Noname manuscript No.
(will be inserted by the editor)

Random access prediction structures for light field video coding with MV-HEVC

Vasileios Avramelos · Johan De Praeter · Glenn Van Wallendael · Peter Lambert

Received: date / Accepted: date

Abstract Computational imaging and light field technology promise to deliver the required six-degrees-of-freedom for natural scenes in virtual reality. Already existing extensions of standardized video coding formats, such as multi-view coding and multi-view plus depth, are the most conventional light field video coding solutions at the moment. The latest multi-view coding format, which is a direct extension of the high efficiency video coding (HEVC) standard, is called multi-view HEVC (or MV-HEVC). MV-HEVC treats each light field view as a separate video sequence, and uses syntax elements similar to standard HEVC for exploiting redundancies between neighboring views. To achieve this, inter-view and temporal prediction schemes are deployed with the aim to find the most optimal trade-off between coding performance and reconstruction quality. The number of possible prediction structures is unlimited and many of them are proposed in the literature. Although some of them are efficient in terms of compression ratio, they complicate random access due to the dependencies on previously decoded pixels or frames. Random access is an important feature in video delivery, and a crucial requirement in multi-view video coding. In this work, we propose and compare different prediction structures for coding light field video using MV-HEVC with a focus on both compression efficiency and random accessibility. Experiments on three different short-baseline light field video sequences show the trade-off between bit-rate and distortion, as well as the average number of decoded views/frames, necessary for displaying any random frame at any time instance. The findings of this work indicate the most appropriate prediction structure depending on the available bandwidth and the required degree of random access.

Keywords Light field video coding · multi-view video coding · MV-HEVC · prediction structures · random access · virtual reality · free navigation

dept. of Electronics and Information Systems | Ghent University - imec, IDLab
Technologiepark-Zwijnaarde 122, 9052, Ghent, Belgium
Tel.: +3293314957 | E-mail: vasileios.avramelos@ugent.be

1 Introduction

Virtual reality (VR) is one of the biggest breakthroughs of our times in the fields of entertainment, edutainment and remote control applications. Camera-captured content in VR (e.g., panorama video) is lagging far behind the computer-generated VR experiences (e.g., computer games). For a fully immersive VR experience, a sense of six degrees-of-freedom (6DoF) around the perpendicular axes of the viewer is needed. 6DoF consists of three rotational movements, such as head rotation and tilts, combined with three translational movements such as walking around and small sideway head movements (see fig. 1). Panorama video allows only rotational head movements but disregards any translational move in that same 3D coordinate space (no parallax effect).

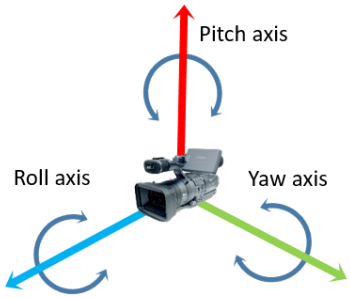

Fig. 1 A sense of 6 degrees-of-freedom (6DoF) around the perpendicular axes of the viewer is required for a fully immersive experience (3 rotational + 3 translational movements).

Light field image modalities are currently a promising solution to help reaching the sense of 6DoF in VR experiences based on natural content. Light fields are the result of advanced capturing of all the light traveling in all directions in a given volume of space. They can create a realistic sense of presence by producing motion parallax, from which we can derive absolute depth information for nearby objects. However, the increase of dimensionality from 2D images and 3D video to 4D light fields and 5D light field video (as in fig. 2) requires the right coding solutions for being able to efficiently transmit camera-captured VR content. This is due to the vast amounts of the resulted data. For instance, less than ten seconds of low resolution 8×8 view light field video consists of several gigabytes (GBs) of data.

