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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 January 2014 (16.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/010537 A1

(51) International Patent Classification:
H04N 13/00 (2006.01) *H04N 7/36* (2006.01)
H04N 7/32 (2006.01)

(21) International Application Number:
PCT/JP2013/068578

(22) International Filing Date:
2 July 2013 (02.07.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
13/544,463 9 July 2012 (09.07.2012) US

(71) Applicant: **MITSUBISHI ELECTRIC CORPORATION** [JP/JP]; 7-3, Marunouchi 2-chome, Chiyoda-ku, Tokyo, 1008310 (JP).

(72) Inventors: **TIAN, Dong**; 574 Burroughs Road, Boxborough, Massachusetts, 01719 (US). **ZOU, Feng**; 117 Harvard St. Apt. 4, Cambridge, Massachusetts, 02139 (US). **VETRO, Anthony**; 30 Gorham Street, Arlington, Massachusetts, 02474 (US).

(74) Agents: **SOGA, Michiharu** et al.; S. Soga & Co., 8th Floor, Kokusai Building, 1-1, Marunouchi 3-chome, Chiyoda-ku, Tokyo, 1000005 (JP).

(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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

(54) Title: METHOD AND SYSTEM FOR PROCESSING MULTIVIEW VIDEOS FOR VIEW SYNTHESIS USING MOTION VECTOR PREDICTOR LIST

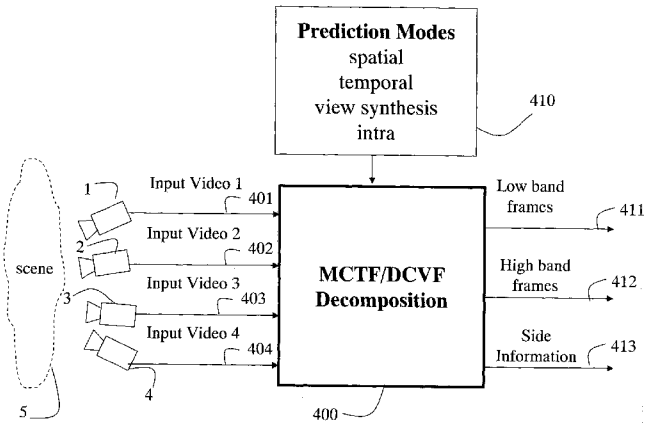


FIG. 4

(57) Abstract: Multiview videos are acquired by overlapping cameras. Side information is used to synthesize multiview videos. A reference picture list is maintained for current frames of the multiview videos, the reference picture indexes temporal reference pictures and spatial reference pictures of the acquired multiview videos and the synthesized reference pictures of the synthesized multiview video. Each current frame of the multiview videos is predicted according to reference pictures indexed by the associated reference picture list with a skip mode and a direct mode, whereby the side information is inferred from the synthesized reference picture. In addition, the skip and merge modes for *single view* video coding are modified to support *multiview* video coding by generating a motion vector prediction list by also considering neighboring blocks that are associated with synthesized reference pictures.

WO 2014/010537 A1

[DESCRIPTION]

[Title of Invention]

METHOD AND SYSTEM FOR PROCESSING MULTIVIEW VIDEOS FOR VIEW SYNTHESIS USING MOTION VECTOR PREDICTOR LIST

[Technical Field]

[0001]

This invention relates generally to encoding and decoding multiview videos, and more particularly to synthesizing multiview videos.

[Background Art]

[0002]

Multiview video encoding and decoding is essential for applications such as three dimensional television (3DTV), free viewpoint television (FTV), and multi-camera surveillance. Multiview video encoding and decoding is also known as *dynamic* light field compression.

[0003]

Figure 1 shows a prior art 'simulcast' system 100 for multiview video encoding. Cameras 1-4 acquire sequences of frames or videos 101-104 of a scene 5. Each camera has a different view of the scene. Each video is encoded 111-114 independently to corresponding encoded videos 121-124. That system uses conventional 2D video encoding techniques. Therefore, that system does not correlate between the different videos acquired by the cameras from the different viewpoints while predicting frames of the encoded video. Independent encoding decreases compression efficiency, and thus network bandwidth and storage are increased.

[0004]

Figure 2 shows a prior art disparity compensated prediction system 200 that does use inter-view correlations. Videos 201-204 are encoded 211-

214 to encoded videos 231-234. The videos 201 and 204 are encoded independently using a standard video encoder such as MPEG-2 or H.264, also known as MPEG-4 Part 10. These independently encoded videos are 'reference' videos. The remaining videos 202 and 203 are encoded using temporal prediction and inter-view predictions based on reconstructed reference videos 251 and 252 obtained from decoders 221 and 222. Typically, the prediction is determined adaptively on a per block basis, S.C. Chan et al., "The data compression of simplified dynamic light fields," Proc. IEEE Int. Acoustics, Speech, and Signal Processing Conf., April, 2003.

[0005]

Figure 3 shows prior art '*lifting-based*' wavelet decomposition, see W. Sweldens, "The data compression of simplified dynamic light fields," J. Appl. Comp. Harm. Anal., vol. 3, no. 2, pp. 186-200, 1996. Wavelet decomposition is an effective technique for static light field compression. Input samples 301 are split 310 into odd samples 302 and even samples 303. The odd samples are predicted 320 from the even samples. A prediction error forms high band samples 304. The high band samples are used to update 330 the even samples and to form low band samples 305. That decomposition is invertible so that linear or non-linear operations can be incorporated into the prediction and update steps.

[0006]

The lifting scheme enables a motion-compensated temporal transform, i.e., motion compensated temporal filtering (MCTF) which, for videos, essentially filters along a *temporal* motion trajectory. A review of MCTF for video coding is described by Ohm et al., "Interframe wavelet coding - motion picture representation for universal scalability," Signal Processing: Image Communication, vol. 19, no. 9, pp. 877-908, October 2004. The

lifting scheme can be based on any wavelet kernel such as Harr or 5/3 Daubechies, and any motion model such as block-based translation or affine global motion, without affecting the reconstruction.

[0007]

For encoding, the MCTF decomposes the video into high band frames and low band frames. Then, the frames are subjected to spatial transforms to reduce any remaining spatial correlations. The transformed low and high band frames, along with associated motion information, are entropy encoded to form an encoded bitstream. MCTF can be implemented using the lifting scheme shown in Figure 3 with the temporally adjacent videos as input. In addition, MCTF can be applied recursively to the output low band frames.

[0008]

MCTF-based videos have a compression efficiency comparable to that of video compression standards such as H.264/AVC. In addition, the videos have inherent temporal scalability. However, that method cannot be used for directly encoding multiview videos in which there is a correlation between videos acquired from multiple views because there is no efficient method for predicting views that accounts for correlation in time.

[0009]

The lifting scheme has also been used to encode *static* light fields, i.e., single multiview images. Rather than performing a motion-compensated temporal filtering, the encoder performs a disparity compensated inter-view filtering (DCVF) across the static views in the spatial domain, see Chang et al., "Inter-view wavelet compression of light fields with disparity compensated lifting," SPIE Conf on Visual Communications and Image Processing, 2003. For encoding, DCVF decomposes the static light field into high and low band images, which are then subject to spatial transforms to

reduce any remaining spatial correlations. The transformed images, along with the associated disparity information, are entropy encoded to form the encoded bitstream. DCVF is typically implemented using the lifting-based wavelet transform scheme as shown in Figure 3 with the images acquired from spatially adjacent camera views as input. In addition, DCVF can be applied recursively to the output low band images. DCVF-based static light field compression provides a better compression efficiency than independently coding the multiple frames. However, that method also cannot encode multiview videos in which both temporal correlation and spatial correlation between views are used because there is no efficient method for predicting views that account for correlation in time.

[0010]

In certain applications, the depth signal can be part of the input of the system as shown in Fig. 25. In a 'simulcast' system 2500, can either acquired at the same time when color video is shot, e.g., using depth cameras 251A, 252A, 253A, and 254A, or be estimated by an offline procedure. Note that the depth is present as an input of the system and the depth is encoded 2500, 2511, 2511A, 2512, 2512A, 2513, 2513A, 2514, and 2514A, and transmitted as part of bitstreams 2521, 2521A, 2522, 2522A, 2523, 2523A, 2524, and 2524A. The depth encoder may or may not be the same as the color encoder.

[Summary of the Invention]

[0011]

A method and system to decompose multiview videos acquired of a scene by multiple cameras is presented.

[0012]

Each multiview video includes a sequence of frames, and each camera provides a different view of the scene.

[0013]

A prediction mode is selected from a temporal, spatial, view synthesis, and intra-prediction mode.

[0014]

The multiview videos are then decomposed into low band frames, high band frames, and side information according to the selected prediction mode.

[0015]

A novel video reflecting a synthetic view of the scene can also be generated from one or more of the multiview videos.

[0016]

More particularly, one embodiment of the invention provides a system and method for encoding and decoding videos. Multiview videos are acquired of a scene with corresponding cameras arranged at a poses such that there is view overlap between any pair of cameras. A synthesized multiview video is generated from the acquired multiview videos for a virtual camera. A reference picture list is maintained in a memory for each current frame of each of the multiview videos and the synthesized video. The reference picture list indexes temporal reference pictures and spatial reference pictures of the acquired multiview videos and the synthesized reference pictures of the synthesized multiview video. Then, each current frame of the multiview videos is predicted according to reference pictures indexed by the associated reference picture list during encoding and decoding.

[0017]

In another embodiment, the depth images corresponding to the multiview videos are part of the input data and this data are encoded as part

of the bitstream. Thus, the reference information to be derived includes a reference picture index refIdxLX and a motion vector (X, Y), depending on a SKIP type.

[0018]

In the HEVC standard, the conventional *skip and direct* modes are modified to *skip and merge* modes, respectively. The skip and merge modes enable selection of different motion vector predictors (MVPs) from neighboring blocks. For this purpose, an MVP list and a syntax element mvp_flag are used to indicate the MVP that is associated with a neighboring block from among a list of candidate MVPs associated with other neighboring blocks. In one embodiment of the invention, the motion information of the current block is derived by copying the motion information of the selected MVP.

[0019]

In another embodiment of the invention, the motion information of the current block is derived based on the motion information of the selected MVP and the motion vector difference information. The motion information includes both reference frame information and motion vector information. In these embodiments, modifications to the skip and merge modes that are supported by the existing HEVC standard for *single view* video coding are described for now also described for *multiview* video coding. In particular, the embodiments consider the specific case in which a reference picture is a synthesized reference picture. Our method generates the MVP list by also considering neighboring blocks that are associated with synthesized reference pictures.

[Brief Description of the Drawings]

[0020]

Figure 1 is a block diagram of a prior art system for encoding multiview videos;

Figure 2 is a block diagram of a prior art disparity compensated prediction system for encoding multiview videos;

Figure 3 is a flow diagram of a prior art wavelet decomposition process;

Figure 4 is a block diagram of a MCTF/DCVF decomposition according to an embodiment of the invention;

Figure 5 is a block diagram of low-band frames and high band frames as a function of time and space after the MCTF/DCVF decomposition according to an embodiment of the invention;

Figure 6 is a block diagram of prediction of high band frame from adjacent low-band frames according to an embodiment of the invention;

Figure 7 is a block diagram of a multiview coding system using macroblock-adaptive MCTF/DCVF decomposition according to an embodiment of the invention;

Figure 8 is a schematic of video synthesis according to an embodiment of the invention;

Figure 9 is a block diagram of a prior art reference picture management;

Figure 10 is a block diagram of multiview reference picture management according to an embodiment of the invention;

Figure 11 is a block diagram of multiview reference pictures in a decoded picture buffer according to an embodiment of the invention;

Figure 12 is a graph comparing coding efficiencies of different multiview reference picture orderings;

Figure 13 is a block diagram of dependencies of view mode on the multiview reference picture list manager according to an embodiment of the invention;

Figure 14 is diagram of a prior art reference picture management for single view coding systems that employ prediction from temporal reference pictures;

Figure 15 is a diagram of a reference picture management for multiview coding and decoding systems that employ prediction from multiview reference pictures according to an embodiment of the invention;

Figure 16 is a block diagram of view synthesis in a decoder using depth information encoded and received as side information according to an embodiment of the invention;

Figure 17 is a block diagram of cost calculations for selecting a prediction mode according to an embodiment of the invention;

Figure 18 is a block diagram of view synthesis in a decoder using depth information estimated by a decoder according to an embodiment of the invention;

Figure 19 is a block diagram of multiview videos using V-frames to achieve spatial random access in the decoder according to an embodiment of the invention;

Figure 20 is a block diagram of view synthesizes using warping and interpolation according to an embodiment of the invention;

Figures 21 is a block diagram of a depth search according to an embodiment of the invention;

Figure 22 is a block diagram of sub-pel reference matching according to an embodiment of the invention;

Figure 23 is a block diagram of conventional skip mode;

Figure 24 is a block diagram of synthetic skip mode according to an embodiment of the invention;

Figure 25 is a block diagram of a prior art system for encoding multiview videos including depth;

Figure 26 is a flow diagram of a decoding process that uses a synthetic skip mode according to an embodiment of the invention;

Figure 27 is a flow diagram of a decoding process that uses an adaptive-reference skip mode according to an embodiment of the invention; and

Figure 28 is a flow diagram of a decoding process that generates a motion vector predictor list according to embodiments of the invention.

[Description of Embodiments]

[0021]

One embodiment of our invention provides a joint temporal/inter-view processing method for encoding and decoding frames of multiview videos. Multiview videos are videos that are acquired of a scene by multiple cameras having different poses. We define a pose camera as both its 3D (x, y, z) position, and its 3D (θ, ρ, φ) orientation. Each pose corresponds to a 'view' of the scene.

[0022]

The method uses temporal correlation between frames within the same video acquired for a particular camera pose, as well as spatial correlation between synchronized frames in different videos acquired from multiple camera views. In addition, '*synthetic*' frames can be correlated, as described below.

[0023]

In one embodiment, the temporal correlation uses motion compensated temporal filtering (MCTF), while the spatial correlation uses disparity compensated inter-view filtering (DCVF).

[0024]

In another embodiment of the invention, spatial correlation uses prediction of one view from synthesized frames that are generated from 'neighboring' frames. Neighboring frames are temporally or spatially adjacent frames, for example, frames before or after a current frame in the temporal domain, or one or more frames acquired at the same instant in time but from cameras having different poses or views of the scene.

[0025]

In another embodiment of the invention, depth images that correspond to the multiview videos are used to generate the synthesize frames. The depth images are considered as part of the input data and are encoded to a bitstream. A decoding process is used to reconstruct the depth images from the bitstream.

[0026]

Each frame of each video includes macroblocks of pixels. Therefore, the method of multiview video encoding and decoding according to one embodiment of the invention is macroblock adaptive. The encoding and decoding of a current macroblock in a current frame is performed using several possible prediction modes, including various forms of temporal, spatial, view synthesis, and intra prediction. To determine the best prediction mode on a macroblock basis, one embodiment of the invention provides a method for selecting a prediction mode. The method can be used for any number of camera arrangements.

[0027]

As used herein, a *reference picture* is defined as any frame that is used during the encoding and decoding to 'predict' a current frame. Typically, reference pictures are spatially or temporally adjacent or 'neighboring' to the current frame.

[0028]

It is important to note that the same operations are applied in both the encoder and decoder because the same set of reference pictures are used at any given time instant to encode and decode the current frame.

[0029]

MCTF/DCVF Decomposition

Figure 4 show a MCTF/DCVF decomposition 400 according to one embodiment of the invention. Frames of input videos 401-404 are acquired of a scene 5 by cameras 1-4 having different posses. Note, as shown in Fig 8, some of the cameras 1a and 1b can be at the same locations but with different orientations. It is assumed that there is some amount of view overlap between any pair of cameras. The poses of the cameras can change while acquiring the multiview videos. Typically, the cameras are synchronized with each other. Each input video provides a different 'view' of the scene. The input frames 401-404 are sent to a MCTF/DCVF decomposition 400. The decomposition produces encoded low-band frames 411, encoded high band frames 412, and associated side information 413. The high band frames encode prediction errors using the low band frames as reference pictures. The decomposition is according to selected prediction modes 410. The prediction modes include spatial, temporal, view synthesis, and intra prediction modes. The prediction modes can be selected adaptively on a per macroblock basis for each current frame. With intra prediction, the current macroblock is predicted from other macroblocks in the same frame.

[0030]

Figure 5 shows a preferred alternating 'checkerboard pattern' of the low band frames (L) 411 and the high band frames (H) 412 for a neighborhood of frames 510. The frames have a spatial (view) dimension 501 and a temporal dimension 502. Essentially, the pattern alternates low band frames and high band frames in the spatial dimension for a single instant in time, and additionally alternates temporally the low band frames and the high band frames for a single video.

[0031]

There are several advantages of this checkerboard pattern. The pattern distributes low band frames evenly in both the space and time dimensions, which achieves scalability in space and time when a decoder only reconstructs the low band frames. In addition, the pattern aligns the high band frames with adjacent low band frames in both the space and time dimensions. This maximizes the correlation between reference pictures from which the predictions of the errors in the current frame are made, as shown in Figure 6.

[0032]

According to a lifting-based wavelet transform, the high band frames 412 are generated by predicting one set of samples from the other set of samples. The prediction can be achieved using a number of modes including various forms of temporal prediction, various forms of spatial prediction, and a view synthesis prediction according to the embodiments of invention described below.

[0033]

The means by which the high band frames 412 are predicted and the necessary information required to make the prediction are referred to as the

side information 413. If a *temporal* prediction is performed, then the temporal mode is signaled as part of the side information along with corresponding *motion* information. If a *spatial* prediction is performed, then the spatial mode is signaled as part of the side information along with corresponding *disparity* information. If *view synthesis* prediction is performed, then the view synthesis mode is signaled as part of the side information along with corresponding *disparity*, *motion* and *depth* information.

[0034]

As shown in Figures 6, the prediction of each current frame 600 uses neighboring frames 510 in both the space and time dimensions. The frames that are used for predicting the current frame are called reference pictures. The reference pictures are maintained in the reference list, which is part of the encoded bitstream. The reference pictures are stored in the decoded picture buffer.

[0035]

In one embodiment of the invention, the MCTF and DCVF are applied adaptively to each current macroblock for each frame of the input videos to yield decomposed low band frames, as well as the high band frames and the associated side information. In this way, each macroblock is processed adaptively according to a 'best' prediction mode. An optimal method for selecting the prediction mode is described below.

[0036]

In one embodiment of the invention, the MCTF is first applied to the frames of each video independently. The resulting frames are then further decomposed with the DCVF. In addition to the final decomposed frames, the corresponding side information is also generated. If performed on a

macroblock-basis, then the prediction mode selections for the MCTF and the DCVF are considered separately. As an advantage, this prediction mode selection inherently supports *temporal* scalability. In this way, lower temporal rates of the videos are easily accessed in the compressed bitstream.

[0037]

In another embodiment, the DCVF is first applied to the frames of the input videos. The resulting frames are then temporally decomposed with the MCTF. In addition to the final decomposed frames, the side information is also generated. If performed on a macroblock-basis, then the prediction mode selections for the MCTF and DCVF are considered separately. As an advantage, this selection inherently supports *spatial* scalability. In this way, a reduced number of the views are easily accessed in the compressed bitstream.

[0038]

The decomposition described above can be applied recursively on the resulting set of low band frames from a previous decomposition stage. As an advantage, our MCTF/DCVF decomposition 400 effectively removes both *temporal* and *spatial* (inter-view) correlations, and can achieve a very high compression efficiency. The compression efficiency of our multiview video encoder outperforms conventional simulcast encoding, which encodes each video for each view independently.

[0039]

Encoding of MCTF/DCVF Decomposition

As shown in Figure 7, the outputs 411 and 412 of decomposition 400 are fed to a signal encoder 710, and the output 413 is fed to a side information encoder 720. The signal encoder 710 performs a transform, quantization and entropy coding to remove any remaining correlations in the

decomposed low band and high band frames 411-412. Such operations are well known in the art, Netravali and Haskell, Digital Pictures: Representation, Compression and Standards, Second Edition, Plenum Press, 1995.

[0040]

The side information encoder 720 encodes the side information 413 generated by the decomposition 400. In addition to the prediction mode and the reference picture list, the side information 413 includes motion information corresponding to the temporal predictions, disparity information corresponding to the spatial predictions and view synthesis and depth information corresponding to the view synthesis predictions.

[0041]

Encoding the side information can be achieved by known and established techniques, such as the techniques used in the MPEG-4 Visual standard, ISO/IEC 14496-2, "Information technology -- Coding of audio-visual objects - Part 2: Visual," 2nd Edition, 2001, or the more recent H.264/AVC standard, and ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," 2004.

[0042]

For instance, motion vectors of the macroblocks are typically encoded using predictive methods that determine a prediction vector from vectors in macroblocks in reference pictures. The difference between the prediction vector and the current vector is then subject to an entropy coding process, which typically uses the statistics of the prediction error. A similar procedure can be used to encode disparity vectors.

[0043]

Furthermore, depth information for each macroblock can be encoded using predictive coding methods in which a prediction from macroblocks in reference pictures is obtained, or by simply using a fixed length code to express the depth value directly. If pixel level accuracy for the depth is extracted and compressed, then texture coding techniques that apply transform, quantization and entropy coding techniques can be applied.

[0044]

The encoded signals 711-713 from the signal encoder 710 and side information encoder 720 can be multiplexed 730 to produce an encoded output bitstream 731.

[0045]

Decoding of MCTF/DCVF Decomposition

The bitstream 731 can be decoded 740 to produce output multiview videos 741 corresponding to the input multiview videos 401-404. Optionally, synthetic video can also be generated. Generally, the decoder performs the inverse operations of the encoder to reconstruct the multiview videos. If all low band and high band frames are decoded, then the full set of frames in both the space (view) dimension and time dimension at the encoded quality are reconstructed and available.

[0046]

Depending on the number of recursive levels of decomposition that were applied in the encoder and which type of decompositions were applied, a reduced number of videos and/or a reduced temporal rate can be decoded as shown in Figure 7.

[0047]

View Synthesis

As shown in Figure 8, view synthesis is a process by which frames 801 of a synthesized video are generated from frames 803 of one or more actual multiview videos. In other words, view synthesis provides a means to synthesize the frames 801 corresponding to a selected novel view 802 of the scene 5. This novel view 802 may correspond to a 'virtual' camera 800 not present at the time the input multiview videos 401-404 were acquired or the view can correspond to a camera view that is acquired, whereby the synthesized view will be used for prediction and encoding/decoding of this view as described below.

[0048]

If one video is used, then the synthesis is based on extrapolation or warping, and if multiple videos are used, then the synthesis is based on interpolation.

[0049]

Given the pixel values of frames 803 of one or more multiview videos and the depth values of points in the scene, the pixels in the frames 801 for the synthetic view 802 can be synthesized from the corresponding pixel values in the frames 803.

[0050]

View synthesis is commonly used in computer graphics for rendering still images for multiple views, see Buehler et al., "Unstructured Lumigraph Rendering," Proc. ACM SIGGRAPH, 2001. That method requires extrinsic and intrinsic parameters for the cameras, incorporated herein by reference

[0051]

View synthesis for compressing multiview videos is novel. In one embodiment of our invention, we generate synthesized frames to be used for predicting the current frame. In one embodiment of the invention,

synthesized frames are generated for designated high band frames. In another embodiment of the invention, synthesized frames are generated for specific views. The synthesized frames serve as reference pictures from which a current synthesized frame can be predicted.

[0052]

When depth signal is not part of the multiview video, one difficulty with this approach is that the depth values of the scene 5 are unknown. Therefore, we estimate the depth values using known techniques, e.g., based on correspondences of features in the multiview videos.

[0053]

Alternatively, for each synthesized video, we generate multiple synthesized frames, each corresponding to a candidate depth value. For each macroblock in the current frame, the best matching macroblock in the set of synthesized frames is determined. The synthesized frame from which this best match is found indicates the depth value of the macroblock in the current frame. This process is repeated for each macroblock in the current frame.

