	11/06/2015	CN 106416250 A	15/02/2017
		EP 3078195 A1	12/10/2016
		US 2015156501 A1	04/06/2015
US 2014254666 A1	11/09/2014	US 9467707 B2	11/10/2016
		CN 105009587 A	28/10/2015
		CN 105009588 A	28/10/2015
		CN 105191310 A	23/12/2015
		CN 105191311 A	23/12/2015
		EP 2965514 A1	13/01/2016
		EP 2965515 A1	13/01/2016
		EP 2965516 A1	13/01/2016
		EP 2965516 B1	24/08/2016
		EP 2965517 A1	13/01/2016
		JP 2016513916 A	16/05/2016
		JP 6113869 B2	12/04/2017
		JP 2016513917 A	16/05/2016
		JP 2016514425 A	19/05/2016
		JP 2016514427 A	19/05/2016
		KR 20150122780 A	02/11/2015
		KR 20150122787 A	02/11/2015
		KR 20150122788 A	02/11/2015
		KR 20150126892 A	13/11/2015
		US 2014254669 A1	11/09/2014
		US 9294776 B2	22/03/2016
		US 2014254667 A1	11/09/2014
		US 9473779 B2	18/10/2016
		US 2014254668 A1	11/09/2014
		US 9578339 B2	21/02/2017
		WO 2014138217 A1	12/09/2014
		WO 2014138222 A1	12/09/2014
		WO 2014138232 A1	12/09/2014
		WO 2014138240 A1	12/09/2014

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2017/050177

CLASSIFICATION OF SUBJECT MATTER

IPC
H04N 19/55 (2014.01)
H04N 19/53 (2014.01)
H04N 19/56 (2014.01)
H04N 19/597 (2014.01)
H04N 19/33 (2014.01)