Multi-view video coding (MVC) standards are the most conventional coding solutions at the moment, since they can compress video sequences simultaneously captured from multiple camera angles. The high efficiency video coding (HEVC) extension, namely multi-view HEVC (MV-HEVC) is the latest MVC standard and it treats each sub-aperture view as a different HEVC video sequence [1–3]. In practice, in addition to the standard motion-compensated algorithm for temporal prediction, MV-HEVC utilizes inter-view dependencies as well. In other words, MV-HEVC makes use of a reference picture from

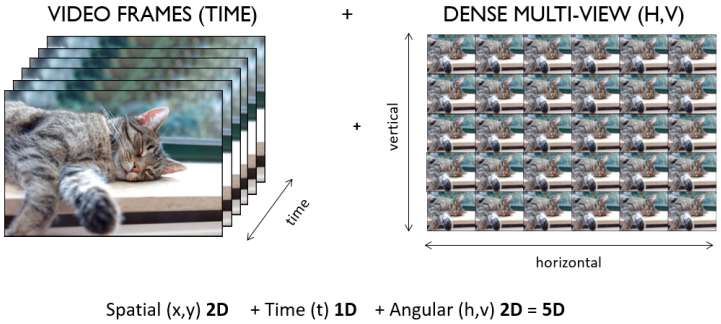

Fig. 2 Representation of the dimensionality increase from a 2D image and 3D video to 4D light fields and 5D light field video.

another position in time, but also a reference picture from a different view in space (typically, with same picture order count). Those inter-view and temporal prediction strategies allow MV-HEVC to compress light field video at acceptable rates.

A similar format, namely multi-view plus depth (MVD), is using the 3D-HEVC extension for coding light field video by treating only a given view (typically the center view) or a set of views as video sequences, and the other sub-aperture views as depth (disparity) maps [3]. Consequently, MVD-based approaches provide additional coding tools such as advanced inter-view prediction tools (for motion and textures), specific depth map coding tools, and view-synthesis coding tools.

The Moving Picture Experts Group (MPEG) has started efforts to standardize a 6DoF video format namely MPEG-I by the year 2021. They aim at a process with two phases: (1) identify the most important 2D views, and (2) rely on view synthesis methods to render other 2D views at the decoder side [4]. Their claim at the moment is that the desired motion parallax feature can be achieved by using color videos, depth information, and associated meta-data, similarly to 3D-HEVC [5]. However, the distortion of the synthesized views is extremely sensitive to the distortion introduced in depth maps [6]. Most depth estimation algorithms have either poor depth estimation performance for light field images, or they add on computational complexity [7]. Therefore, the view synthesis methods which are expected to be used may require considerable computational complexity at the decoder side.

While the coding solutions using MV-HEVC and the ones using 3D-HEVC perform similarly, MV-HEVC solutions do not require disparity maps or any other additional prediction tools in contrast to techniques built upon 3D-HEVC. They rather use syntax elements similar to standard HEVC for enabling inter-view prediction, which facilitates the whole process to a great extent. Due to this facts, we argue that MV-HEVC offers a more convenient and straight-forward coding solution (in terms of implementation) for data-intensive light field video.

In video applications, random access is one important feature which allows the viewer to navigate through a video. For a video to be fully random accessed in time for example, all video frames must be able to be independently decoded, and consequently independently accessed. Besides random access in time, VR and free viewpoint television (FTV) applications also have view random access as a crucial requirement for delivering high quality content at any possible instance both in time and in space (free navigation) [8,9]. For this work, we define view random access as the ability of a decoder to switch to a different view immediately at any point in time. Predictive coding structures complicate random access because pixels within a frame directly depend on previously decoded frames. This is an issue for light field video and the desired free navigation feature, since the user movement changes the access patterns in complex ways. Therefore, it is plausible to seek for a prediction scheme which offers low-cost random access in time and space for light field video coding. There is a large level of freedom in choosing an optimal prediction structure. In previously published work, we have stressed the importance of choosing the appropriate scanning topology and the correct view indexing in the spatial domain [10]. In this work we go one step further by combining those inter-view prediction structures with temporal prediction for an optimal MV-HEVC light field video coding scheme which exploits every possible redundancy between different frames. The main goal of this work is to combine inter-view and temporal prediction structures for MV-HEVC, aiming towards an optimal rate-distortion coding scheme with regard to the additional requirement of cost-effective random accessibility.