[0054]

A difference between the current macroblock and the synthesized block is encoded and compressed by the signal encoder 710. The side information for this multiview mode is encoded by the side information encoder 720. The side information includes a signal indicating the view synthesis prediction mode, the depth value of the macroblock, and an optional displacement vector that compensates for any misalignments between the macroblock in the current frame and the best matching macroblock in the synthesized frame to be compensated.

[0055]

Prediction Mode Selection

In the macroblock-adaptive MCTF/DCVF decomposition, the prediction mode m for each macroblock can be selected by minimizing a cost function adaptively on a per macroblock basis:

$$m^* = \arg\min_m J(m),$$

[0056]

where $J(m) = D(m) + \lambda R(m)$, and D is distortion, λ is a weighting parameter, R is rate, m indicates the set of candidate prediction modes, and m^* indicates the optimal prediction mode that has been selected based on a minimum cost criteria.

[0057]

The candidate modes m include various modes of temporal, spatial, view synthesis, and intra prediction. The cost function $J(m)$ depends on the rate and distortion resulting from encoding the macroblock using a specific prediction mode m.

[0058]

The distortion D measures a difference between a reconstructed macroblock and a source macroblock. The reconstructed macroblock is obtained by encoding and decoding the macroblock using the given prediction mode m. A common distortion measure is a sum of squared difference. The rate R corresponds to the number of bits needed to encode the macroblock, including the prediction error and the side information. The weighting parameter λ controls the rate-distortion tradeoff of the macroblock coding, and can be derived from a size of a quantization step.

[0059]

Detailed aspects of the encoding and decoding processes are described in further detail below. In particular, the various data structures that are used by the encoding and decoding processes are described. It should be understood that the data structures, as described herein, that are used in the encoder are identical to corresponding data structures used in the decoder. It should also be understood that the processing steps of the decoder essentially follow the same processing steps as the encoder, but in an inverse order.
[0060]

Reference Picture Management

Figure 9 shows a reference picture management for prior art single-view encoding and decoding systems. Temporal reference pictures 901 are managed by a single-view reference picture list (RPL) manager 910, which determines insertion 920 and removal 930 of temporal reference pictures 901 to a decoded picture buffer (DPB) 940. A reference picture list 950 is also maintained to indicate the frames that are stored in the DPB 940. The RPL is used for reference picture management operations such as insert 920 and remove 930, as well as temporal prediction 960 in both the encoded and the decoder.
[0061]

In single-view encoders, the temporal reference pictures 901 are generated as a result of applying a set of typical encoding operations including prediction, transform and quantization, then applying the inverse of those operations including inverse quantization, inverse transform and motion compensation. Furthermore, temporal reference pictures 901 are only inserted into the DPB 940 and added to the RPL 950 when the temporal pictures are required for the prediction of a current frame in the encoder.
[0062]

In single-view decoders, the same temporal reference pictures 901 are generated by applying a set of typical decoding operations on the bitstream including inverse quantization, inverse transform and motion compensation. As in the encoder, the temporal reference pictures 901 are only inserted 920 into the DPB 940 and added to the RPL 950 if they are required for prediction of a current frame in the decoder.

[0063]

Figure 10 shows a reference picture management for multiview encoding and decoding. In addition to temporal reference pictures 1003, the multiview systems also include spatial reference pictures 1001 and synthesized reference pictures 1002. These reference pictures are collectively referred to as multiview reference pictures 1005. The multiview reference pictures 1005 are managed by a multiview RPL manager 1010, which determines insertion 1020 and removal 1030 of the multiview reference pictures 1005 to the multiview DPB 1040. For each video, a multiview reference picture list (RPL) 1050 is also maintained to indicate the frames that are stored in the DPB. That is, the RPL is an index for the DPB. The multiview RPLs are used for reference picture management operations such as insert 1020 and remove 1030, as well as prediction 1060 of the current frame.

[0064]

It is noted that prediction 1060 for the multiview system is different than prediction 960 for the single-view system because prediction from different types of multiview reference pictures 1005 is enabled. Further details on the multiview reference picture management 1010 are described below.

[0065]

Multiview Reference Picture List Manager

Before encoding a current frame in the encoder or before decoding the current frame in the decoder, a set of multiview reference pictures 1005 can be indicated in the multiview RPL 1050. As defined conventionally and herein, a set can have zero (null set), one or multiple elements. Identical copies of the RPLs are maintained by both the encoder and decoder for each current frame.

[0066]

All frames inserted in the multiview RPLs 1050 are initialized and marked as usable for prediction using an appropriate syntax. According to the H.264/AVC standard and reference software, the 'used_for_reference' flag is set to '1'. In general, reference pictures are initialized so that a frame can be used for prediction in a video encoding system. To maintain compatibility with conventional single-view video compression standards, such as H.264/AVC, each reference picture is assigned a picture order count (POC). Typically, for single-view encoding and decoding systems, the POC corresponds to the temporal ordering of a picture, e.g., the frame number. For multiview encoding and decoding systems, temporal order alone is not sufficient to assign a POC for each reference picture. Therefore, we determine a unique POC for every multiview reference picture according to a convention. One convention is to assign a POC for temporal reference pictures based on temporal order, and then to reserve a sequence of very high POC numbers, e.g., 10,000 - 10,100, for the spatial and synthesized reference pictures. Other POC assignment conventions, or simply "ordering" conventions, are described in further detail below.

[0067]

All frames used as multiview reference pictures are maintained in the RPL and stored in the DPB in such a way that the frames are treated as conventional reference pictures by the encoder 700 or the decoder 740. This way, the encoding and decoding processes can be conventional. Further details on storing multiview reference pictures are described below. For each current frame to be predicted, the RPL and DPB are updated accordingly.

[0068]

Defining and Signaling Multiview Conventions

The process of maintaining the RPL is coordinated between the encoder 700 and the decoder 740. In particular, the encoder and decoder maintain identical copies of multiview reference picture list when predicting a particular current frame.

[0069]

A number of conventions for maintaining the multi-frame reference picture list are possible. Therefore, the particular convention that is used is inserted in the bitstream 731, or provided as sequence level side information, e.g., configuration information that is communicated to the decoder. Furthermore, the convention allows different prediction structures, e.g., 1-D arrays, 2-D arrays, arcs, crosses, and sequences synthesized using view interpolation or warping techniques.

[0070]

For example, a synthesized frame is generated by warping a corresponding frame of one of the multiview videos acquired by the cameras. Alternatively, a conventional model of the scene can be used during the synthesis. In other embodiments of our invention, we define several multiview reference picture maintenance conventions that are dependent on view type, insertion order, and camera properties.

[0071]

The view type indicates whether the reference picture is a frame from a video other than the video of the current frame, or whether the reference picture is synthesized from other frames, or whether the reference picture depends on other reference pictures. For example, synthesized reference pictures can be maintained differently than reference pictures from the same video as the current frame, or reference pictures from spatially adjacent videos.

[0072]

The insertion order indicates how reference pictures are ordered in the RPL. For instance, a reference picture in the same video as the current frame can be given a lower order value than a reference picture in a video taken from an adjacent view. In this case, the reference picture is placed earlier in the multiview RPL.

[0073]

Camera properties indicate properties of the camera that is used to acquire the reference picture, or the virtual camera that is used to generate a synthetic reference picture. These properties include translation and rotation relative to a fixed coordinate system, i.e., the camera 'pose', intrinsic parameters describing how a 3-D point is projected into a 2-D image, lens distortions, color calibration information, illumination levels, etc. For instance, based on the camera properties, the proximity of certain cameras to adjacent cameras can be determined automatically, and only videos acquired by adjacent cameras are considered as part of a particular RPL.

[0074]

As shown in Figure 11, one embodiment of our invention uses a convention that reserves a portion 1101 of each reference picture list for

temporal reference pictures 1003, reserves another portion 1102 for synthesized reference pictures 1002 and a third portion 1103 for spatial reference pictures 1001. This is an example of a convention that is dependent only on the view type. The number of frames contained in each portion can vary based on a prediction dependency of the current frame being encoded or decoded.

[0075]

The particular maintenance convention can be specified by standard, explicit or implicit rules, or in the encoded bitstream as side information.

[0076]

Storing Pictures in the DPB

The multiview RPL manager 1010 maintains the RPL so that the order in which the multiview reference pictures are stored in the DPB corresponds to their 'usefulness' to improve the efficiency of the encoding and decoding. Specifically, reference pictures in the beginning of the RPL can be predicatively encoded with fewer bits than reference pictures at the end of the RPL.

[0077]

As shown in Figure 12, optimizing the order in which multiview references pictures are maintained in the RPL can have a significant impact on coding efficiency. For example, following the POC assignment described above for initialization, multiview reference pictures can be assigned a very large POC value because they do not occur in the normal temporal ordering of a video sequence. Therefore, the default ordering process of most video codecs can place such multiview reference pictures earlier in the reference picture lists.

[0078]

Because temporal reference pictures from the same sequence generally exhibit stronger correlations than spatial reference pictures from other sequences, the default ordering is undesirable. Therefore, the multiview reference pictures are either explicitly reordered by the encoder, whereby the encoder then signals this reordering to the decoder, or the encoder and decoder implicitly reorder multiview reference pictures according to a predetermined convention.

[0079]

As shown in Figure 13, the order of the reference pictures is facilitated by a view mode 1300 to each reference picture. It is noted that the view mode 1300 also affects the multiview prediction process 1060. In one embodiment of our invention, we use three different types of view modes, I-view, P-view and B-view, which are described in further detail below.

[0080]

Before describing the detailed operation of multiview reference picture management, prior art reference picture management for single video encoding and decoding systems is shown in Figure 14. Only temporal reference pictures 901 are used for the temporal prediction 960. The temporal prediction dependency between temporal reference pictures of the video in acquisition or display order 1401 is shown. The reference pictures are reordered 1410 into an encoding order 1402, in which each reference picture is encoded or decoded at a time instants t_0 - t_6. Block 1420 shows the ordering of the reference pictures for each instant in time. At time t_0, when an intra-frame I_0 is encoded or decoded, there are no temporal reference pictures used for temporal prediction, hence the DBP/RPL is empty. At time t_1, when the uni-directional inter-frame P_1 is encoded or

decoded, frame I_0 is available as a temporal reference picture. At times t_2 and t_3, both frames I_0 and P_1 are available as reference frames for bi-directional temporal prediction of inter-frames B_1 and B_2. The temporal reference pictures and DBP/RPL are managed in a similar way for future pictures.

[0081]

To describe the multiview case according to an embodiment of the invention, we consider the three different types of views described above and shown in Figure 15: I-view, P-view, and B-view. The multiview prediction dependency between reference pictures of the videos in display order 1501 is shown. As shown in Figure 15, the reference pictures of the videos are reordered 1510 into a coding order 1502 for each view mode, in which each reference picture is encoded or decoded at a given time instant denoted t_0 - t_2. The order of the multiview reference pictures is shown in block 1520 for each time instant.

[0082]

The I-view is the simplest mode that enables more complex modes. I-view uses conventional encoding and prediction modes, without any spatial or synthesized prediction. For example, I-views can be encoded using conventional H.264/AVC techniques without any multiview extensions. When spatial reference pictures from an I-view sequence are placed into the reference lists of other views, these spatial reference pictures are usually placed after temporal reference pictures.

[0083]

As shown in Figure 15, for the I-view, when frame I_0 is encoded or decoded at t_0, there are no multiview reference pictures used for prediction.

Hence, the DBP/RPL is empty. At time t_1, when frame P_0 is encoded or decoded, I_0 is available as a temporal reference picture. At time t_2, when the frame B_0 is encoded or decoded, both frames I_0 and P_0 are available as temporal reference pictures.

[0084]

P-view is more complex than I-view in that P-view allows prediction from another view to exploit the spatial correlation between views. Specifically, sequences encoded using the P-view mode use multiview reference pictures from other I-view or P-view. Synthesized reference pictures can also be used in the P-view. When multiview reference pictures from an I-view are placed into the reference lists of other views, P-views are placed after both temporal reference pictures and after multiview references pictures derived from I-views.

[0085]

As shown in Figure 15, for the P-view, when frame I_2 is encoded or decoded at t_0, a synthesized reference picture S_{20} and the spatial reference picture I_0 are available for prediction. Further details on the generation of synthesized pictures are described below. At time t_1, when P_2 is encoded or decoded, I_2 is available as a temporal reference picture, along with a synthesized reference picture S_{21} and a spatial reference picture P_0 from the I-view. At time t_2, there exist two temporal reference pictures I_2 and P_2,

as well as a synthesized reference picture S_{22} and a spatial reference picture B_0, from which predictions can be made.

[0086]

B-views are similar to P-views in that the B-views use multiview reference pictures. One key difference between P-views and B-views is that P-views use reference pictures from its own view as well as one other view, while B-views may reference pictures in multiple views. When synthesized reference pictures are used, the B-views are placed before spatial reference pictures because synthesized views generally have a stronger correlation than spatial references.

[0087]

As shown in Figure 15, for the B-view, when I_1 is encoded or decoded at t_0, a synthesized reference picture S_{10} and the spatial reference pictures I_0 and I_2 are available for prediction. At time t_1, when P_1 is encoded or decoded, I_1 is available as a temporal reference picture, along with a synthesized reference picture S_{11} and spatial reference pictures P_0 and P_2 from the I-view and P-view, respectively. At time t_2, there exist two temporal reference pictures I_1 and P_1, as well as a synthesized reference picture S_{12} and spatial reference pictures B_0 and B_2, from which predictions can be made.

[0088]

It must be emphasized that the example shown in Figure 15 is only for one embodiment of the invention. Many different types of prediction

dependencies are supported. For instance, the spatial reference pictures are not limited to pictures in different views at the same time instant. Spatial reference pictures can also include reference pictures for different views at different time instants. Also, the number of bi-directionally predicted pictures between intra-pictures and uni-directionally predicted inter-pictures can vary. Similarly, the configuration of I-views, P-views, and B-views can also vary. Furthermore, there can be several synthesized reference pictures available, each generated using a different set of pictures or different depth map or process.

[0089]

Compatibility

One important benefit of the multiview picture management according to the embodiments of the invention is that it is compatible with existing single-view video coding systems and designs. Not only does this provide minimal changes to the existing single-view video coding standards, but it also enables software and hardware from existing single view video coding systems to be used for multiview video coding as described herein.

[0090]

The reason for this is that most conventional video encoding systems communicate encoding parameters to a decoder in a compressed bitstream. Therefore, the syntax for communicating such parameters is specified by the existing video coding standards, such as the H.264/AVC standard. For example, the video coding standard specifies a prediction mode for a given macroblock in a current frame from other temporally related reference pictures. The standard also specifies methods used to encode and decode a resulting prediction error. Other parameters specify a type or size of a transform, a quantization method, and an entropy coding method.

[0091]

Therefore, our multiview reference pictures can be implemented with only limited number of modifications to standard encoding and decoding components such as the reference picture lists, decoded picture buffer, and prediction structure of existing systems. It is noted that the macroblock structure, transforms, quantization and entropy encoding remain unchanged.

[0092]

View Synthesis

As described above for Figure 8, view synthesis is a process by which frames 801 corresponding to a synthetic view 802 of a virtual camera 800 are generated from frames 803 acquired of existing videos. In other words, view synthesis provides a means to synthesize the frames corresponding to a selected novel view of the scene by a virtual camera not present at the time the input videos were acquired. Given the pixel values of frames of one or more actual video and the depth values of points in the scene, the pixels in the frames of the synthesized video view can be generated by extrapolation and/or interpolation.

[0093]

Prediction from Synthesized Views

Figure 16 shows a process for generating a reconstructed macroblock using the view-synthesis mode, when depth 1901 information is included in the encoded multiview bitstream 731. The depth for a given macroblock is decoded by a side information decoder 1910. The depth 1901 and the spatial reference pictures 1902 are used to perform view synthesis 1920, where a synthesized macroblock 1904 is generated. A reconstructed macroblock 1903 is then formed by adding 1930 the synthesized macroblock 1904 and a decoded residual macroblock 1905.

[0094]

Details on Multiview Mode Selection at Encoder

It is assumed in the following that depth images corresponding to the multiview videos are not part of the input data. However, the process can be modified when depth images are present as part of the input data.

[0095]

Figure 17 shows a process for selecting the prediction mode while encoding or decoding a current frame. Motion estimation 2010 for a current macroblock 2011 is performed using temporal reference pictures 2020. The resultant motion vectors 2021 are used to determine 2030 a first coding cost, $cost_1$ 2031, using temporal prediction. The prediction mode associated with this process is m_1.

[0096]

Disparity estimation 2040 for the current macroblock is performed using spatial reference pictures 2041. The resultant disparity vectors 2042 are used to determine 2050 a second coding cost, $cost_2$ 2051, using spatial prediction. The prediction mode associated with this process is denoted m_2.

[0097]

Depth estimation 2060 for the current macroblock is performed based on the spatial reference pictures 2041. View synthesis is performed based on the estimated depth. The depth information 2061 and the synthesized view 2062 are used to determine 2070 a third coding cost, $cost_3$ 2071, using view-synthesis prediction. The prediction mode associated this process is m_3.

[0098]

Adjacent pixels 2082 of the current macroblock are used to determine 2080 a fourth coding cost, $cost_4$ 2081, using intra prediction. The prediction mode associated with process is m_4.

[0099]

The minimum cost among $cost_1$, $cost_2$, $cost_3$ and $cost_4$ is determined 2090, and one of the modes m_1, m_2, m_3 and m_4 that has the minimum cost is selected as the best prediction mode 2091 for the current macroblock 2011.

[0100]

View Synthesis Using Depth Estimation

Using the view synthesis mode 2091, the depth information and displacement vectors for synthesized views can be estimated from decoded frames of one or more multiview videos. The depth information can be per-pixel depth estimated from stereo cameras, or it can be per-macroblock depth estimated from macroblock matching, depending on the process applied.

[0101]

An advantage of this approach is a reduced bandwidth because depth values and displacement vectors are not needed in the bitstream, as long as the encoder has access to the same depth and displacement information as the decoder. The encoder can achieve this as long as the decoder uses exactly the same depth and displacement estimation process as the encoder. Therefore, in this embodiment of the invention, a difference between the current macroblock and the synthesized macroblock is encoded by the encoder.

[0102]

The side information for this mode is encoded by the side information encoder 720. The side information includes a signal indicating the view

synthesis mode and the reference view(s). The side information can also include depth and displacement correction information, which is the difference between the depth and displacement used by the encoder for view synthesis and the values estimated by the decoder.

[0103]

Figure 18 shows the decoding process for a macroblock using the view-synthesis mode when the depth information is estimated or inferred in the decoder and is not conveyed in the encoded multiview bitstream. The depth 2101 is estimated 2110 from the spatial reference pictures 2102. The estimated depth and the spatial reference pictures are then used to perform view synthesis 2120, where a synthesized macroblock 2121 is generated. A reconstructed macroblock 2103 is formed by the addition 2130 of the synthesized macroblock and the decoded residual macroblock 2104.

[0104]

Spatial Random Access

In order to provide random access to frames in a conventional video, intra-frames, also known as I-frames, are usually spaced throughout the video. This enables the decoder to access any frame in the decoded sequence, although at a decreased compression efficiency.

[0105]

For our multiview encoding and decoding system, we provide a new type of frame, which we call a 'V-frame' to enable random access and increase compression efficiency. A V-frame is similar to an I-frame in the sense that the V-frame is encoded without any temporal prediction. However, the V-frame also allows prediction from other cameras or prediction from synthesized videos. Specifically, V-frames are frames in the compressed bitstream that are predicted from spatial reference pictures or synthesized

reference pictures. By periodically inserting V-frames, instead of I-frames, in the bitstream, we provide temporal random access as is possible with I-frames, but with a better encoding efficiency. Therefore, V-frames do not use temporal reference frames. Figure 19 shows the use of I-frames for the initial view and the use of V-frames for subsequent views at the same time instant 1900. It is noted that for the checkerboard configuration shown in Figure 5, V-frames would not occur at the same time instant for all views. Any of the low-band frames could be assigned a V-frame. In this case, the V-frames would be predicted from low-band frames of neighboring views.

[0106]

In H.264/AVC video coding standard, IDR frames, which are similar to MPEG-2 I-frames with closed GOP, imply that all reference pictures are removed from the decoder picture buffer. In this way, the frame before an IDR frame cannot be used to predict frames after the IDR frame.

[0107]

In the multiview decoder as described herein, V-frames similarly imply that all temporal reference pictures can be removed from the decoder picture buffer. However, spatial reference pictures can remain in the decoder picture buffer. In this way, a frame in a given view before the V-frame cannot be used to perform temporal prediction for a frame in the same view after the V-frame.

[0108]

To gain access to a particular frame in one of the multiview videos, the V-frame for that view must first be decoded. As described above, this can be achieved through prediction from spatial reference pictures or synthesized reference pictures, without the use of temporal reference pictures.

[0109]

After the V-frame of the select view is decoded, subsequent frames in that view are decoded. Because these subsequent frames are likely to have a prediction dependency on reference pictures from neighboring views, the reference pictures in these neighboring views are also be decoded.

[0110]

Multiview Encoding and Decoding

The above sections describe view synthesis for improved prediction in multiview coding and depth estimation. We now describe implementation for variable block-size depth and motion search, rate-distortion (RD) decision, sub-pel reference depth search, and context-adaptive binary arithmetic coding (CABAC) of depth information. The coding can include encoding in an encoder and decoding in a decoder. CABAC is specified by the H.264 standard PART 10, incorporated herein by reference.

[0111]

View Synthesis Prediction

To capture the correlations that exist both across cameras and across time, we implemented two methods of block prediction:

disparity compensated view prediction (DCVP); and

view synthesis prediction (VSP).

[0112]

DCVP

The first method, DCVP, corresponds to using a frame from a different camera (view) at the same time to predict a current frame, instead of using a frame from a different time of the same (view) camera. DCVP provides gains when a temporal correlation is lower than the spatial

correlation, e.g., due to occlusions, objects entering or leaving the scene, or fast motion.

[0113]

VSP

The second method, VSP, synthesizing frames for a virtual camera to predict a sequence of frames. VSP is complementary to DCVP due to the existence of non-translational motion between camera views and provide gains when the camera parameters are sufficiently accurate to provide high quality virtual views, which is often the case in real applications.

[0114]

As shown in Figure 20, we exploit these features of multiview video by synthesizing a virtual view from already encoded views and then performing predictive coding using the synthesized views. Figure 20 shows time on the horizontal axis, and views on the vertical axis, with view synthesis and warping 2001, and view synthesis and interpolation 2002.

[0115]

Specifically, for each camera *c,* we first synthesize a virtual frame I'[*c, t, x, y*] based on the unstructured lumigraph rendering technique of Buehler et al., see above, and then predicatively encode the current sequence using the synthesized view.

[0116]

To synthesize a frame I'[*c, t, x, y*], we require a depth map D[*c, t, x, y*] that indicates how far the object corresponding to pixel (*x, y*) is from camera *c* at time *t*, as well as an intrinsic matrix *A(c)*, rotation matrix *R(c)*, and a translation vector *T(c)* describing the location of camera *c* relative to some world coordinate system.

[0117]

Using these quantities, we can apply the well-known pinhole camera model to project the pixel location (x, y) into world coordinates $[u, v, w]$ via

$$[u, v, w] = R(c) \cdot A^{-1}(c) \cdot [x, y, 1] \cdot D[c, t, x, y] + T(c) \quad (1)$$

[0118]

Next, the world coordinates are mapped to a target coordinates $[x', y', z']$ of the frame in camera c', which we wish to predict from via

$$[x', y', z'] = A(c') \cdot R^{-1}(c') \cdot \{[u, v, w] - T(c')\} \quad (2)$$

[0119]

Finally, to obtain a pixel location, the target coordinates are converted to homogenous from $[x'/z', y'/z', 1]$, and an intensity for pixel location (x, y) in the synthesized frame is $I'[c, t, x, y] = I[c', t, x'/z', y'/z']$.

[0120]

Variable Block-Size Depth/Motion Estimation

Above, we described a picture buffer-management method that enables for the use of DCVP without changing the syntax. The disparity vectors between camera views were found by using the motion estimation step and could be used just as an extended reference type. In order to use VSP as another type of reference, we extend the typical motion-estimation process as follows.

[0121]

Given a candidate macroblock type *mb_type* and N possible reference frames, possibly including synthesized multiview reference frames, i.e., VSP, we find, for each sub-macroblock, a reference frame along with either a motion vector $\vec{m}$ or a depth/correction-vector pair J that minimizes a

following Lagrangian cost J using a Lagrange multiplier λ_{motion} or λ_{depth}, respectively,

$$J = \min(J_{motion}, J_{depth}),$$

with

$$J_{depth}(d, \overrightarrow{m_c}, \text{mul_ref} \mid \lambda_{depth}) = \sum_{X \in \text{sub-MB}} | X - X_{p_synth}(d, \overrightarrow{m_c}, \text{mul_ref}) | + \lambda_{depth}(R_d + R_{m_c} + R_{mul_ref})$$

and

$$J_{motion}(\vec{m}, \text{ref} \mid \lambda_{motion}) = \sum_{X \in \text{sub-MB}} | X - X_{p_motion}(\vec{m}, \text{ref}) | + \lambda_{motion}(R_m + R_{ref})$$

,

where the sum is taken over all the pixels in the sub-macroblock being considered (sub-MB), and X_{p_synth} or X_{p_motion} refer to the intensity of a pixel in the reference sub-macroblock.

[0122]

Note that here 'motion' not only refers to temporal motion but also inter-view motion resulting from the disparity between the views.

[0123]

Depth Search

We use a block-based depth search process to find the optimal depth for each variable-size sub-macroblock. Specifically, we define minimum, maximum, and incremental depth values D_{min}, D_{max}, D_{step}. Then, for each variable-size sub-macroblock in the frame we wish to predict, we select the depth to minimize the error for the synthesized block

$$D(c, t, x, y) = \min_{d \in \{D_{\min}, D_{\min} + D_{step}, D_{\min} + 2D_{step} + \ldots + D_{\max}\}} // I[c, t, x, y] - I[c', t, x', y'] //,$$

(3)

where $\| I[c, t, x, y] - I[c', t, x', y'] \|$ denotes an average error between the sub-macroblock centered at (x, y) in camera c at time t, and the corresponding block from which we predict.

[0124]

As an additional refinement to improve the performance of the basic VSP process, we find that due to slight inaccuracies in the camera parameters, non-idealities that are not captured by the pin hole camera model, adding a synthesis correction vector significantly improved the performance of VSP.

[0125]

Specifically as shown in Figure 21 for each macroblock 2100, we map the target frame 2101 to the reference frame 2102, and then to the synthesized frame 2103. However, instead of computing the coordinates to interpolate from using Equation (1), we compute $[u, v, w]$ by adding a synthesis correction vector (C_x, C_y) 2110 to each set of original pixel coordinates to obtain

$$[u, v, w] = R(v) \cdot A^{-1}(v) \cdot [x - C_x, y - C_y, 1] \cdot D[v, t, x - C_x, y - C_y] + T(v).$$

(4)

[0126]

We discovered that with a correction-vector search range of as small as +/-2 often significantly improves the quality of the resulting synthesized reference frame.

[0127]

Sub-Pixel Reference Matching

Because the disparity of two corresponding pixels in different cameras is, in general, not given by an exact multiple of integers, the target coordinates [x', y', z'] of the frame in camera c' which we wish to predict from given by Equation (2) does not always fall on an integer-grid point. Therefore, we use interpolation to generate pixel-values for sub-pel positions in the reference frame. This enables us to select a nearest sub-pel reference point, instead of integer-pel, thereby more accurately approximating the true disparity between the pixels.

[0128]

Figure 22 shows this process, where "oxx...ox" indicate pixels. The same interpolation filters adopted for sub-pel motion estimation in the H.264 standard are used in our implementation.

[0129]

Sub-Pixel Accuracy Correction Vector

We can further improve the quality of synthesis by allowing the use of sub-pel accuracy correction vector. This is especially true when combined with the sub-pel reference matching described above. Note that there is a subtle difference between the sub-pel motion-vector search and the current sub-pel correction-vector search.

[0130]

In motion-vector cases, we usually search through the sub-pel positions in the *reference* picture and selects a sub-pel motion-vector pointing to the sub-pel position minimizing the RD cost. However, in correction-vector cases, after finding the optimal depth-value, we search through the sub-pel positions in the *current* picture and select the correction-vector minimizing the RD cost.

[0131]

Shifting by a sub-pel correction-vector in the current picture does not always lead to the same amount of shift in the reference picture. In other words, the corresponding match in the reference picture is always found by rounding to a nearest sub-pel position after geometric transforms of Equations (1) and (2).

[0132]

Although coding a sub-pel accuracy correction-vector is relatively complex, we observe that the coding significantly improves the synthesis quality and often leads to improved RD performance.

[0133]

YUV-Depth Search

In depth estimation, regularization can achieve smoother depth maps. Regularization improves visual quality of synthesized prediction but slightly degrades its prediction quality as measured by sum of absolute differences (SAD).

[0134]

The conventional depth search process uses only the Y luminance component of input images to estimate the depth in the depth map. Although this minimizes the prediction error for Y component, it often results in visual artifacts, e.g., in the form of color mismatch, in the synthesized prediction. This means that we are likely to have a degraded objective, i.e., PSNRs for U,V, as well as subjective quality, in the form of color mismatch, in the final reconstruction.

[0135]

In order to cope with this problem, we extended the depth search process to use the Y luminance component *and* the U and V chrominance components. Using only the Y component can lead to visual artifacts

because a block can find a good match in a reference frame by minimizing the prediction error, but these two matching areas could be in two totally different colors. Therefore, the quality of U and V prediction and reconstruction can improve by incorporating the U and V components in the depth-search process.

[0136]

RD Mode Decision

A mode decision can be made by selecting the *mb_type* that minimizes a Lagrangian cost function J_{mode} defined as

$$J_{\text{mode}}(mb_type \mid \lambda_{\text{mode}}) = \sum_{X \in \text{MB}} (X - X_p)^2 + \lambda_{\text{mode}} (R_{side-\text{inf}o} + R_{residual})$$

,

where X_p refers to the value of a pixel in a reference MB, i.e., either a MB of a synthesized multiview, pure multiview or temporal reference, and $R_{side-\text{inf}o}$ includes the bit rate for encoding the reference index and the depth value/correction-value, or the motion-vector depending on the type of the reference frame.

[0137]

The method described above does not assume that depth images corresponding to the multiview video are part of the input data. If depth images are part of the content to be encoded and decoded, then the rate for the depth images should not be counted as part of the side information rate, i.e., $R_{side-info}$ is removed from the above Lagrangian cost function.

[0138]

CABAC Encoding of Side-Information

Note that we have to encode a depth-value and a correction vector for each synthesized MB when the MB is selected to be the best reference via RD mode-decision. Both the depth value and the correction vector are quantized using concatenated unary/3rd-order Exp-Golomb (UEG3) binarization with signedValFlag=1, and cut off parameter uCoff=9 in exactly the same manner as with the motion-vector.

[0139]

Then, different context-models are assigned to the bins of the resulting binary representation. The assignment of ctxIdxInc for the depth and the correction-vector components is essentially the same as that for motion-vectors as specified in Table 9-30 of ITU-T Recommendation H.264 & ISO/IEC 14496-10 (MPEG-4) AVC, "Advanced Video Coding for Generic Audiovisual Services," version 3: 2005, incorporated herein by reference, except that we do not apply sub-clause 9.3.3.1.1.7 to the first bin.

[0140]

In our embodiment, we predicatively encode the depth value and the correction vector using the same prediction scheme for motion-vectors. Because a MB or a sub-MB of size down to 8x8 can have its own reference picture either from a temporal, multiview, or synthesized multiview frame, the types of side-information can vary from MB to MB. This implies the number of neighboring MBs with identical reference picture can be reduced, potentially leading to less efficient side-information (motion-vector or depth/correction-vector) prediction.

[0141]

When a sub-MB is selected to use a synthesized reference, but has no surrounding MBs with the same reference, its depth/correction-vector is independently coded without any prediction. In fact, we found that it is often

good enough to binarize the correction-vector components with a fixed-length representation followed by CABAC encoding of the resulting bins. This is because MBs selected to use synthesized references tend to be isolated, i.e., the MBs do not have neighboring MBs with the same reference picture, and correction-vectors are usually less correlated with their neighbors than in the case of motion-vector cases.

[0142]

Syntax and Semantics

As described above, we incorporate a synthesized reference picture, in addition to the temporal and pure multiview references. Above, we described a multiview reference picture list management method that is compatible with the existing reference picture list management in the H.264/AVC standard, referenced above.

[0143]

The synthesized references in this embodiment are regarded as a special case of multiview reference, and thus, are handled in exactly the same way.

[0144]

We define a new high level syntax element called a view_parameter_set to describe the multiview identification and prediction structure. By slightly modifying the parameter, we can identify whether the current reference picture is of synthetic type or not. Hence, we can either decode the depth/correction-vector or the motion-vector for the given (sub-)MB depending on the type of the reference. Therefore, we can integrate the use of this new type of prediction by extending the macroblock-level syntax as specified in Appendix A.

[0145]

Skip Mode

In conventional skip mode, motion-vector information and reference indices are derived from co-located or neighboring macroblocks. Considering inter-view prediction based on view synthesis, an analogous mode that derives depth and correction-vector information from its co-located or neighboring macroblocks is considered as well. We refer to this new coding mode as synthetic skip mode.

[0146]

In the conventional skip mode as shown in Figure 23, which applies to both P-slices and B-slices, no residual data are coded for the current macroblock (X) 2311. For skip in P-slices, the first entry 2304 in the reference list 2301 is chosen as the reference to predict and derive information from, while for skip in B-slices, the earliest entry 2305 in the reference list 2301 among neighboring macroblocks (A, B, C) 2312-2314 is chosen as the reference from which to predict and derive information.

[0147]

Assuming that the view synthesis reference picture is not ordered as the first entry in the reference picture list, e.g., as shown in Figure 11, the reference picture for skip modes in both P-slices and B-slices would never be a view synthesis picture with existing syntax, and following conventional decoding processes. However, since the view synthesized picture may offer better quality compared to a disparity-compensated picture or a motion-compensated picture, we describe a change to the existing syntax and decoding processes to allow for skip mode based on the view synthesis reference picture.

[0148]

To utilize the skip mode with respect to a view synthesis reference, we provide a synthetic skip mode that is signaled with modifications to the existing mb_skip_flag. Currently, when the existing mb_skip_flag is equal to 1, the macroblock is skipped, and when it is equal to 0, the macroblock is not skipped.

[0149]

In a first embodiment, an additional bit is added in the case when mb_skip_flag is equal to 1 to distinguish the conventional skip mode with the new synthetic skip mode. If the additional bit equals 1, this signals the synthetic skip mode, otherwise if the additional bit equals 0, the conventional skip mode is used.

[0150]

The above signaling scheme would work well at relatively high bit-rates, where the number of skipped macroblocks tends to be less. However, for lower bit-rates, the conventional skip mode is expected to be invoked more frequently. Therefore, the signaling scheme that includes the synthetic skip modes should not incur additional overhead to signal the conventional skip. In a second embodiment, an additional bit is added in the case when mb_skip_flag is equal to 0 to distinguish the conventional non-skip mode with the new synthetic skip mode. If the additional bit equals 1, this signals the synthetic skip mode, otherwise if the additional bit equals 0, the conventional non-skip mode is used.

[0151]

When a high percentage of synthetic skip modes are chosen for a slice or picture, overall coding efficiency could be improved by reducing the overhead to signal the synthetic skip mode for each macroblock. In a third embodiment, the synthetic skip mode is signaled collectively for all

macroblocks in a slice. This is achieved with a slice_skip_flag included in the slice layer syntax of the bitstream. The signaling of the slice_skip_flag would be consistent with the mb_skip_flag described in the first and second embodiments.

[0152]

As shown in Figure 24 when a synthetic skip mode is signaled for P-slices, the first view synthesis reference picture 2402 in the reference picture list 2401 is chosen as the reference instead of the first entry 2304 in the reference picture list 2301 in the case of conventional skip. When a synthetic skip mode is signaled for B-slices, the earliest view synthesis reference picture 2403 in the reference picture list 2401 is chosen as the reference instead of the earliest entry 2305 in the reference picture list 2301 in the case of conventional skip.

[0153]

The reference information is derived from the neighboring partitions: A 2312, B 2313 and C 2314. In one embodiment, where the depth images corresponding to the multiview videos are not part of the input data and a depth value is estimated for each partition, the depth and correction-vector information for synthetic skip mode is derived as follows.

[0154]

A depth vector dpthLXN includes three components (Depth, CorrX, CorrY), where Depth is a scalar value representing the depth associated with a partition, and CorrX and CorrY are the horizontal and vertical components of a correction vector associated with the partition, respectively. The reference index refIdxLX of the current partition for which the depth vector is derived is assigned as the first or earliest view synthesis reference picture

depending on whether the current slices is a P-slice or B-slice, as described above.

[0155]

Inputs to this process are the neighboring partitions A 2312, B 2313 and C 2314, the depth vector for each of the neighboring partitions denoted dpthLXN (with N being replaced by A, B, or C), the reference indices refIdxLXN (with N being replaced by A, B, or C) of the neighboring partitions, and the reference index refIdxLX of the current partition.

[0156]

Output of this process is the depth vector prediction dpthpLX. The variable dpthpLX is derived as follows. When neither the B 2313 nor C 2314 neighboring partition is available and the A 2312 neighboring partition is available, then the following assignments apply: dpthLXB=dpthLXA and dpthLXC=dpthLXA, refIdxLXB=refIdxLA and refIdxLXC=refIdxLA.

[0157]

If refIdxLXN is a reference index of a multiview reference picture from which a synthetic multiview reference picture with the reference index refIdxLX is to be synthesized, refIdxLXN is considered equal to refIdxLX and its associated depth vector dpthLXN is derived by converting the disparity from the reference picture with the reference index refIdxLXN to an equivalent depth vector associated with the reference picture with the reference index refIdxLX.

[0158]

Depending on reference indices refIdxLXA, refIdxLXB, or refIdxLXC, the following applies. If one and only one of the reference indices refIdxLXA, refIdxLXB, or refIdxLXC is equal to the reference index refIdxLX of the current partition, the following applies. Let refIdxLXN be

the reference index that is equal to refIdxLX, the depth vector dpthLXN is assigned to the depth vector prediction dpthpLX. Otherwise, each component of the depth vector prediction dpthpLX is given by the median of the corresponding vector components of the depth vector dpthLXA, dpthLXB, and dpthLXC.

[0159]

In another embodiment, the depth images corresponding to the multiview videos are part of the input data and this data are encoded as part of the bitstream. Thus, the reference information to be derived includes a reference picture index refIdxLX and a motion vector (X, Y).

[0160]

Two instances of the derivation process are described with respect to Fig. 26 and Fig. 27.

[0161]

For both instances, inputs are the SKIP type, the neighboring partitions A 2312, B 2313 and C 2314, the reference indices refIdxLXN, and the motion vector for each of the neighboring partitions (XN, YN) with N equal to A, B or C. Output of this process is the reference picture index refIdxLX and a motion vector (X, Y) for the current partition.

[0162]

One embodiment is shown in Fig 26. The procedure starts 2601 by setting 2602 the macroblock (MB) index, and parsing 2603 the header. If 2610 MB indicates not to skip, decode 2620 the non-skip MB, and determine if 2670 if more MB need to processed, and end 2680 if no. Otherwise, increment 2671 the MB index.

[0163]

If 2630 Skip_Type is 0, set 2640 RefIdxLX point to the first synthetic reference picture, set 2650 motion vector (X,Y) to be a zero vector, and use 2660 the derived RefIdxLX and motion vector (X,Y) to do the prediction, continue at 2670.

[0164]

Otherwise, derive 2621 the reference picture index RefIdxLX, motion vector (X,Y) from neighbors A, B and C, and continue at 2660.

[0165]

In another embodiment shown in Fig 27, the SKIP_Type is interpreted differently. It can be one of two values: value 0 indicates a conventional SKIP and value 1 indicates an adaptive-reference SKIP, which may be signaled as described earlier. With value 0, the refIdxLX and the motion vector (X,Y) are derived 2720 from the neighboring partitions in the conventional way. Note that a synthetic picture may be selected in this procedure if the synthetic reference picture is selected by the neighboring partitions. In other words, value 0 will keep the conventional decoding process for the SKIP mode unchanged. With this design, the prediction method could easily propagate to the subsequent partitions, regardless of whether it refers to a non-synthetic reference or a synthetic reference.

[0166]

However, value 1 invokes a new process for the adaptive-reference skip mode as defined below. A reference index refIdxLX is first derived from neighboring partitions. If refIdxLX points 2740 to a synthetic reference picture, the refIdxLX is forced 2750 to point to the first non-synthetic reference picture, either a temporal reference picture or inter-view (spatial) reference picture. Otherwise, refIdxLX is modified to point 2745 to the first synthetic reference picture in the list. The MV predictor is always set to a

zero vector. In other words, value 1 serves as a trigger to modify the derived prediction method from the neighboring partitions.

[0167]

With this embodiment, instead of directly signaling synthetic prediction or non-synthetic prediction with SKIP mode, means to propagate a prediction method to subsequent partitions or means to modify prediction methods is enabled.

[0168]

Direct Mode

Similar to skip mode, the conventional direct mode for B-slices also derives motion-vector information and reference indices from neighboring macroblocks. Direct mode differs from skip mode in that residual data are also present. For the same reasons that we provide the synthetic skip mode, we also describe analogous extensions of direct mode that we refer to as synthetic direct mode and adaptive-reference direct mode.

[0169]

To invoke the conventional direct mode, a macroblock is coded as non-skip. Direct mode could then be applied to both 16x16 macroblocks and 8x8 blocks. Both of these direct modes are signaled as a macroblock mode.

[0170]

In another embodiment, the method for signaling the synthetic direct mode is done by adding an additional mode to the list of candidate macroblock modes.

[0171]

In a second embodiment, the method for signaling the synthetic direct mode or the adaptive-reference direct mode is done by signaling an

additional flag indicating that the 16x16 macroblock or 8x8 block is coded as synthetic direct mode, or adaptive-reference direct mode.

[0172]

When a synthetic direct mode is signaled for B-slices, the earliest view synthesis reference pictures in the reference picture lists are chosen as the references instead of the earliest entries in the reference picture lists in the case of conventional direct mode.

[0173]

When an adaptive-reference direct mode is signaled for B-slices and the derived reference picture index is associated with a non-synthesized reference picture, the earliest view synthesis reference pictures in the reference picture lists are chosen as the references instead of the earliest entries in the reference picture lists in the case of conventional direct mode.

[0174]

On the other hand, when an adaptive-reference direct mode is signaled for B-slices and the derived reference picture index is associated with a synthesized reference picture, the earliest non-synthesized reference picture in the reference picture list is chosen as the reference picture.

[0175]

Regardless of whether there is depth image that corresponds to the multiview video available as part of the input data, the same processes for deriving the depth and correction vector information as synthetic skip mode is done for synthetic direct modes. Similarly, the same processes for deriving the reference picture index information as adaptive-reference skip mode is done for adaptive-reference direct modes.

[0176]

In applications where depth images are available for multiple viewpoints, the multiple depth maps can be encoded in a similar way as the multiview color video. The processes described in this invention to use a synthetic picture as reference can be also applied to encode a second depth map. That is, the depth map from a first viewpoint can be warped to a second viewpoint using its own data to generate a synthesized depth image; and then the synthesized depth image is used a reference picture to encode and decode the second depth image.

[0177]

Motion Vector Predictor List with View Synthesis Reference Pictures

High Efficiency Video Coding (HEVC), also known as H.265 and MPEG-H Part 2, is a *draft* video compression standard, a successor to H.264/MPEG-4 AVC (Advanced Video Coding), currently under joint development by the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). HEVC is improves video quality and doubles the compression ratio compared to H.264.

[0178]

In the HEVC standard, the conventional *skip and direct* modes have been modified to *skip and merge* modes, respectively. The new skip and merge modes as supported in the HEVC standard enable selection of different motion vector predictors (MVPs) from neighboring blocks. For this purpose, an MVP list and a syntax element mvp_flag indicate that the MVP that is associated with a neighboring block from among a list of candidate MVPs associated with other neighboring blocks.

[0179]

In the case of skip and merge modes, the MVP list is also referred to as a *merging candidate list* and the index to the selected motion vector predictor is indicate by the merge_idx syntax element.

[0180]

In the case of non-skip and non-merge modes, the MVP list is also referred to as an *adaptive motion vector predictor* (AMVP) list, and the index to the selected motion vector predictor is indicate by the mvp_flag.

[0181]

In one embodiment of the invention, the motion information of the current block is derived by copying the motion information of the selected MVP. In another embodiment of the invention, the motion information of the current block is derived based on the motion information of the selected MVP and the motion vector difference information. Note that the motion information includes both reference frame information and motion vector information.

[0182]

In these embodiments, modifications to the skip and merge modes that are supported by the existing HEVC standard for single view video coding are described for multiview video coding. In particular, we consider the specific case in which a reference picture is a synthesized reference picture. Our method generates the MVP list by also considering neighboring blocks that are associated with synthesized reference pictures.

[0183]

When the MVP list indexes a motion vector associated with the synthesized reference picture, the motion vector predictor is set to zero. To generate an MVP list, i.e., a merging candidate list or an AMVP list, the spatial and temporal neighbors are first added into the MVP list as specified

in the HEVC standard. Then, the motion vector predictor associated with the synthesized reference picture is appended to the end of the MVP list if a synthesized reference picture is present in the reference picture buffer. Alternatively, the motion vector predictor associated with the synthesized reference picture can be placed at any position in the MVP list.

[0184]

Fig. 28 shows the decoding process incorporating the MVP generation method from start 2801 to end 2899. The process is invoked when the current prediction unit (PU) is coded using any Inter mode.

[0185]

In step 2810, the MVP list is constructed based on motion vectors associated with neighboring blocks in the current picture, spatial reference pictures and temporal reference pictures. Then, only if 2820 there is a synthetic reference picture in the reference picture buffer, then we append 2825 an MVP candidate pointing to the co-located block of the synthetic reference picture.

[0186]

In step 2830, we parse an index that selects the MVP from the MVP list. If the Inter mode is a skip or merge mode, then the corresponding syntax is merge_idx. If the Inter mode is a non-skip or non-merge mode, the corresponding syntax is amvp_l0_flag and amvp_l1_flag. According to the embodiments described herein, the semantics of the merge_idx, amvp_l0_flag and amvp_l1_flag are modified to account for the motion vector predictor associated with the synthesized reference picture.

[0187]

In step 2840, a check is made if the index points to a synthetic reference picture. If true, then set 2481 the reference picture as the synthetic

reference picture and the value of the MVP to zero. If false, then set 2842 the reference picture as the spatial or temporal reference picture and set the MVP according to the index. In either case, the MVP is then output 2850.

[0188]

When there are several synthesized reference pictures available in the buffer, the reference picture whose original position is nearest to the current view in terms of camera distance is selected. If there are multiple views at the same camera distance, then a predetermined selection order can be defined. For example, the left view is selected. Alternatively, the synthesized reference picture to be selected can be explicitly signaled in the bitstream.

[0189]

As another alternative, multiple motion vector predictors associated with multiple synthesized reference pictures can be appended to the MVP list.