We tested and compared a number of different light field video coding scenarios for MV-HEVC, which differ in terms of random access capabilities in space and in time. For the evaluation of the results, we focused on compression efficiency and view random access. The main contribution of this work is twofold. First, a coding structure (namely *CenterView* - see sec. 3.3), which offers sufficient random access freedom for free navigation applications while remaining efficient in terms of compression gain, is evaluated and tested. Second, we assess a coding structure (namely *Full* - see sec. 3.4), which efficiently combines inter-view and temporal prediction for offering drastically better compression performance than conventional methods at the cost of minimum view random access. For measuring random access, a simulation of a simple light field video streaming scenario for a VR application is presented.

In the next section (sec. 2) we briefly present related work on MV-HEVC light field coding and their different approaches for exploiting redundancies between views in the 2D space. In sec. 3 we present in detail the different coding structures proposed in this work. Then, in sec. 4 we describe the experimental setup and we provide necessary information for recreating the demonstrated results. Finally, sec. 5 discusses those results along with useful conclusions we were able to draw with respect to light field video compression and its applications.

2 Related work

Several studies in the literature propose methods of restructuring light field images into 2D video frames and using standardized video codecs and their extensions for compression purposes. In that way, they are able to exploit spatial redundancies between views in space by treating them as consecutive frames in time. For instance, in [11], the authors used the sub-aperture views of a light field image as video frames and code them using HEVC. In [12], pseudo multi-view sequences were created from light field images and coded with standard MV-HEVC using a 2D bi-directional prediction scheme. Liu et al. propose an HEVC scalable coding approach (HEVC-SC) by using a sparse interlaced view image set and disparity maps [13]. A hybrid linear weighted prediction scheme for coding light field images using the screen content coding extension of HEVC (HEVC-SCC) is proposed in [14].

While numerous efforts have been made to efficiently compress light field image data similar to conventional video, light field video coding has been an emerging field of research as well. Both MVD and MVC techniques are typically deployed and deliver acceptable results. While MVC offers a more straightforward implementation, it focuses on the compression of a given number of camera views and it is not able to facilitate the generation of additional views. Therefore, the new immersive video coding standard (MPEG-I) is expected to be MVD-based. With the right additional coding tools MVD can slightly outperform current MVC methods. However, in the absence of readily available depth maps for each camera view and/or the absence of depth coding tools and view synthesizers, MVC methods are convenient solutions for light field video coding. Dricot et al. propose an efficient full parallax prediction order for the optimization and the exploitation of inter-view dependencies [15]. Their method, namely *Central2D* outperforms conventional MVC prediction structures used in the past (see fig. 3) for the purpose of multi-view video coding [16]. Wang et al. propose an MVC-based approach (see fig. 4) by enabling a two-directional (horizontal and vertical) inter-view prediction scheme [17] to eventually outperform the *Central2D* method [15]. On the other hand, Conti et al. experiment with MVD light field video coding and they recommend a geometry-based disparity compensation scheme aiming to only code the depth map of the base view [18].

Due to the high dimensionality of light fields and the resulting amount of data which is prohibitive for real-time immersive applications, alternative approaches which depart significantly from traditional coding techniques have been proposed as well. For example, steered mixture-of-experts (SMoE) is a promising framework for delivering the required 6DoF for camera-captured content [19, 20]. SMoE focuses on the generic representation of multidimensional image modalities while operating in the spatial domain and offering pixel-parallel decoding capabilities [21].