[0190]

Appendix A

Macroblock Syntax and Semantics

mb_pred(mb_type) {	C	Descriptor
.....		
} else if(MbPartPredMode(mb_type, 0) != Direct) {		
.....		
for(mbPartIdx = 0; mbPartIdx < NumMbPart(mb_type); mbPartIdx++)		
if(MbPartPredMode (mb_type, mbPartIdx) != Pred_L1) {		
if (multiview_type(view_id(ref_idx_l0[mbPartIdx]))== 1) {		
depthd_l0[mbPartIdx][0]	2	se(v) | ae(v)
for(compIdx = 0; compIdx < 2; compIdx++)		
corr_vd_l0[mbPartIdx][0][compIdx]	2	se(v) | ae(v)
}		
else		
for(compIdx = 0; compIdx < 2; compIdx++)		
mvd_l0[mbPartIdx][0][compIdx]	2	se(v) | ae(v)
}		
for(mbPartIdx = 0; mbPartIdx < NumMbPart(mb_type); mbPartIdx++)		
if(MbPartPredMode(mb_type, mbPartIdx) != Pred_L0) {		
If (multiview_type (view_id(ref_idx_l1[mbPartIdx]))== 1) {		
depthd_l1[mbPartIdx][0]	2	se(v) | ae(v)
for(compIdx = 0; compIdx < 2; compIdx++)		
corr_vd_l1[mbPartIdx][0][compIdx]	2	se(v) | ae(v)
}		
else		
for(compIdx = 0; compIdx < 2; compIdx++)		
mvd_l1[mbPartIdx][0][compIdx]	2	se(v) | ae(v)
}		
}		
}		

sub_mb_pred(mb_type) {	C	Descriptor
.....		

for(mbPartIdx = 0; mbPartIdx < 4; mbPartIdx++)		
if(sub_mb_type[mbPartIdx] != B_Direct_8x8 && SubMbPredMode(sub_mb_type[mbPartIdx]) != Pred_L1)		

for(subMbPartIdx = 0;
subMbPartIdx <
NumSubMbPart(sub_mb_type[mbPartIdx]);
subMbPartIdx++) {		

if (multiview_type (view_id(ref_idx_l0[mbPartIdx]))== 1) {		
depthd_l0[mbPartIdx][subMbPartIdx]	2	se(v) | ae(v)
for(compIdx = 0; compIdx < 2; compIdx++)		
corr_vd_l0[mbPartIdx][subMbPartIdx][compIdx]	2	se(v) | ae(v)
}		
else		
for(compIdx = 0; compIdx < 2; compIdx++)		
mvd_l0[mbPartIdx][subMbPartIdx][compIdx]	2	se(v) | ae(v)
}		
for(mbPartIdx = 0; mbPartIdx < 4; mbPartIdx++)		
if(sub_mb_type[mbPartIdx] != B_Direct_8x8 && SubMbPredMode(sub_mb_type[mbPartIdx]) != Pred_L0)		

for(subMbPartIdx = 0;
subMbPartIdx <
NumSubMbPart(sub_mb_type[mbPartIdx]);
subMbPartIdx++) {		

if (multiview_type (view_id(ref_idx_l1[mbPartIdx]))==1) {		
depthd_l1[mbPartIdx][subMbPartIdx]	2	se(v) | ae(v)
for(compIdx = 0; compIdx < 2; compIdx++)		
corr_vd_l1[mbPartIdx][subMbPartIdx][compIdx]	2	se(v) | ae(v)
}		
else		
for(compIdx = 0; compIdx < 2; compIdx++)		
mvd_l1[mbPartIdx][subMbPartIdx][compIdx]	2	se(v) | ae(v)
}		

[0191]

Additions to subclause 7.4.5.1 macroblock prediction semantics :

depthd_l0[mbPartIdx][0] specifies the difference between a depth value to be used and its prediction. The index mbPartIdx specifies to which macroblock partition depthd_l0 is assigned. The partitioning of the macroblock is specified by mb_type.

depthd_l1[mbPartIdx][0] has the same semantics as depthd_l0, with l0 replaced by l1.

corr_vd_l0[mbPartIdx][0][compIdx] specifies the difference between a correction-vector component to be used and its prediction. The index mbPartIdx specifies to which macroblock partition corr_vd_l0 is assigned. The partitioning of the macroblock is specified by mb_type. The horizontal correction vector component difference is decoded first in decoding order and is assigned CompIdx = 0. The vertical correction vector component is decoded second in decoding order and is assigned CompIdx = 1.

corr_vd_l1[mbPartIdx][0][compIdx] has the same semantics as corr_vd_l0, with l0 replaced by l1.

[0192]

Additions to subclause 7.4.5.2 sub-macroblock prediction semantics :

depthd_l0[mbPartIdx][subMbPartIdx] has the same semantics as depthd_l0, except that it is applied to the sub-macroblock partition index with subMbPartIdx. The indices mbPartIdx and subMbPartIdx specify to which macroblock partition and sub-macroblock partition depthd_l0 is assigned.

depthd_l1[mbPartIdx][subMbPartIdx] has the same semantics as depthd_l0, with l0 replaced by l1.

corr_vd_l0[mbPartIdx][subMbPartIdx][compIdx] has the same semantics as corr_vd_l0, except that it is applied to the sub-macroblock partition index with subMbPartIdx. The indices mbPartIdx and subMbPartIdx specify to which macroblock partition and sub-macroblock partition corr_vd_l0 is assigned.

corr_vd_l1[mbPartIdx **][** subMbPartIdx **][** compIdx **]** has the same semantics as corr_vd_l1, with l0 replaced by l1.

[0193]

View parameter set syntax :

View_parameter_set() {	**Descriptor**
view_id	ue(v)
multiview_type	u (1)
if (multiview_type) {	
multiview_synth_ref0	ue(v)
multiview_synth_ref1	ue(v)
}	
num_multiview_refs_for_list0	ue(v)
num_multiview_refs_for_list1	ue(v)
for(i = 0; i < num_multiview_refs_for_list0; i++) {	
reference_view_for_list_0[i]	ue(v)
}	
for(i = 0; i < num_multiview_refs_for_list1; i++) {	
reference_view_for_list_1[i]	ue(v)
}	
}	

[0194]

Additions to view parameter set semantics :

multiview_type equal to 1 specifies that the current view is synthesized from other views. multiview_type equal to 0 specifies that the current view is not a synthesized one.

multiview_synth_ref0 specifies an index for the first view to be used for synthesis.

multiview_synth_ref1 specifies an index for the second view to be used for synthesis.

[Industrial Applicability]

[0195]

The method and system of this invention are applicable to processing of a plurality of multiview videos of a scene in many kinds of fields.

[CLAIMS]

[Claim 1]

A method for processing a plurality of multiview videos of a scene, comprising the steps of:

obtaining side information for synthesizing a particular view of the multiview video;

synthesizing a synthesized multiview video from the plurality of multiview videos and the side information;

generating, for each block of a current picture, a motion vector predictor (MVP) list, wherein the MVP list indexes a set of motion vectors associated with neighboring blocks in:

the current picture;

spatial reference pictures:

temporal reference pictures; and

synthesized reference pictures;

determining, for each block of the current picture, a motion vector predictor from the set of motion vectors indexed by the MVP list;

deriving, for each block of the current picture, the motion vector based on the motion vector predictor;

maintaining a reference picture list for each current picture of each of the plurality of multiview videos, the reference picture indexing temporal reference pictures and spatial reference pictures of the plurality of acquired multiview videos and the synthesized reference pictures of the synthesized multiview video; and

predicting each block of the current picture of the plurality of multiview videos according to reference pictures indexed by the associated

reference picture list and the derived motion vector, wherein the steps are performed in a decoder.

[Claim 2]

The method of claim 1 in which the motion vector predictor associated with the synthesized reference picture is set to zero.

[Claim 3]

The method of claim 1, wherein the derived motion vector is equal to the motion vector predictor.

[Claim 4]

The method of claim 1, wherein a motion vector difference is added to the motion vector predictor to obtain the derived motion vector.

[Claim 5]

The method of claim 1, further comprising:

deriving the reference picture index for each block of each current picture using the reference picture indices of the neighboring blocks.

[Claim 6]

The method of claim 1, in which the side information includes depth values.

[Claim 7]

The method of claim 1, in which the side information is obtained in an encoder.

[Claim 8]

The method of claim 1, in which the side information is obtained in a decoder.

[Claim 9]

The method of claim 1, in which the determining of the motion vector predictor, further comprises:

parsing an index from the bitstream to indicate the motion vector predictor from the motion vector predictor list.

[Claim 10]

A system for processing a plurality of multiview videos of a scene, comprising:

a plurality of cameras, each camera configured to acquire a multiview video of a scene;

a decoder further comprising:

means for obtaining side information for synthesizing a particular view of the multiview video;

means for synthesizing a synthesized multiview video from the plurality of multiview videos and the side information;

means for generating, for each block of the current picture, a motion vector predictor (MVP) list, wherein the MVP list indexes a set of motion vectors associated with neighboring blocks in:

the current picture;

spatial reference pictures;

temporal reference pictures; and

synthesized reference pictures;

means for determining, for each block of the current picture, a motion vector predictor from the set of motion vectors in the MVP list according to an index;

means for deriving, for each block of the current picture, a motion vector from the motion vector predictor;

a memory buffer configured to maintain a reference picture list for each current picture of each of the plurality of multiview videos, the reference picture indexing temporal reference pictures and spatial reference pictures of the plurality of acquired multiview videos and the synthesized reference pictures of the synthesized multiview video; and

means for predicting each block of each current picture of the plurality of multiview videos according to reference pictures indexed by the associated reference picture list and the motion vector.

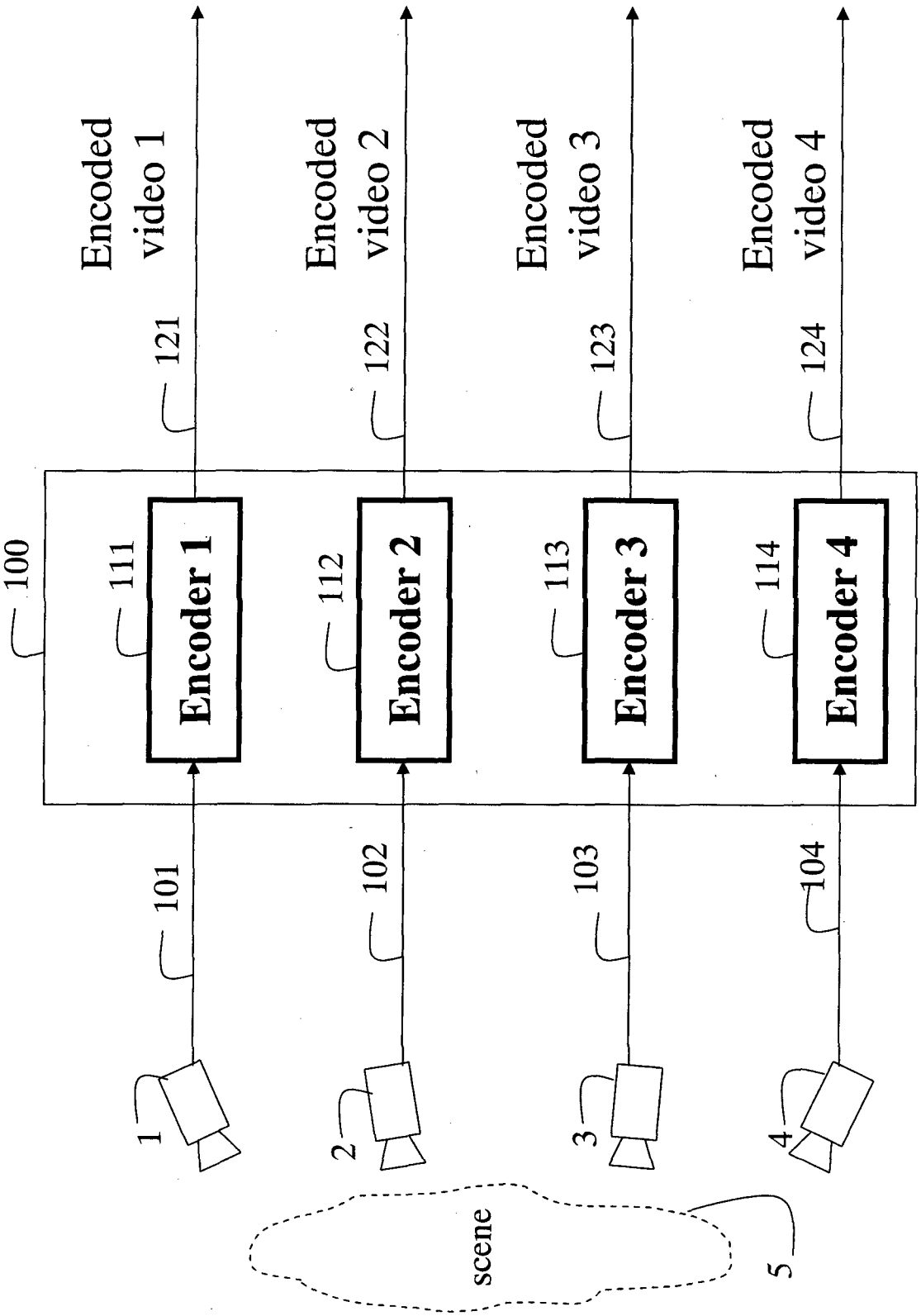

Encoded video 1
Encoded video 2
Encoded video 3
Encoded video 4
121
122
123
124
100
111
112
113
114
Encoder 1
Encoder 2
Encoder 3
Encoder 4
101
102
103
104
1
2
3
4
scene
5

FIG. 1

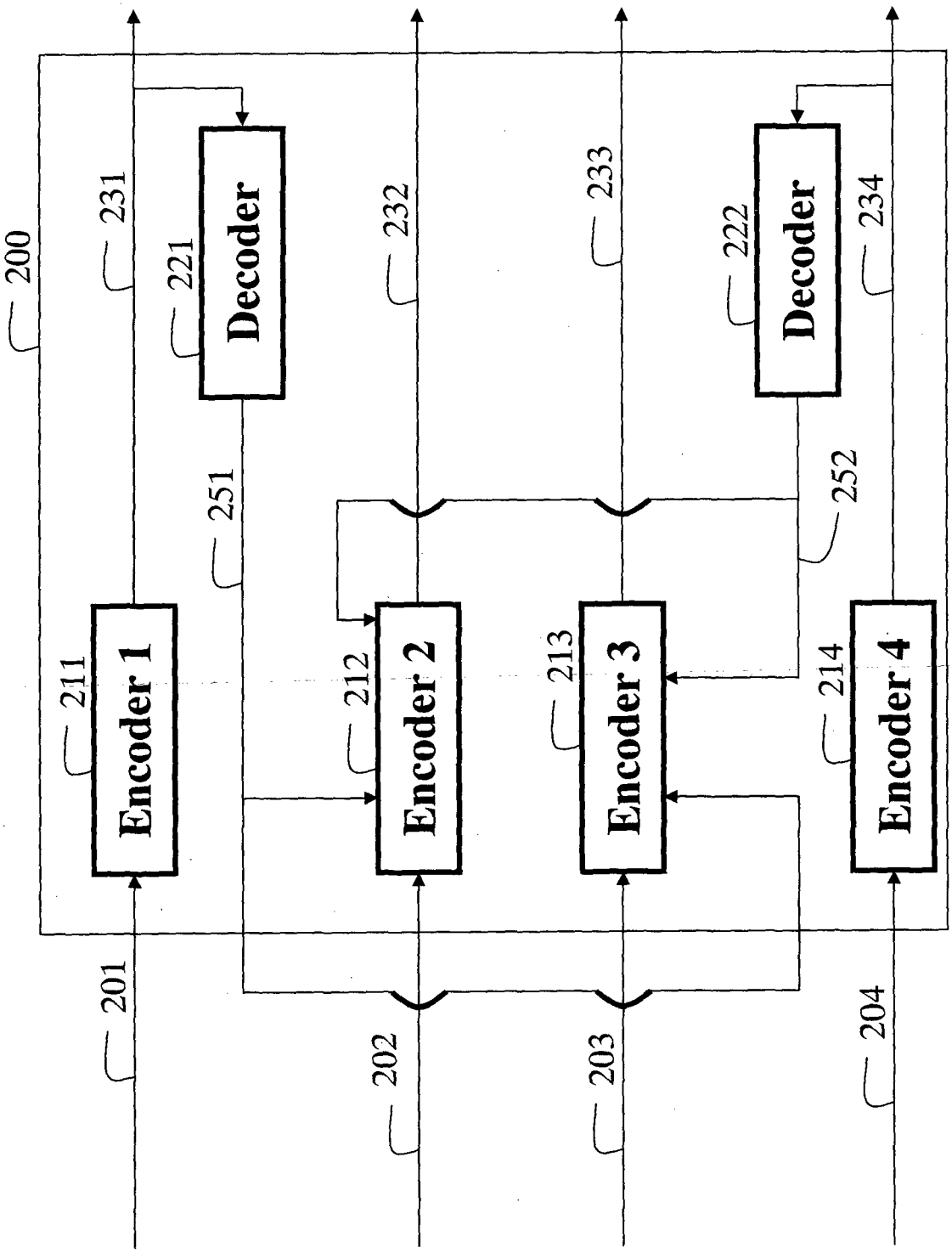

200
231
221
Decoder
232
233
222
Decoder
234
251
252
211
Encoder 1
212
Encoder 2
213
Encoder 3
214
Encoder 4
201
202
203
204

FIG. 2

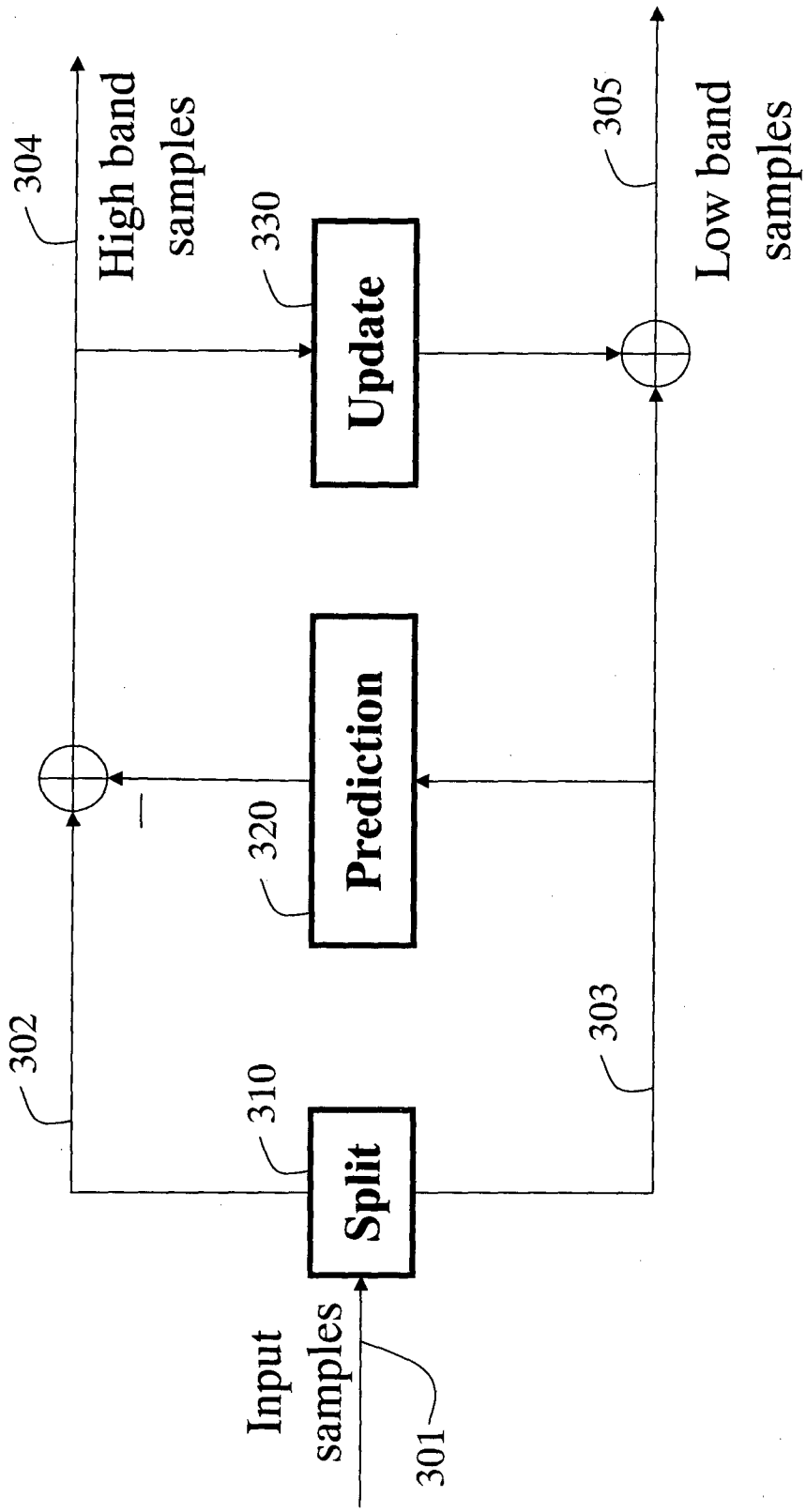

Input samples
301
Split
310
302
303
Prediction
320
Update
330
304
High band samples
305
Low band samples
−

FIG. 3

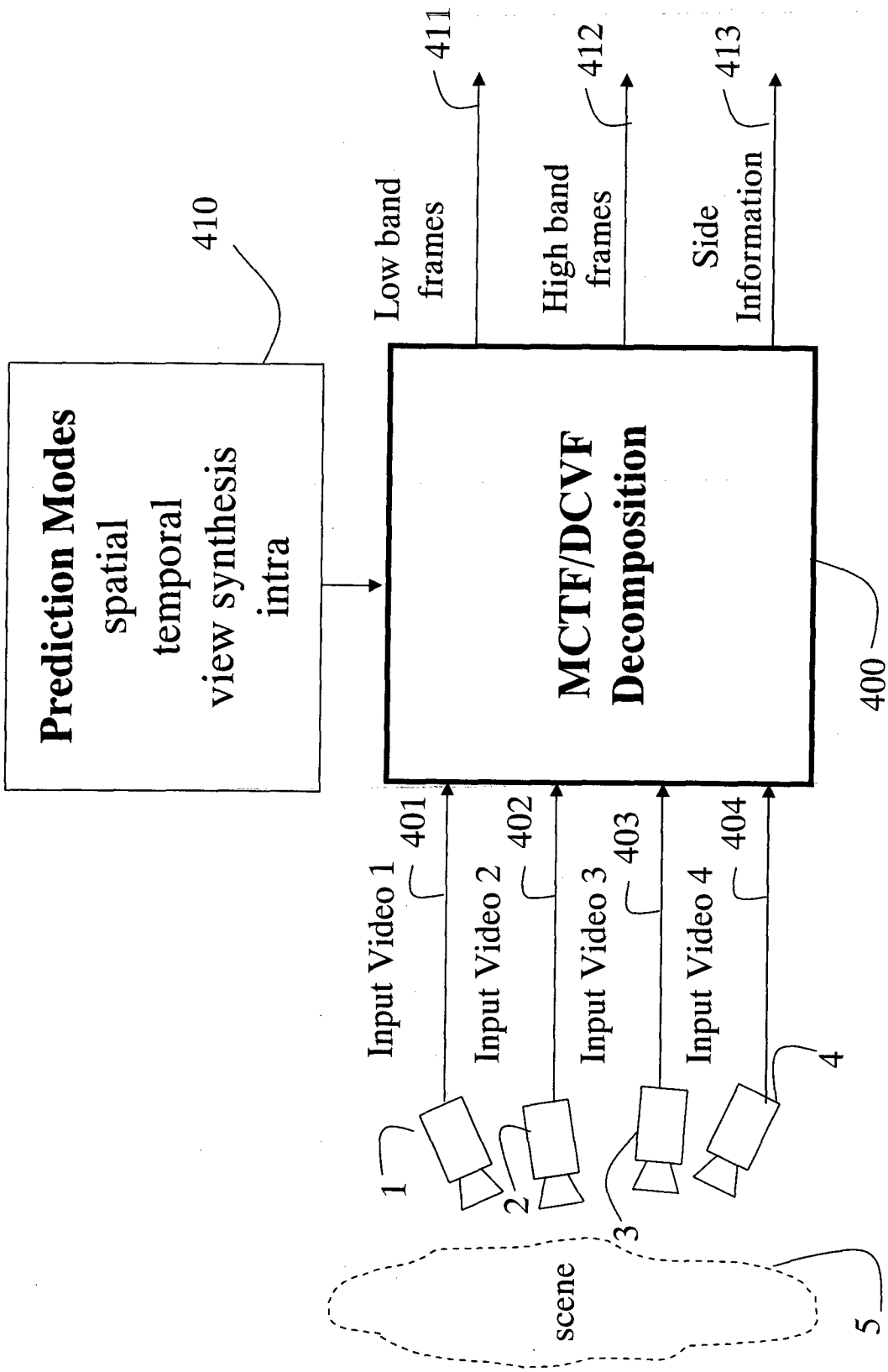
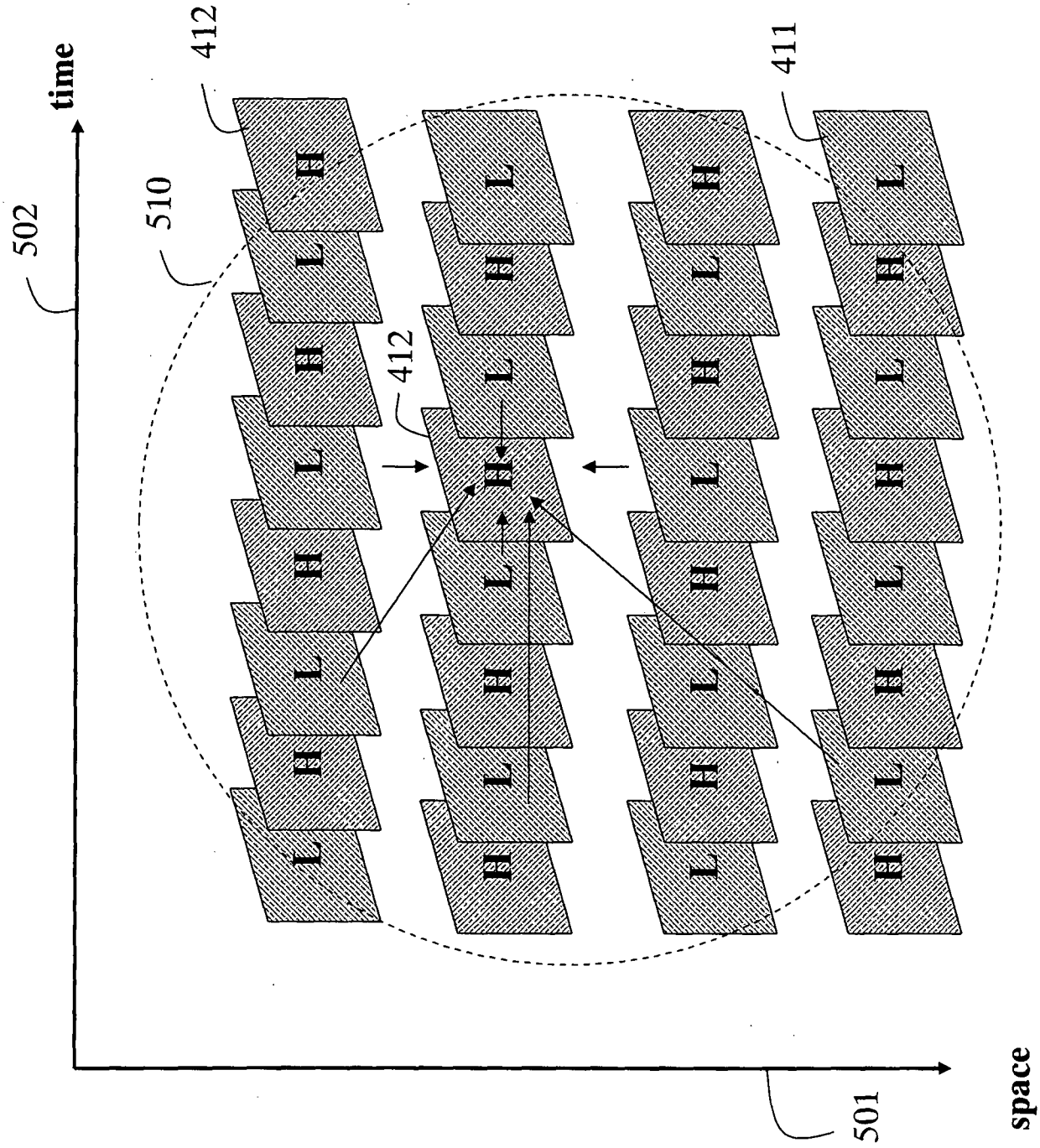

Prediction Modes
spatial
temporal
view synthesis
intra
410
MCTF/DCVF
Decomposition
400
Low band
frames
411
High band
frames
412
Side
Information
413
Input Video 1
401
Input Video 2
402
Input Video 3
403
Input Video 4
404
1
2
3
4
scene
5

FIG. 4

FIG. 5



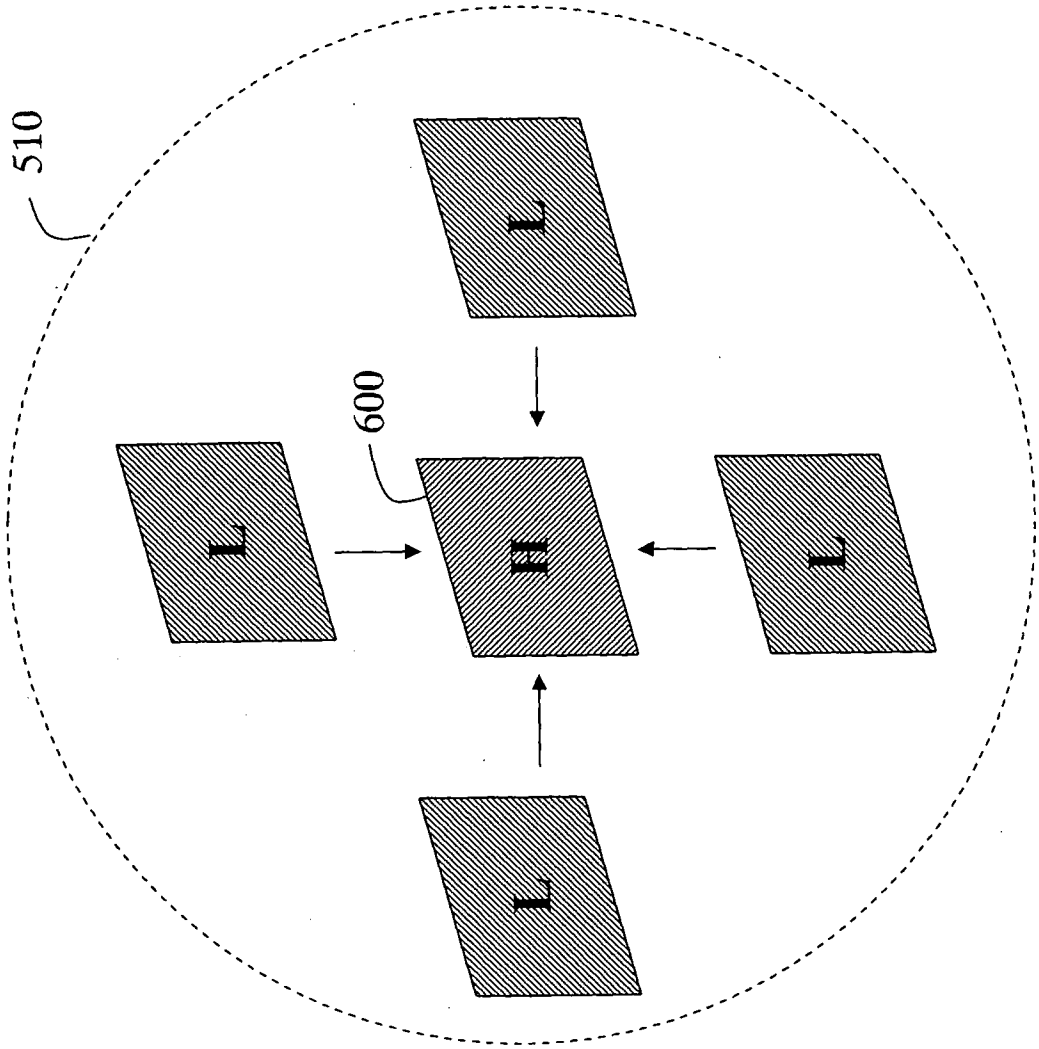


FIG. 6

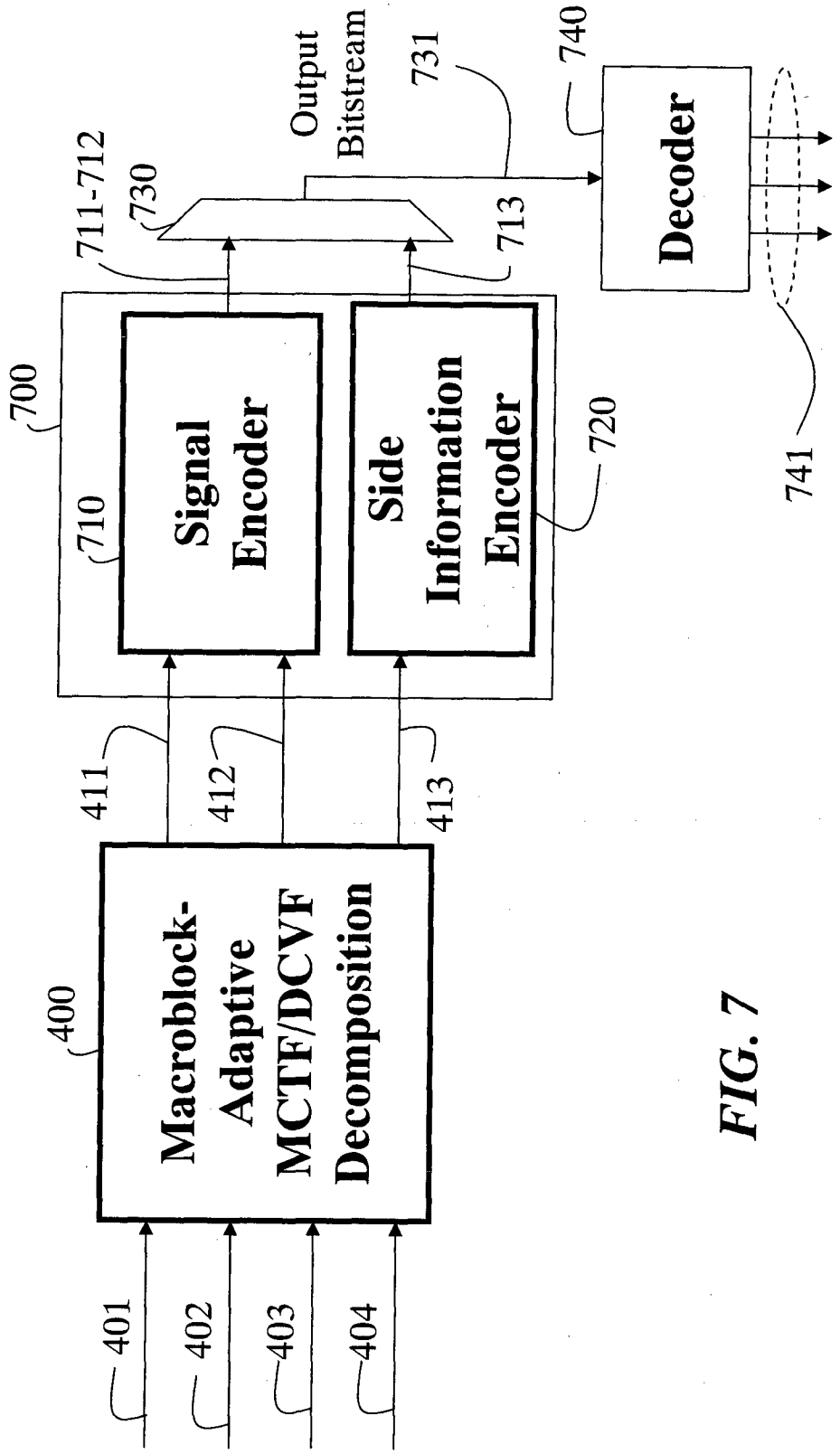

Macroblock-Adaptive MCTF/DCVF Decomposition
Signal Encoder
Side Information Encoder
Output Bitstream
Decoder
400
401
402
403
404
411
412
413
700
710
711-712
720
730
731
740
741
713

FIG. 7

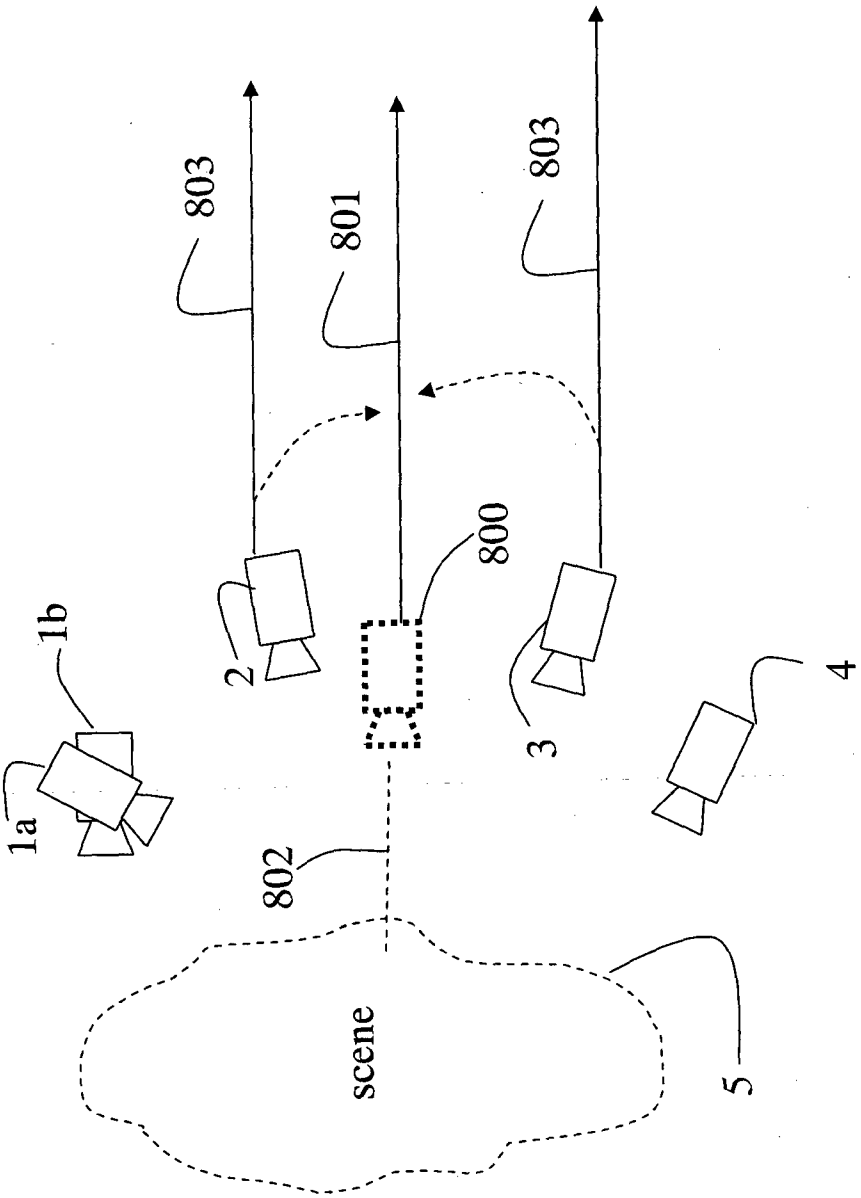


FIG. 8

FIG. 9

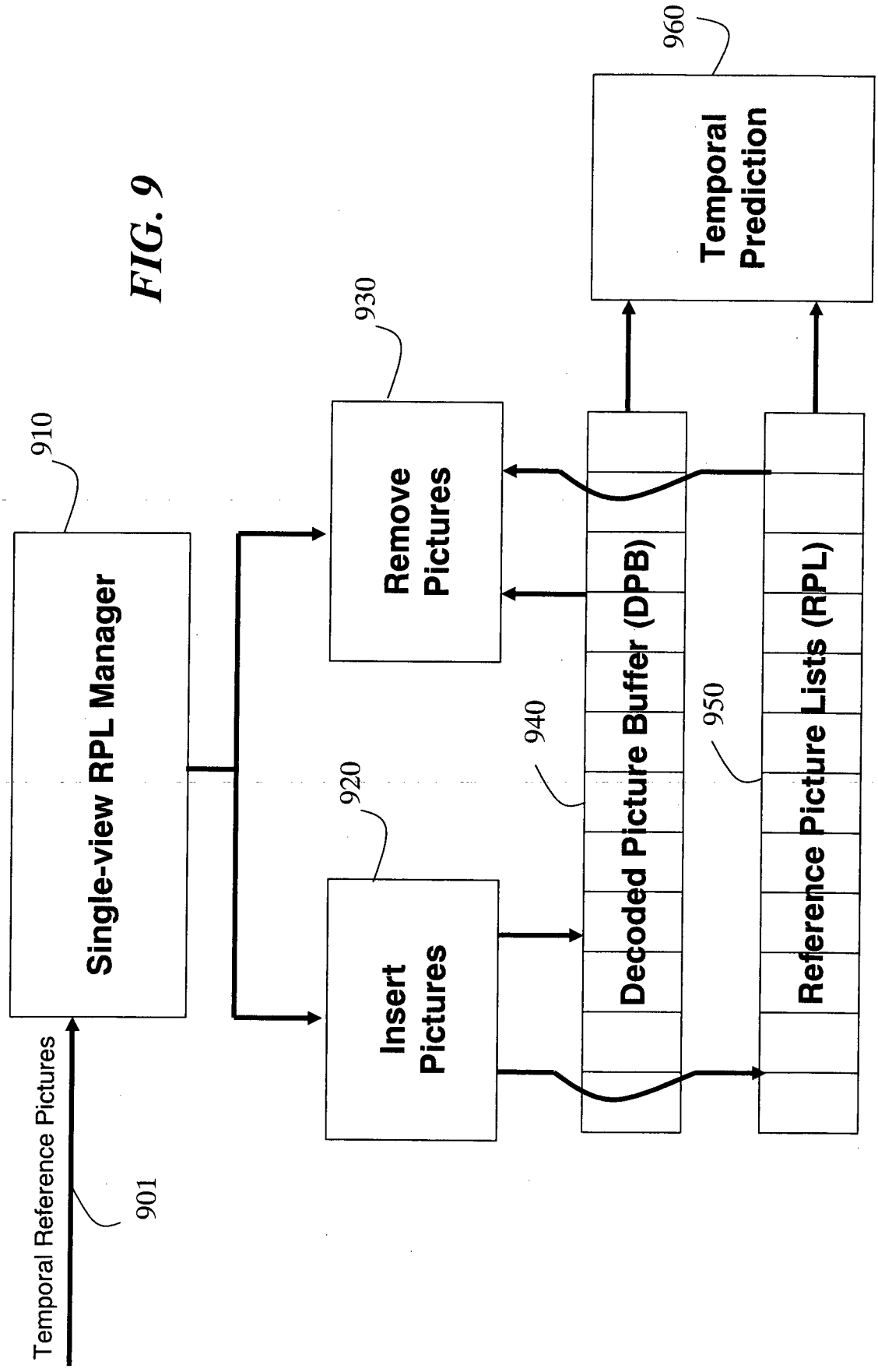
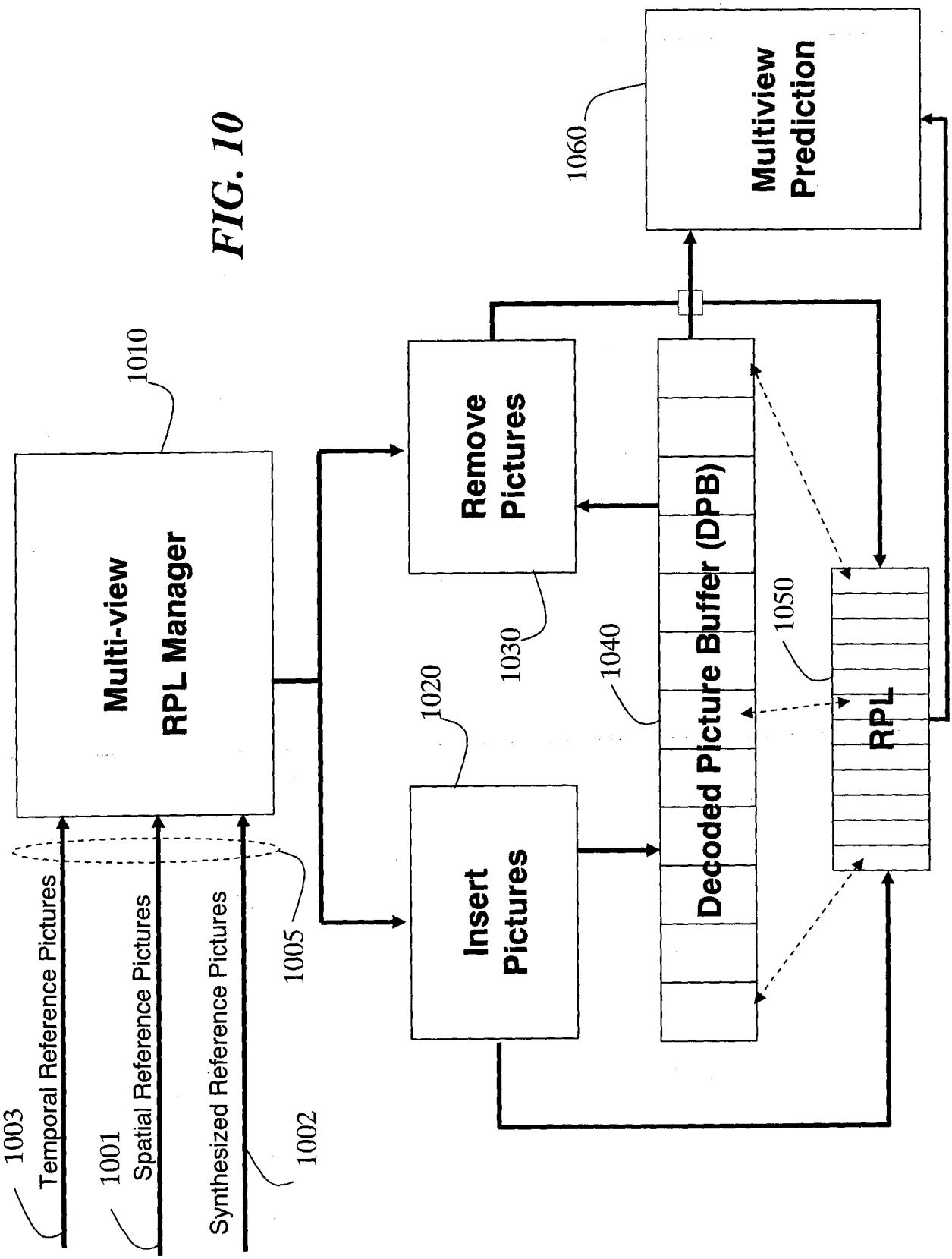

Temporal Reference Pictures
901
Single-view RPL Manager
910
Insert Pictures
920
Remove Pictures
930
Decoded Picture Buffer (DPB)
940
Reference Picture Lists (RPL)
950
Temporal Prediction
960

FIG. 10


1003
Temporal Reference Pictures
1001
Spatial Reference Pictures
Synthesized Reference Pictures
1002
1005
1010
Multi-view
RPL Manager
Remove
Pictures
1030
Insert
Pictures
1020
1040
Decoded Picture Buffer (DPB)
1050
RPL
1060
Multiview
Prediction

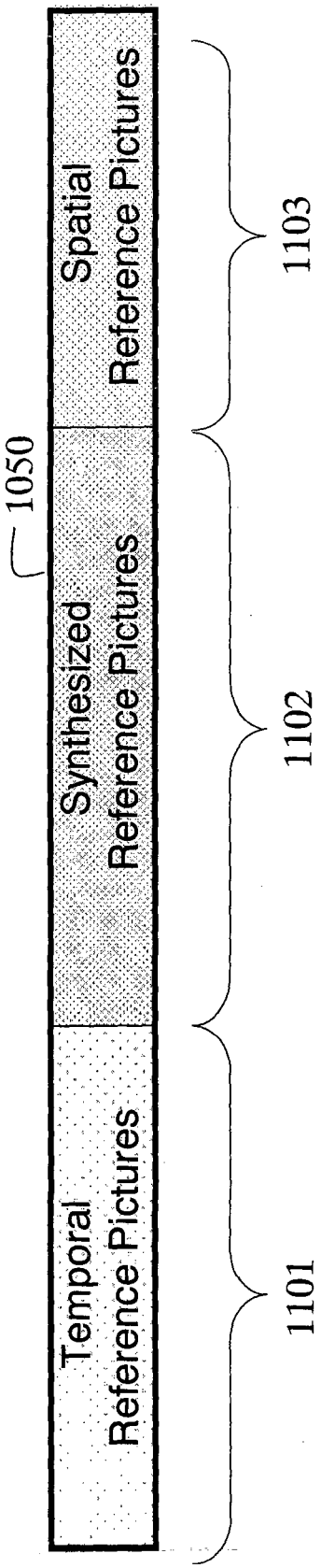

1050
Temporal Reference Pictures
Synthesized Reference Pictures
Spatial Reference Pictures
1101
1102
1103

FIG. 11

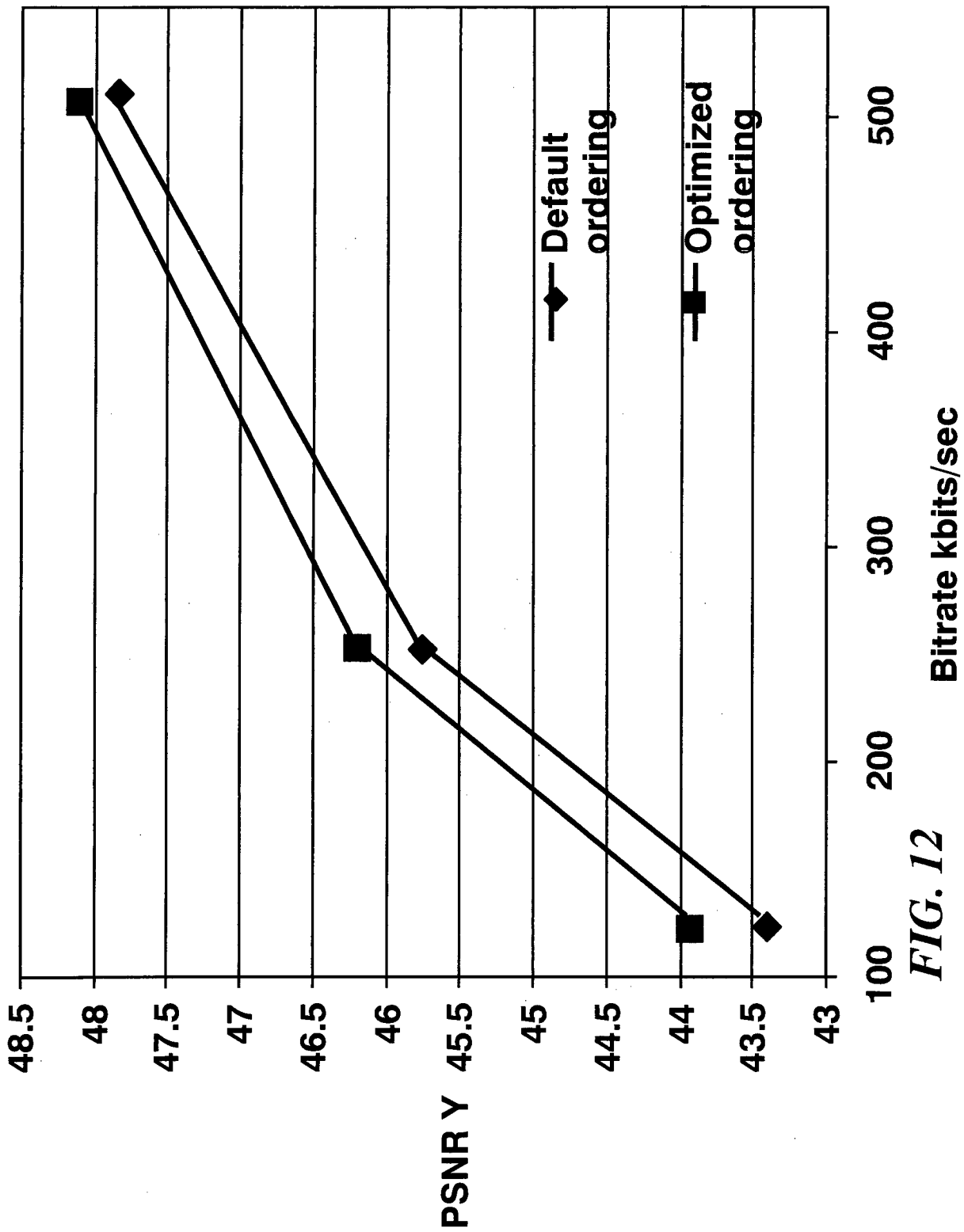

PSNR Y
48.5
48
47.5
47
46.5
46
45.5
45
44.5
44
43.5
43
100
200
300
400
500
Bitrate kbits/sec
Default ordering
Optimized ordering

FIG. 12

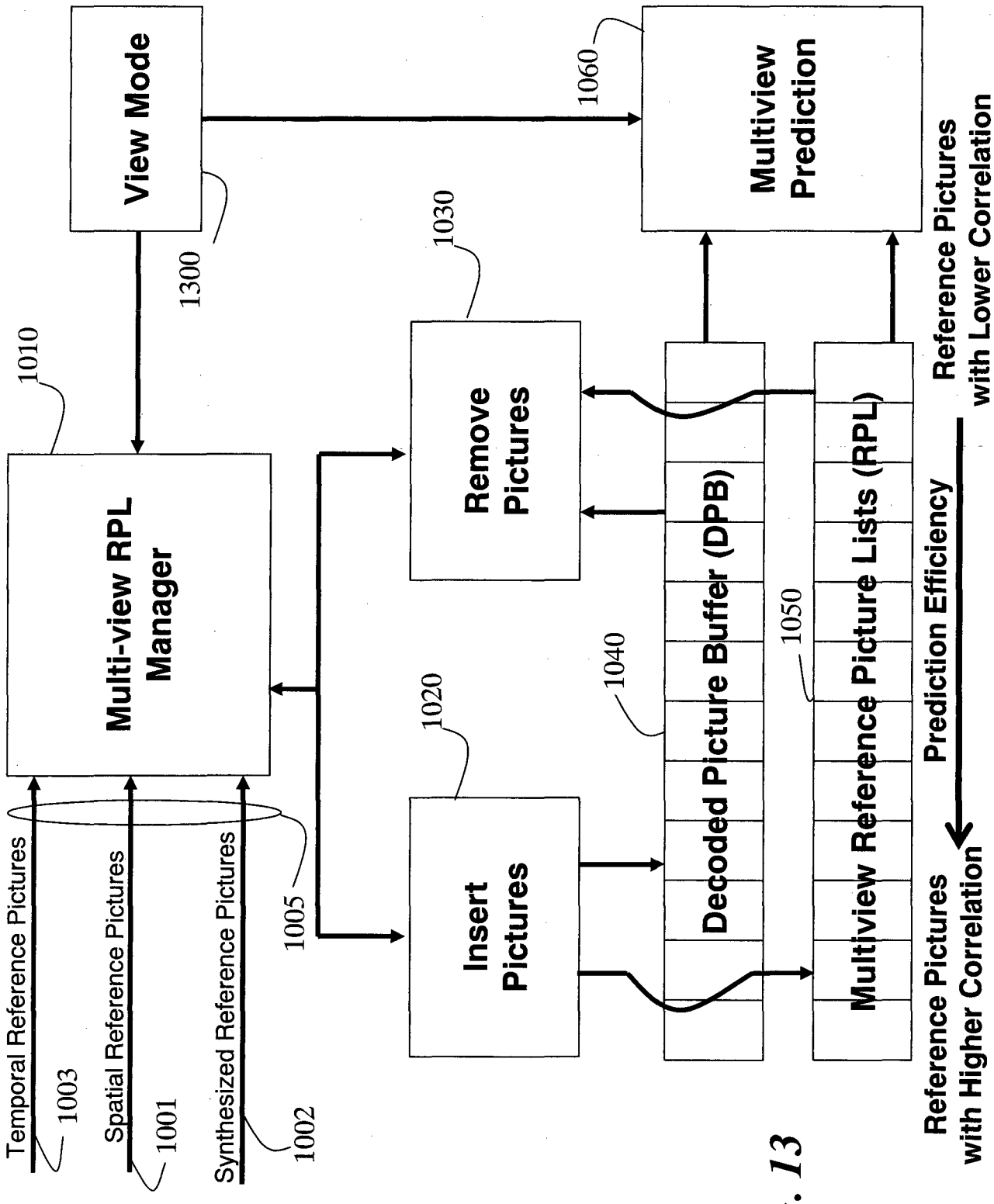

View Mode
1300
1060
Multiview Prediction
1030
1010
Remove Pictures
Multi-view RPL Manager
1020
Insert Pictures
1040
Decoded Picture Buffer (DPB)
1050
Multiview Reference Picture Lists (RPL)
Reference Pictures with Lower Correlation
Prediction Efficiency
Reference Pictures with Higher Correlation
Temporal Reference Pictures
1003
Spatial Reference Pictures
1001
Synthesized Reference Pictures
1002
1005

FIG. 13

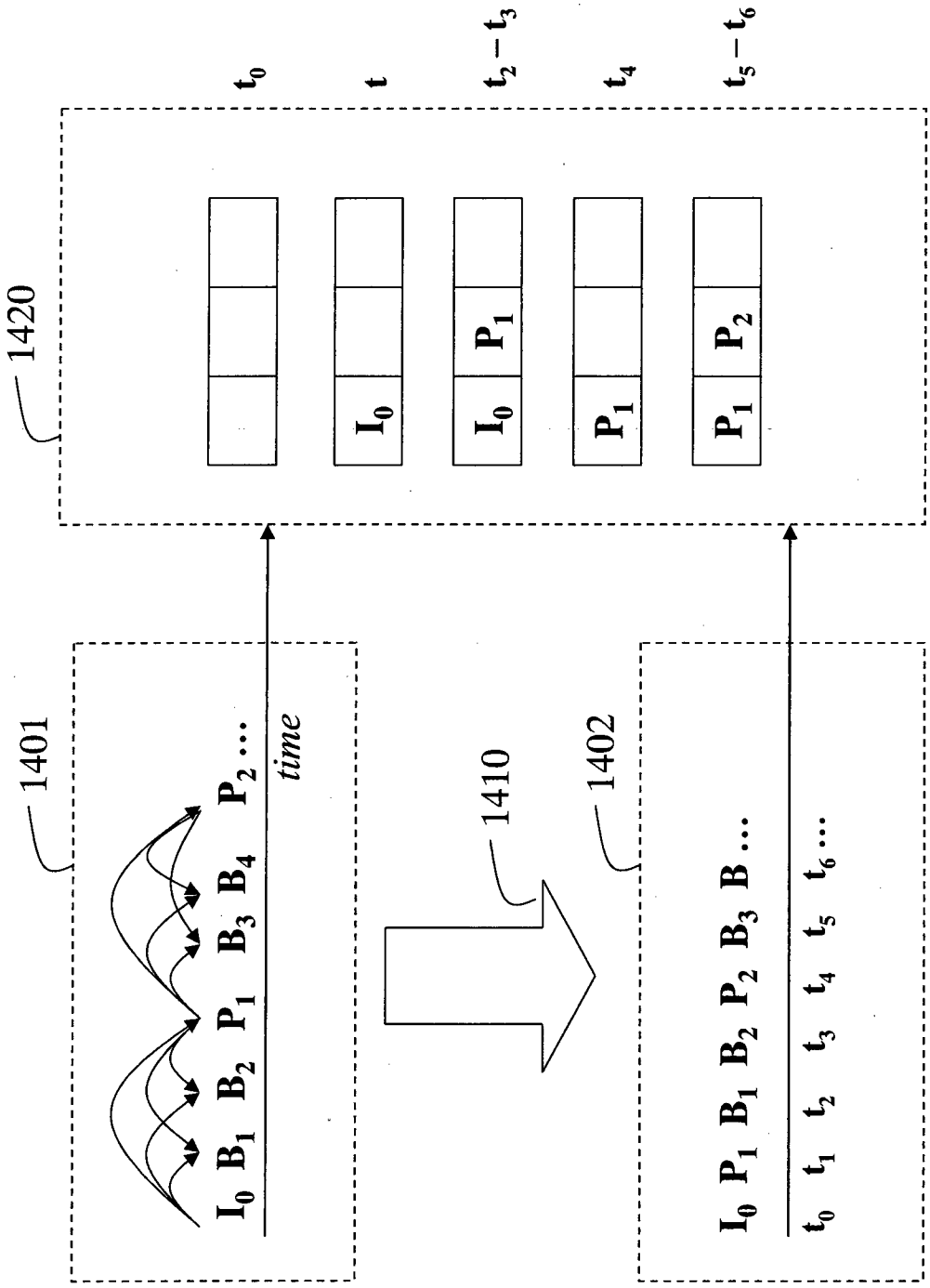

1420
t_0
t
$t_2 - t_3$
t_4
$t_5 - t_6$
I_0
I_0 P_1
P_1
P_1 P_2
1401
I_0 B_1 B_2 P_1 B_3 B_4 P_2 …
time
1410
1402
I_0 P_1 B_1 B_2 P_2 B_3 B…
t_0 t_1 t_2 t_3 t_4 t_5 t_6…

FIG. 14

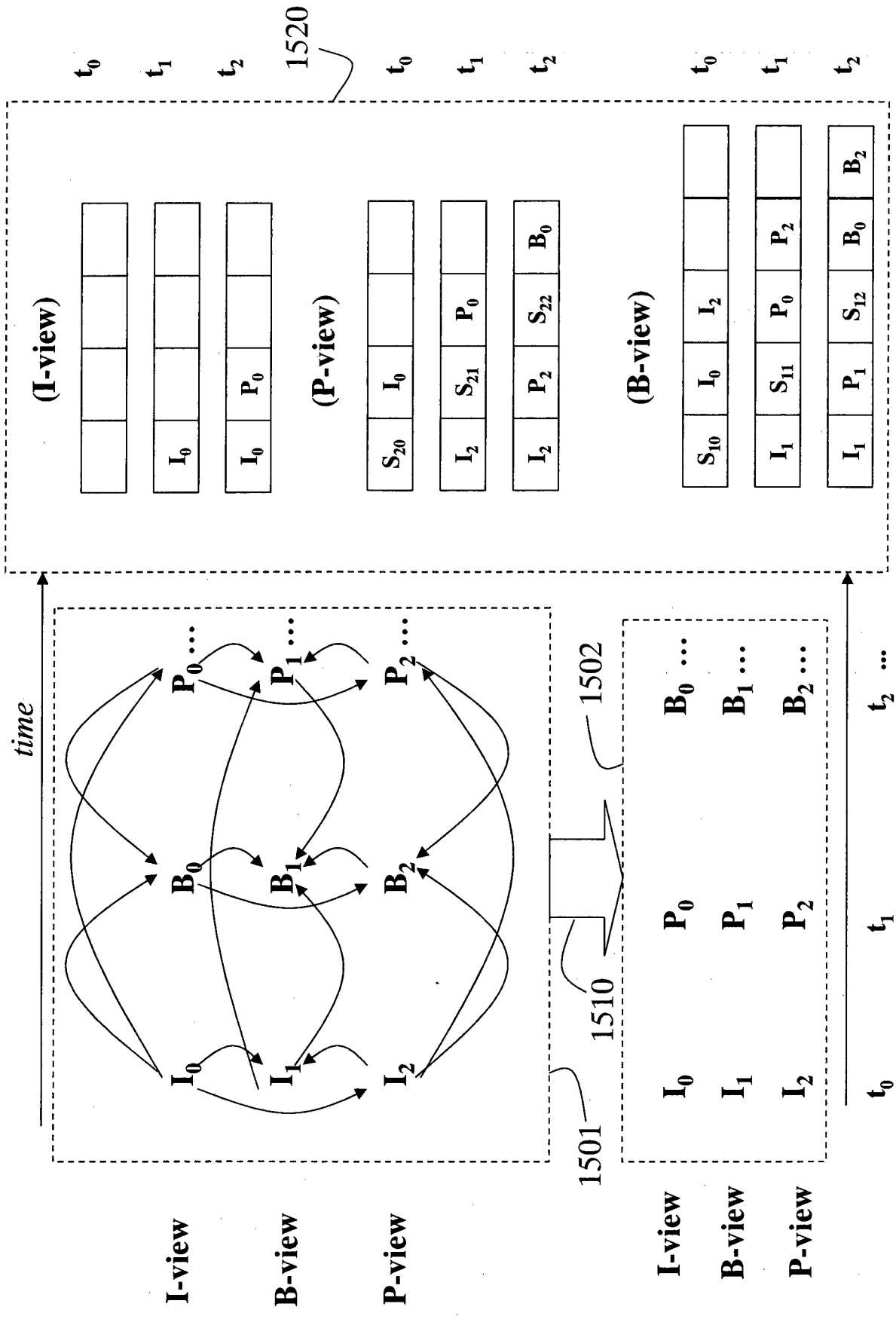

time
I-view
B-view
P-view
1501
1502
1510
1520
(I-view)
(P-view)
(B-view)
t0
t1
t2

FIG. 15

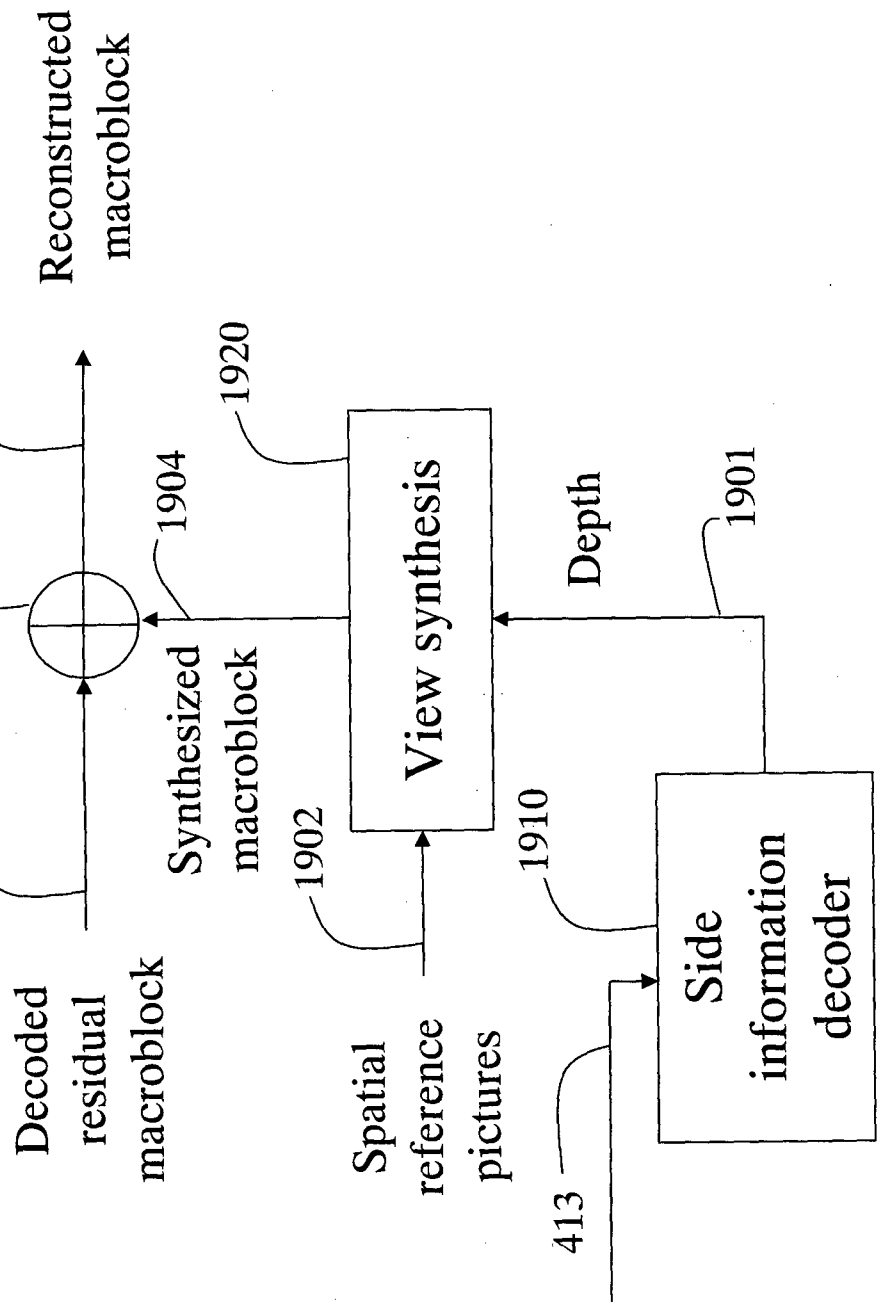

Reconstructed macroblock
1903
1930
1904
1920
1905
Synthesized macroblock
View synthesis
Depth
1901
1902
1910
Decoded residual macroblock
Spatial reference pictures
Side information decoder
413

FIG. 16

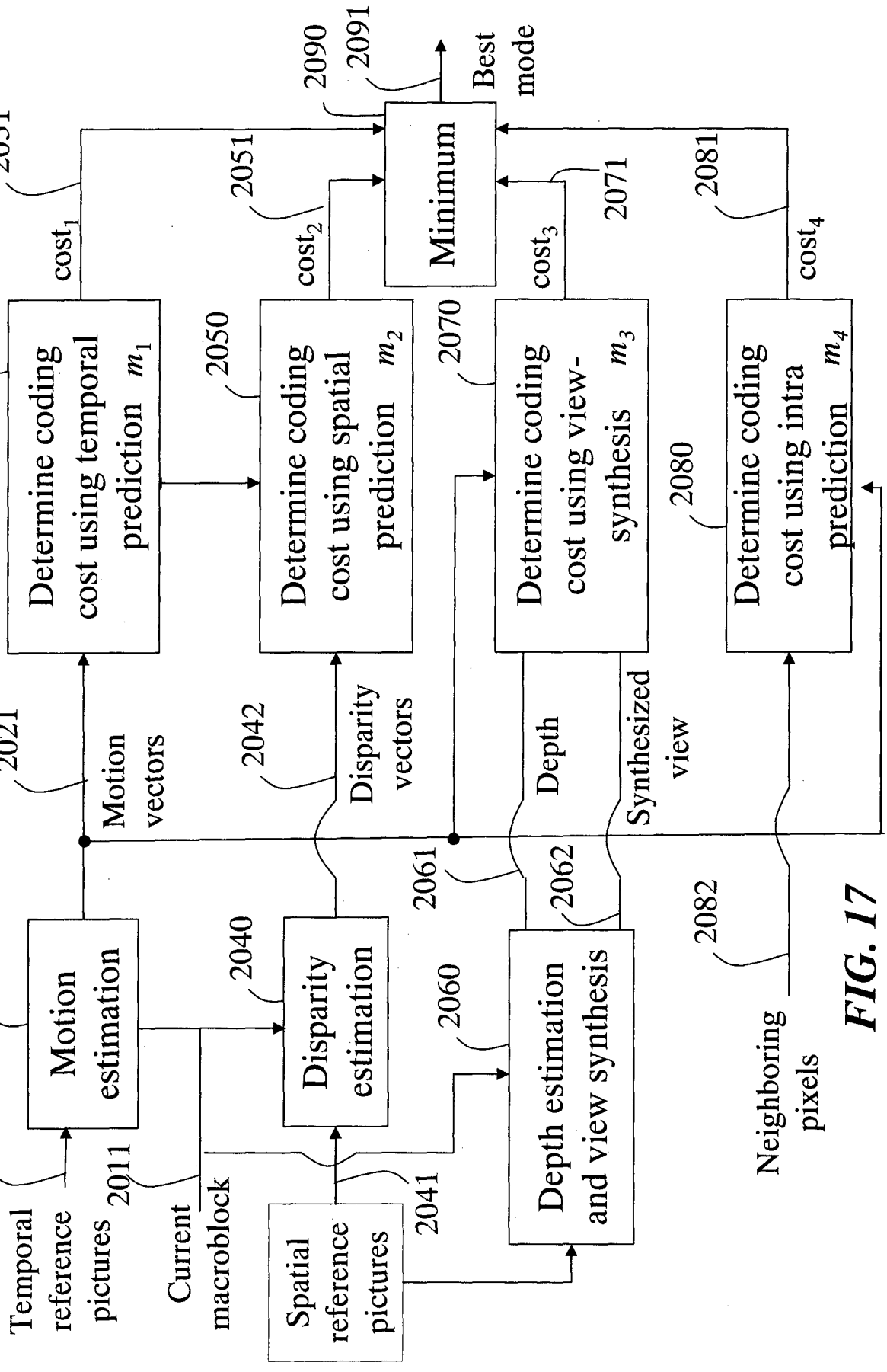

Temporal reference pictures
2020
Motion estimation
2010
Current macroblock
2011
Spatial reference pictures
2041
Disparity estimation
2040
Depth estimation and view synthesis
2060
Neighboring pixels
2082
2061
2062
Motion vectors
2021
Disparity vectors
2042
Depth
Synthesized view
Determine coding cost using temporal prediction m_1
2030
Determine coding cost using spatial prediction m_2
2050
Determine coding cost using view-synthesis m_3
2070
Determine coding cost using intra prediction m_4
2080
$cost_1$
2031
$cost_2$
2051
$cost_3$
2071
$cost_4$
2081
Minimum
2090
2091
Best mode

FIG. 17

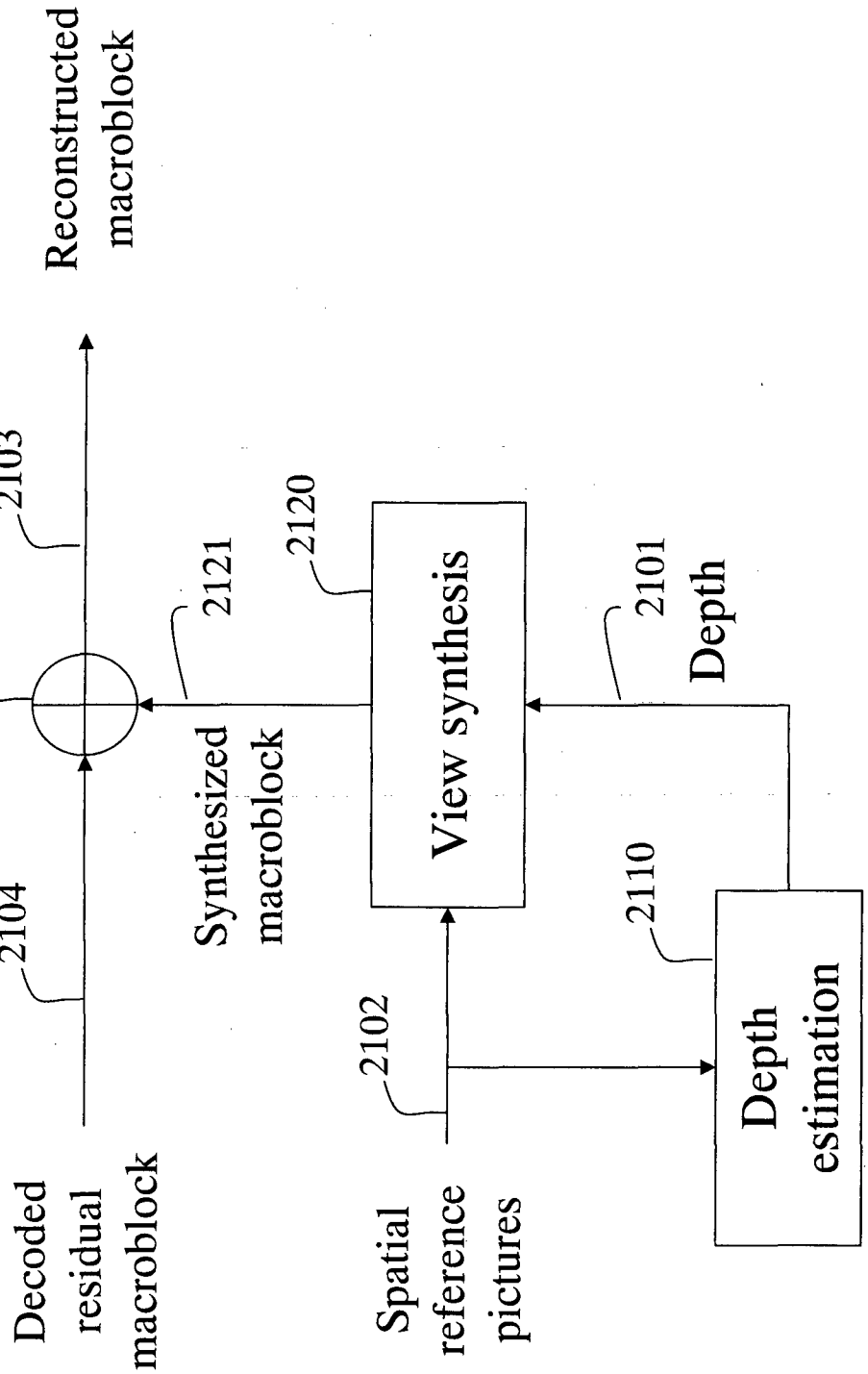

Reconstructed macroblock
2103
2130
2121
2120
2101
Depth
Synthesized macroblock
View synthesis
2104
2110
2102
Decoded residual macroblock
Spatial reference pictures
Depth estimation

FIG. 18

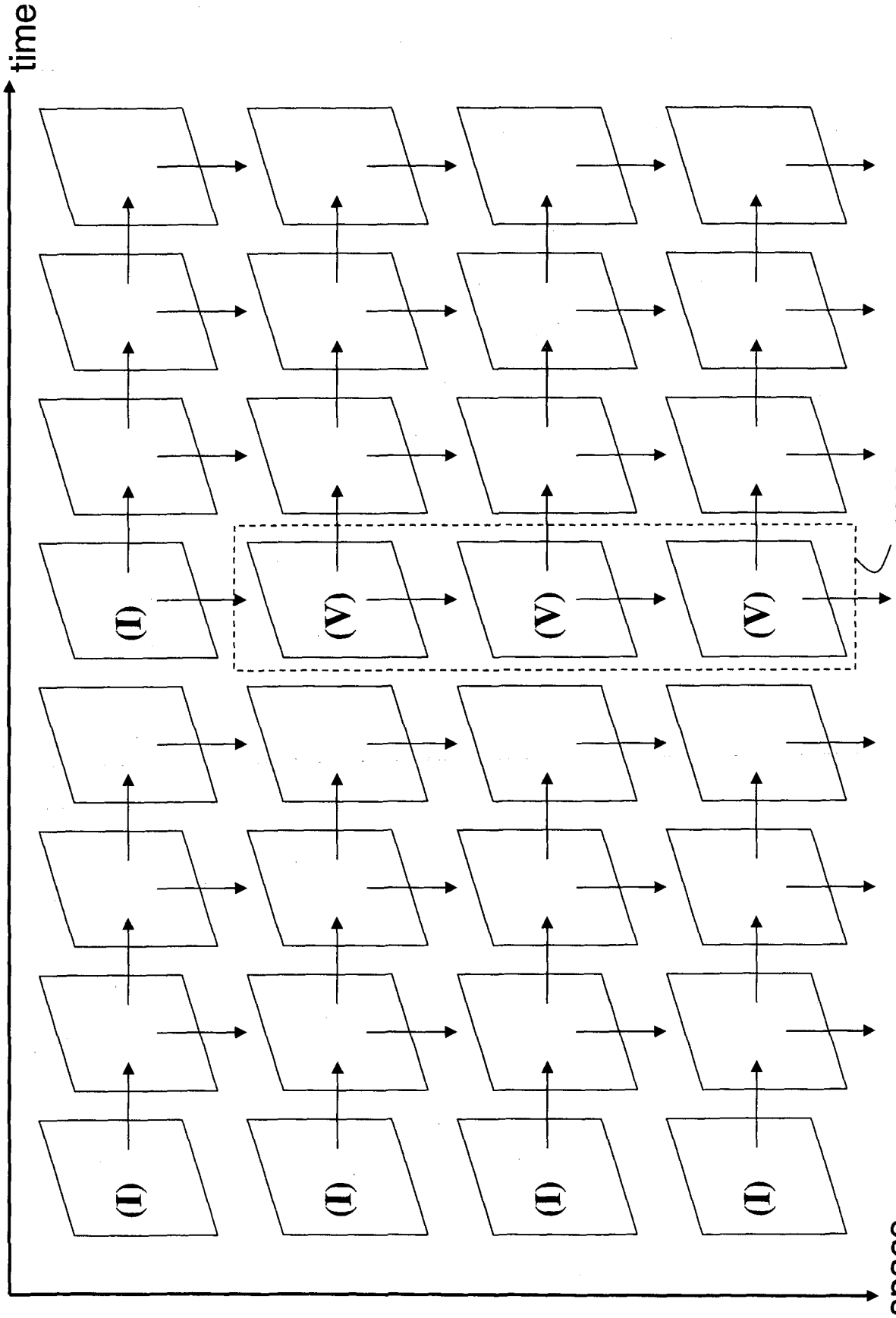


FIG. 19

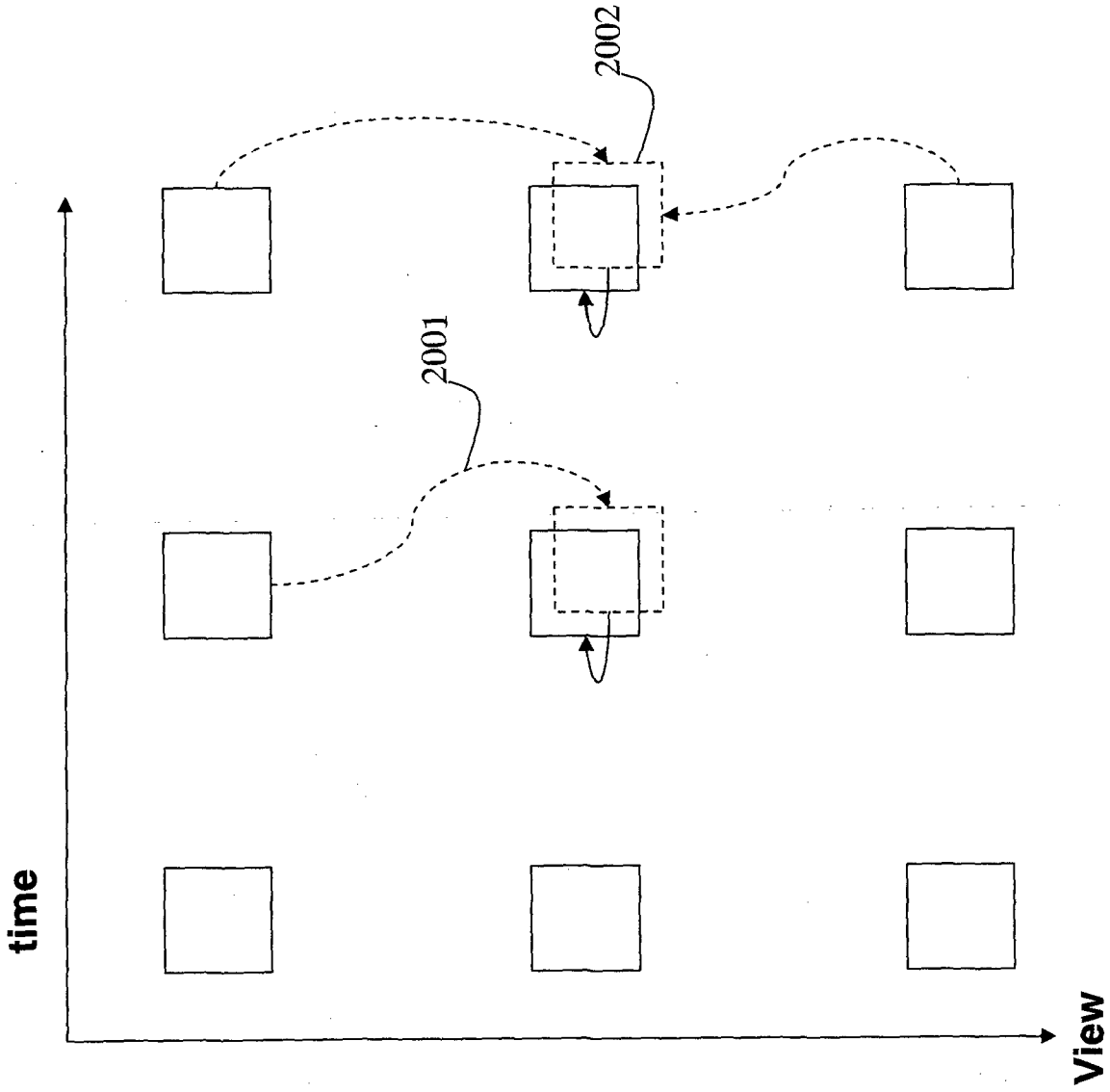


FIG. 20

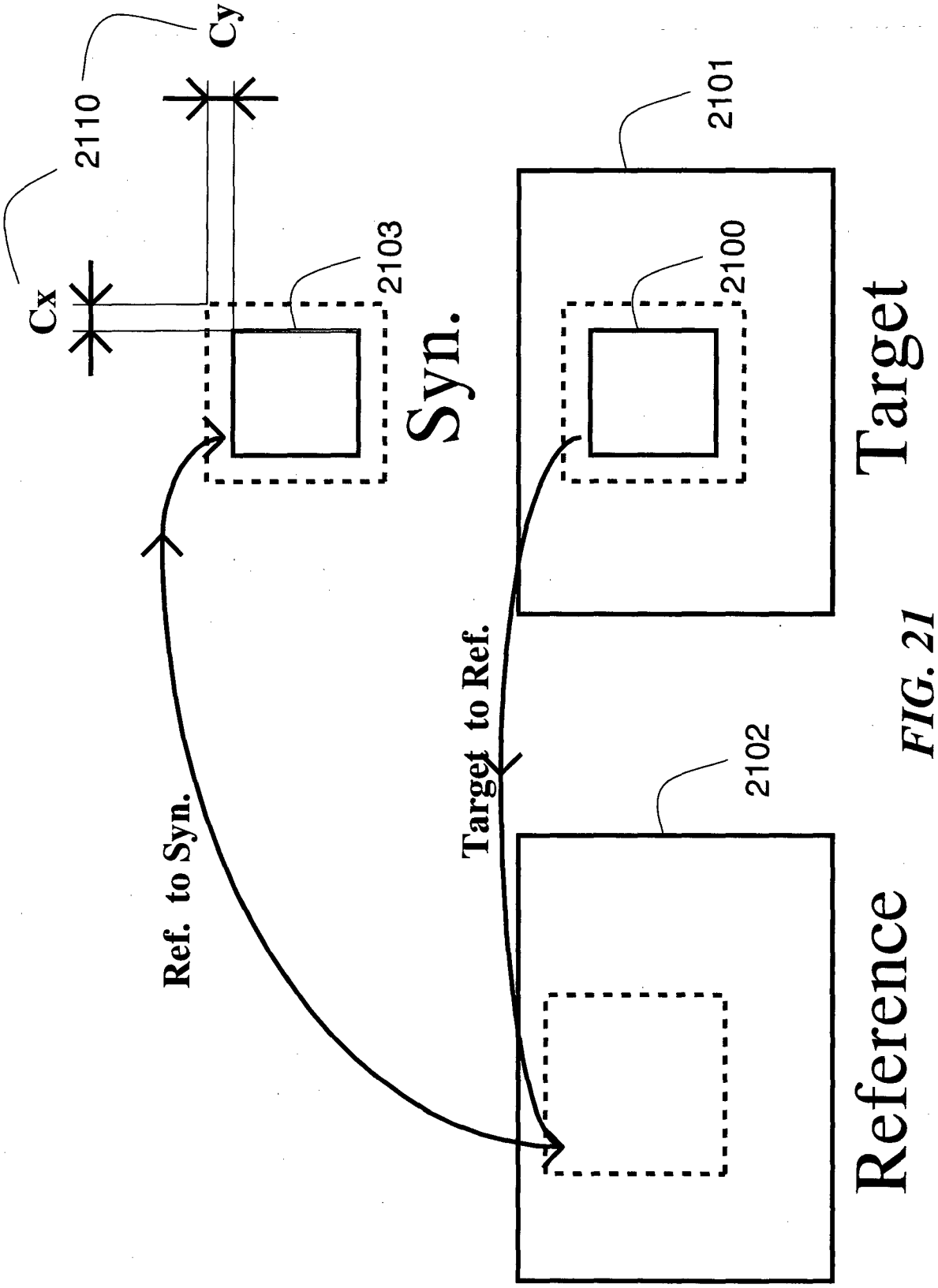

Cy
2110
Cx
2103
Syn.
2101
2100
Target
Ref. to Syn.
Target to Ref.
2102
Reference

FIG. 21

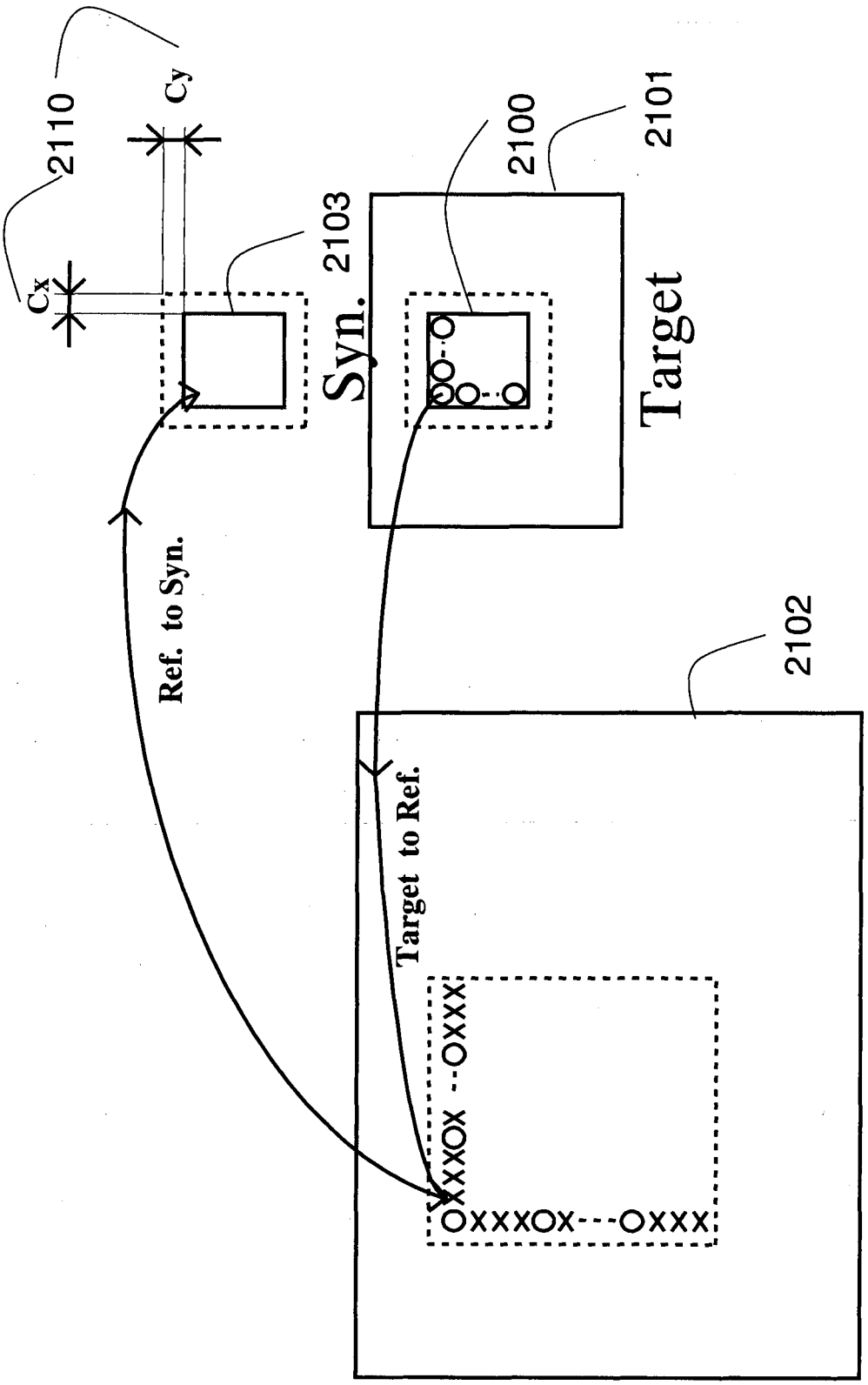

2110
Cy
Cx
2103
Syn.
2100
2101
Target
Ref. to Syn.
2102
Target to Ref.
OXXXOX---OXXX
OXXOX --OXXX
Reference

FIG. 22

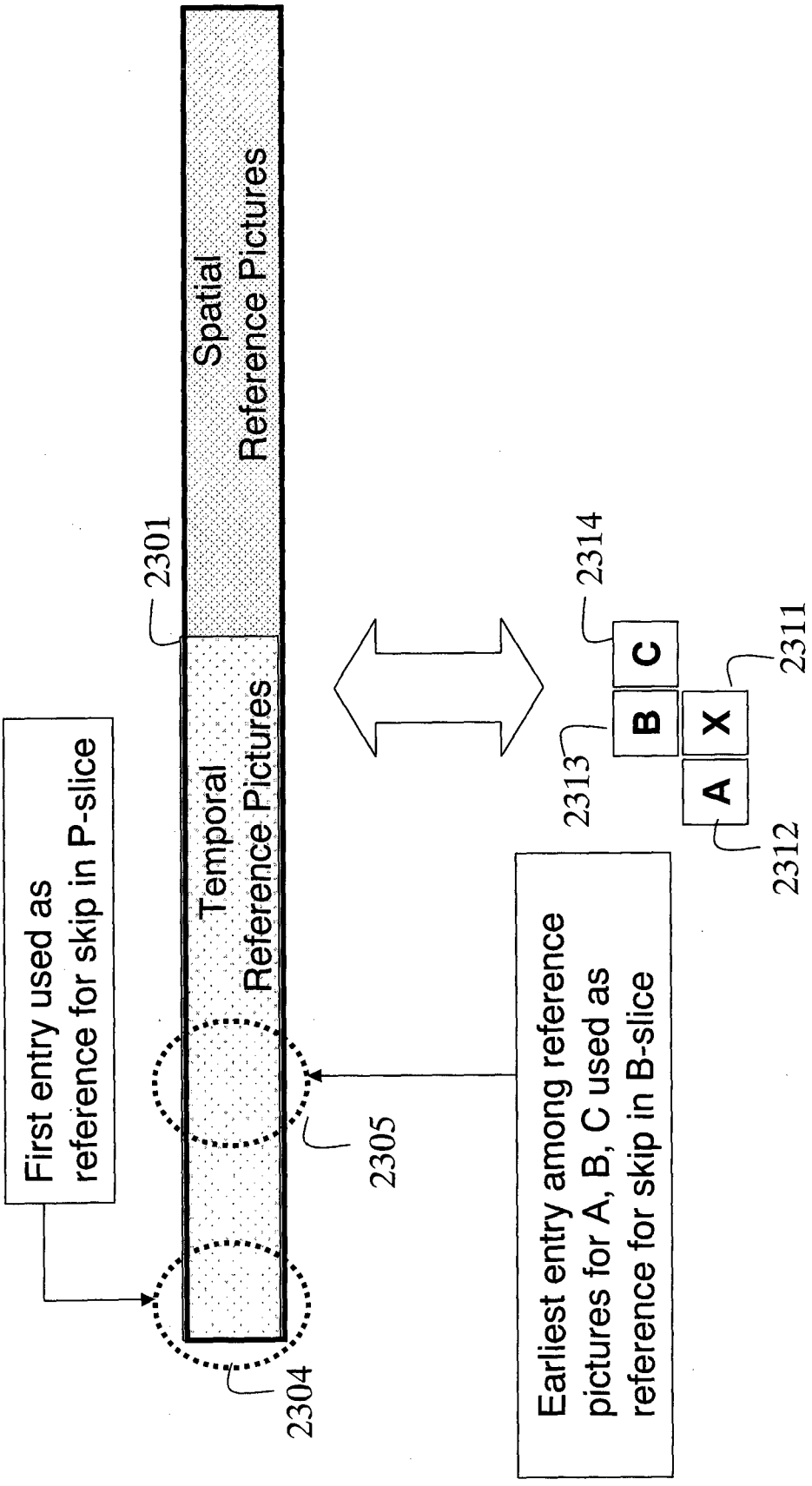

First entry used as reference for skip in P-slice
Temporal Reference Pictures
Spatial Reference Pictures
2301
2304
2305
2311
2312
2313
2314
A
X
B
C
Earliest entry among reference pictures for A, B, C used as reference for skip in B-slice

FIG. 23

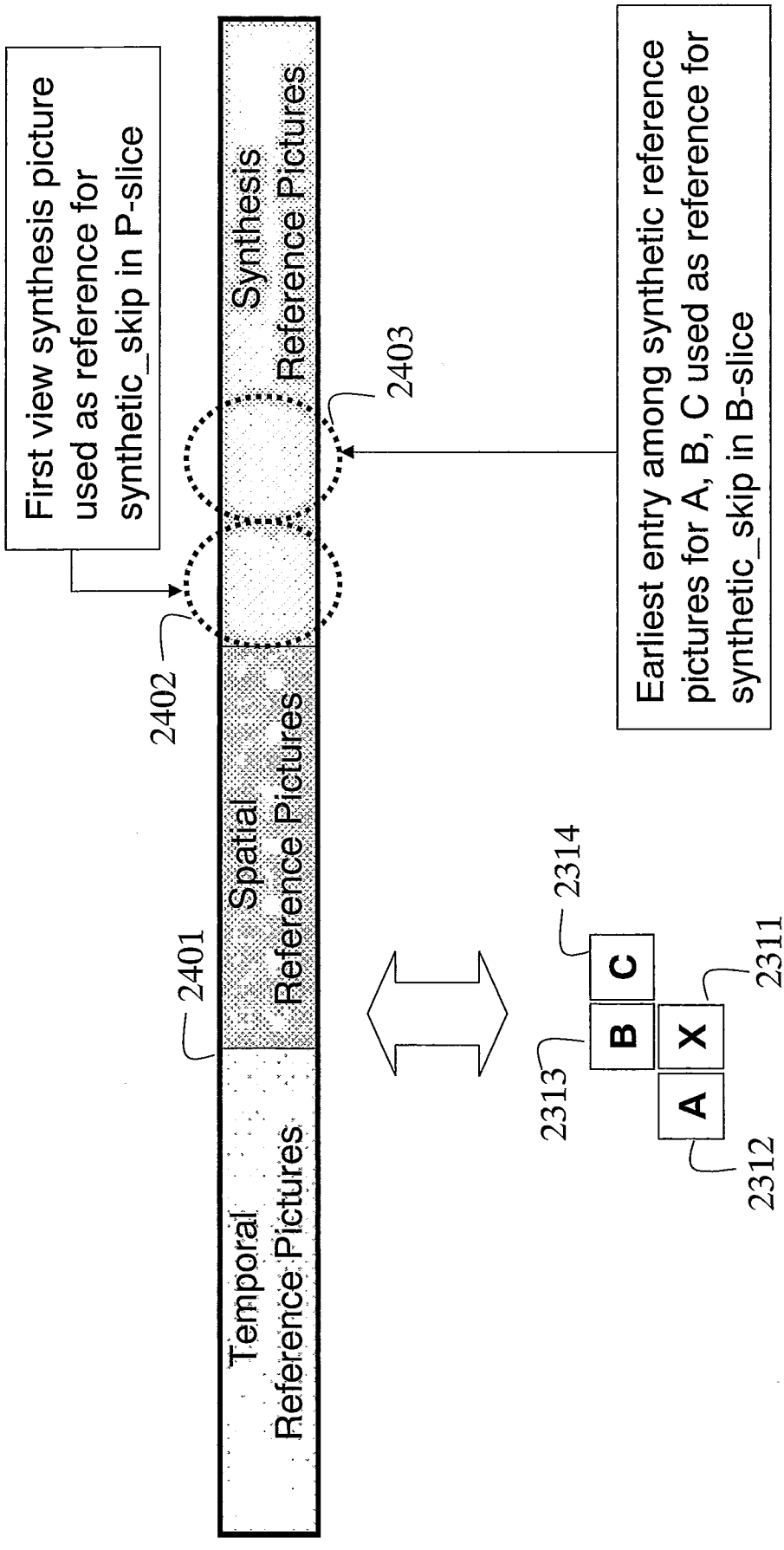

First view synthesis picture used as reference for synthetic_skip in P-slice
Synthesis Reference Pictures
2403
2402
Spatial Reference Pictures
2401
Temporal Reference Pictures
Earliest entry among synthetic reference pictures for A, B, C used as reference for synthetic_skip in B-slice
2314
C
2313
B
X
2311
A
2312

FIG. 24

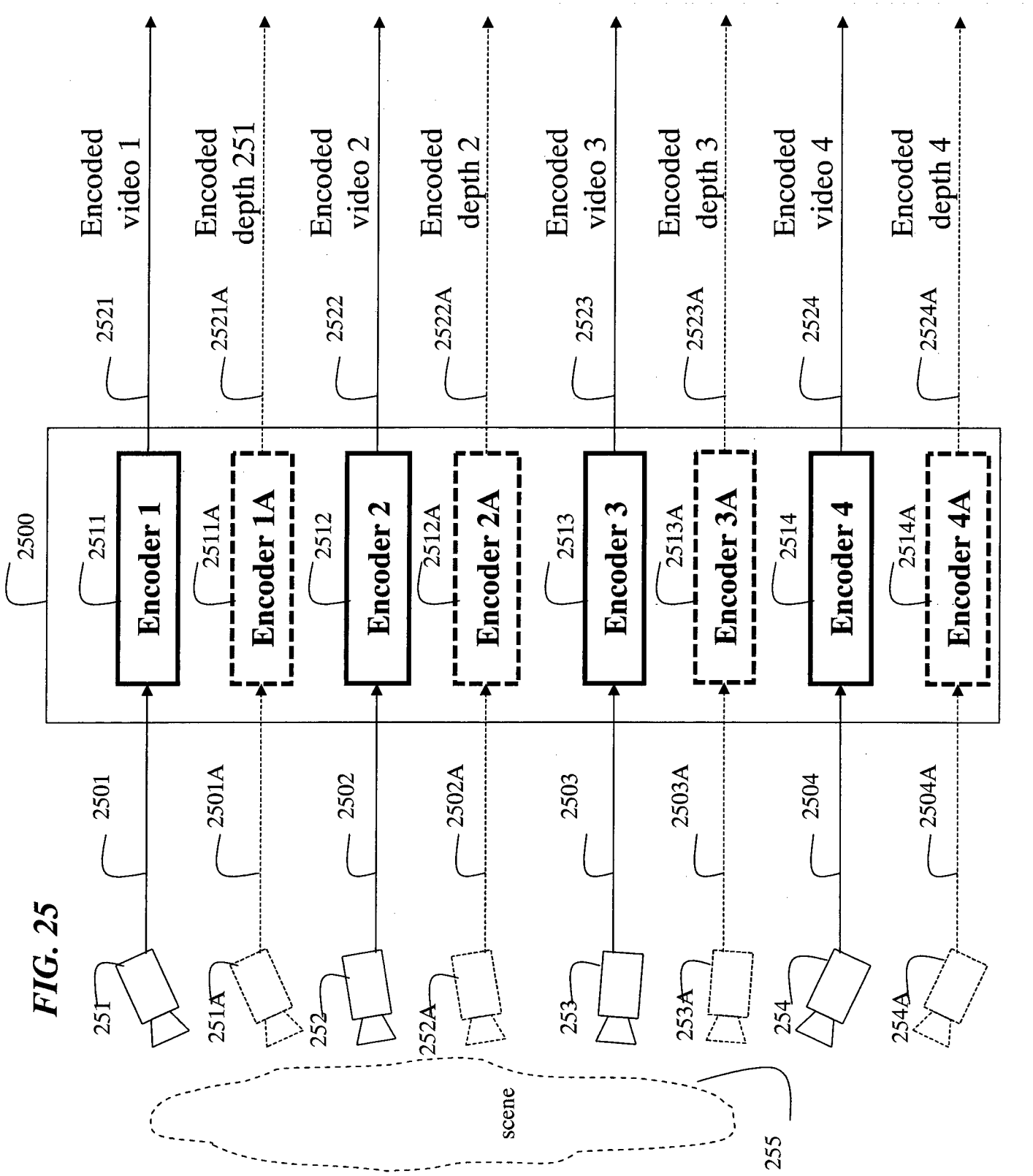

2500
2511
Encoder 1
2511A
Encoder 1A
2512
Encoder 2
2512A
Encoder 2A
2513
Encoder 3
2513A
Encoder 3A
2514
Encoder 4
2514A
Encoder 4A
2521
Encoded video 1
2521A
Encoded depth 251
2522
Encoded video 2
2522A
Encoded depth 2
2523
Encoded video 3
2523A
Encoded depth 3
2524
Encoded video 4
2524A
Encoded depth 4
2501
2501A
2502
2502A
2503
2503A
2504
2504A
251
251A
252
252A
253
253A
254
254A
scene
255

FIG. 25

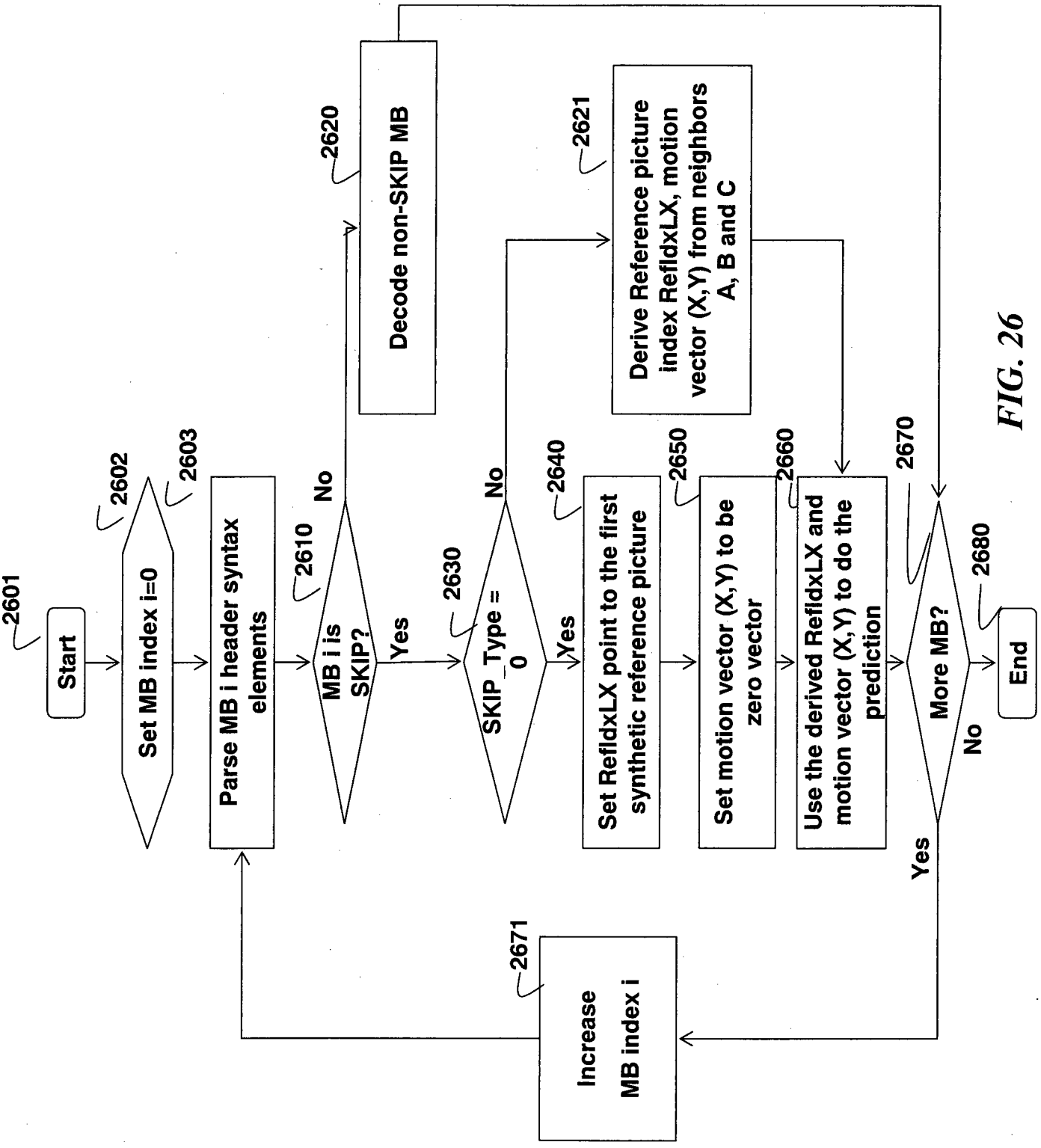

2601
Start
2602
Set MB index i=0
2603
Parse MB i header syntax elements
2610
MB i is SKIP?
No
Yes
2620
Decode non-SKIP MB
2630
SKIP_Type = 0
No
Yes
2621
Derive Reference picture index RefIdxLX, motion vector (X,Y) from neighbors A, B and C
2640
Set RefIdxLX point to the first synthetic reference picture
2650
Set motion vector (X,Y) to be zero vector
2660
Use the derived RefIdxLX and motion vector (X,Y) to do the prediction
2670
More MB?
Yes
No
2680
End
2671
Increase MB index i

FIG. 26

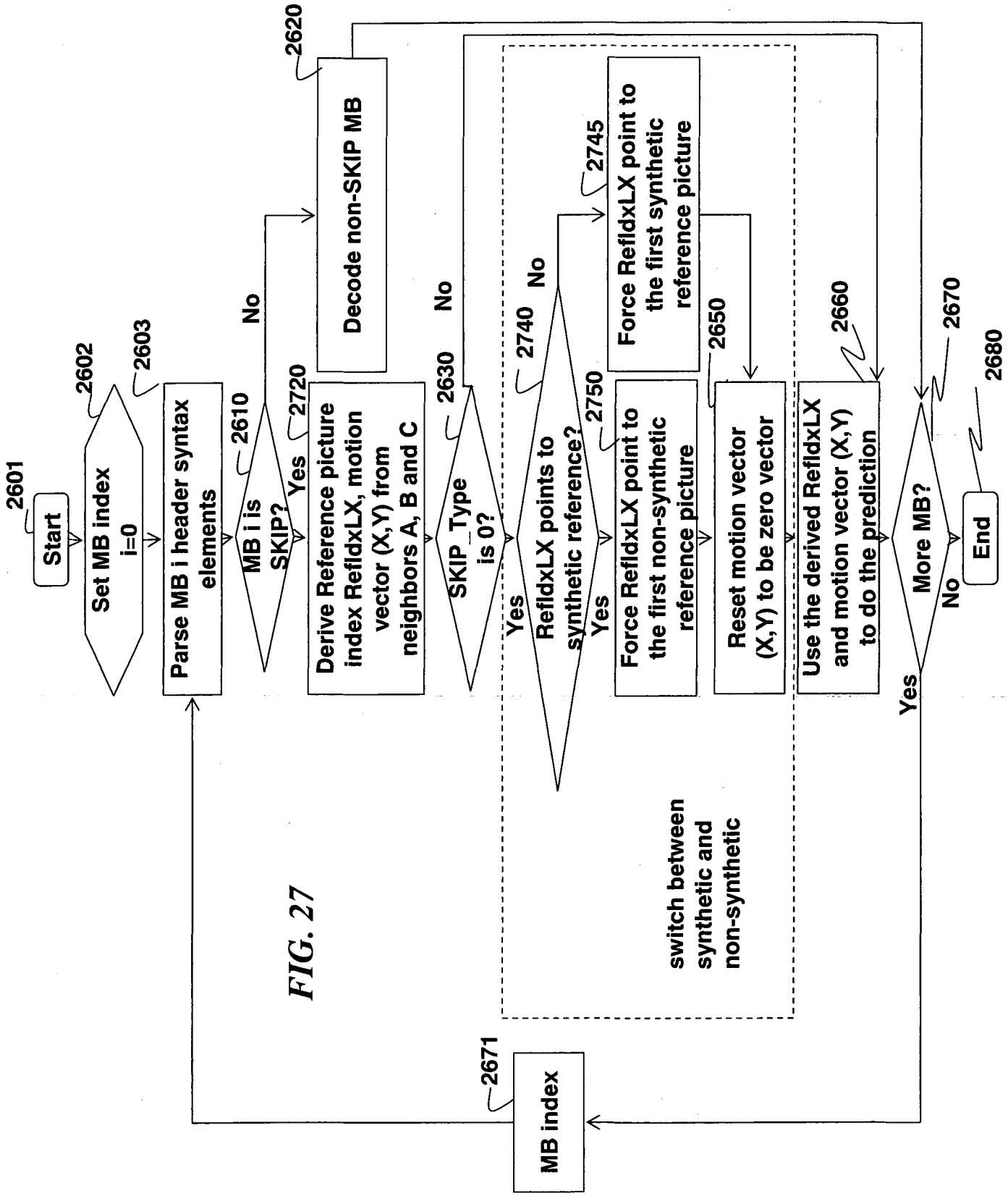

2601
Start
2602
Set MB index i=0
2603
Parse MB i header syntax elements
2610
MB i is SKIP?
No
Yes
2620
Decode non-SKIP MB
2720
Derive Reference picture index RefIdxLX, motion vector (X,Y) from neighbors A, B and C
2630
SKIP_Type is 0?
No
Yes
2740
RefIdxLX points to synthetic reference?
No
Yes
2745
Force RefIdxLX point to the first synthetic reference picture
2750
Force RefIdxLX point to the first non-synthetic reference picture
2650
Reset motion vector (X,Y) to be zero vector
2660
Use the derived RefIdxLX and motion vector (X,Y) to do the prediction
2670
More MB?
No
Yes
2680
End
switch between synthetic and non-synthetic
2671
MB index

FIG. 27

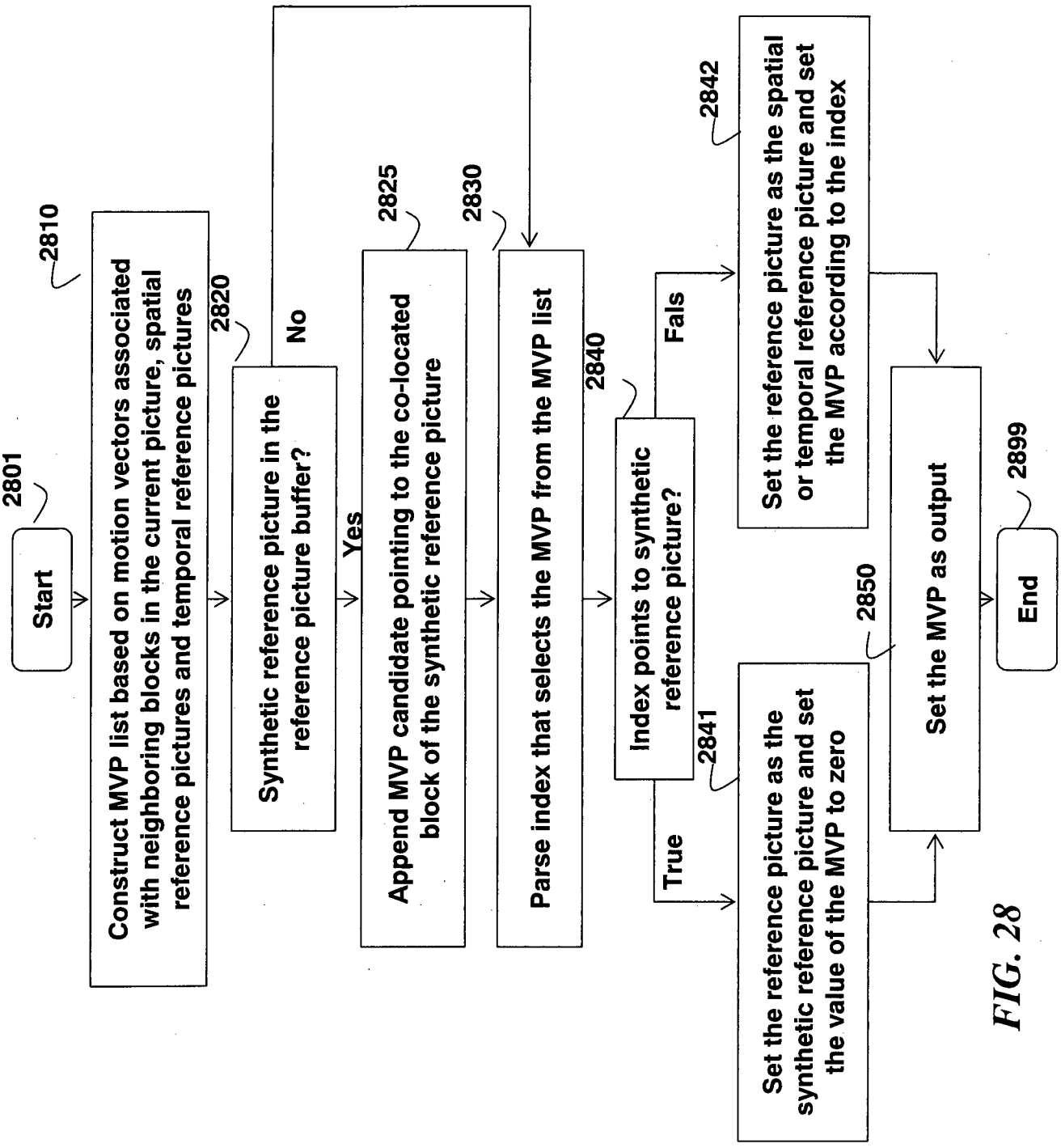

Start
2801
Construct MVP list based on motion vectors associated with neighboring blocks in the current picture, spatial reference pictures and temporal reference pictures
2810
Synthetic reference picture in the reference picture buffer?
2820
No
Yes
Append MVP candidate pointing to the co-located block of the synthetic reference picture
2825
Parse index that selects the MVP from the MVP list
2830
Index points to synthetic reference picture?
2840
True
Fals
Set the reference picture as the synthetic reference picture and set the value of the MVP to zero
2841
Set the reference picture as the spatial or temporal reference picture and set the MVP according to the index
2842
Set the MVP as output
2850
End
2899

FIG. 28

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2013/068578

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04N13/00 H04N7/32 H04N7/36
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04N

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/062756 A1 (TIAN DONG [US] ET AL) 15 March 2012 (2012-03-15) paragraph [0075] - paragraph [0083] paragraph [0199] - paragraph [0233] paragraph [0017] paragraph [0141] paragraph [0066] - paragraph [0068] -----	1-10
A	US 2007/109409 A1 (YEA SEHOON [US] ET AL) 17 May 2007 (2007-05-17) abstract -----	1-10
A	US 2006/146138 A1 (XIN JUN [US] ET AL) 6 July 2006 (2006-07-06) abstract -----	1-10

☐ 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
25 September 2013	07/10/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Cooke, Edward

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/JP2013/068578

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012062756 A1	15-03-2012	US 2012062756 A1	15-03-2012
		WO 2013073282 A1	23-05-2013
US 2007109409 A1	17-05-2007	EP 1978750 A2	08-10-2008
		JP 5116394 B2	09-01-2013
		JP 2008172749 A	24-07-2008
		US 2007109409 A1	17-05-2007
US 2006146138 A1	06-07-2006	EP 1793611 A2	06-06-2007
		JP 5274766 B2	28-08-2013
		JP 2007159112 A	21-06-2007
		JP 2012114942 A	14-06-2012
		US 2006146138 A1	06-07-2006