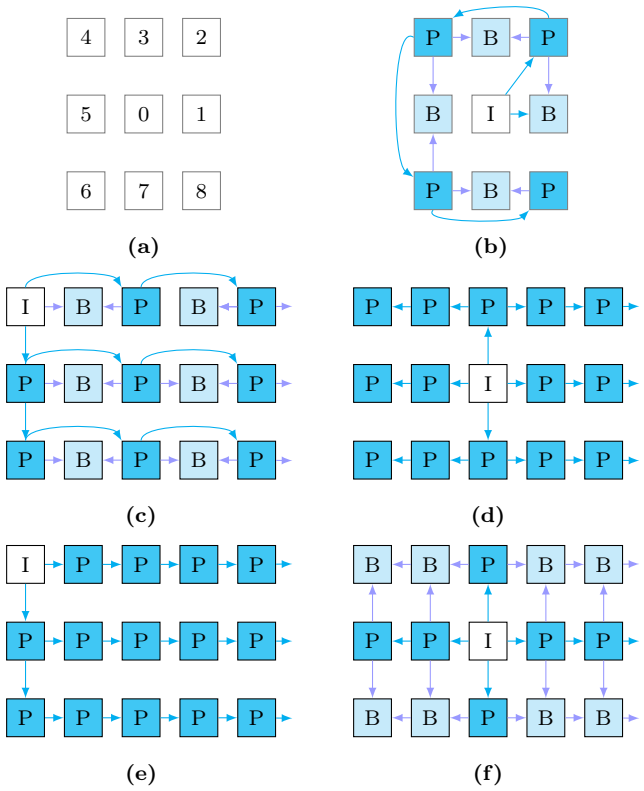

Fig. 3 Typical coding structures found in the state-of-the-art literature for inter-view prediction (spatial ordering and inter-dependencies). On top, a spiral scanning topology (a) for a 3×3 light field for which an IBPBP prediction structure (b) is applied. Beneath, customized coding structures (c),(d) and (e) for a 5×3 camera array proposed in [16], and the *Central2D* method (f) proposed in [15].

3 Proposed Coding Structures for Light Field Video Coding

In this section, the tested light field video coding structures for MV-HEVC are presented in detail. Four different coding structures have been implemented, aiming to find an optimal solution in terms of rate-distortion performance and random access capabilities. As such, the presented implementations are covering a set of MVC solutions ranging from true random access in time and space (free navigation) to most optimal coding performance (lowest bit-rate for the highest video quality). The four configurations consist of:

- an all intra-coded anchor scheme (*AllIntra*) for allowing full independence between frames/views (true random access),
- an anchor scheme where inter prediction between frames in time (*Inter-Frame*) is allowed and every view is an independent HEVC sequence,

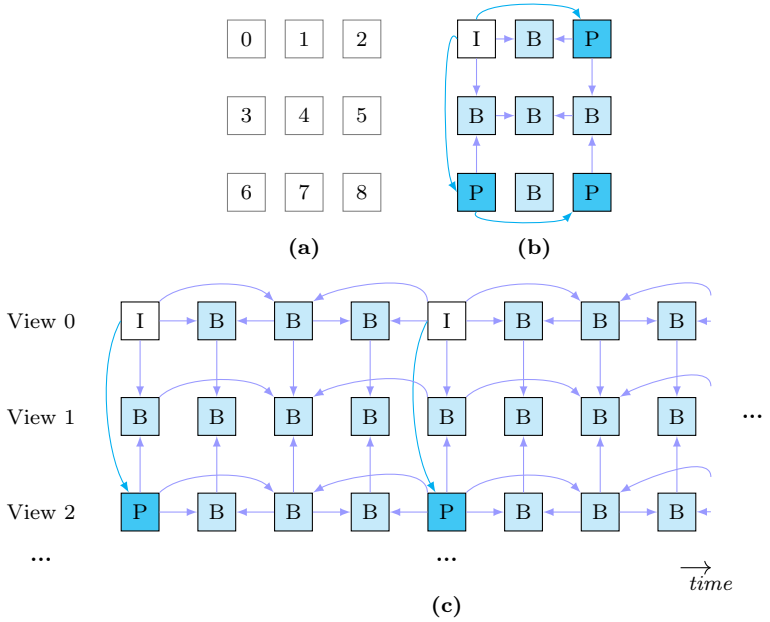

Fig. 4 Coding structure proposed in [17]. On top, the indexing (a) for a 3×3 light field (raster scan), for which the proposed prediction structure (b) is applied. Beneath, the temporal prediction structure is visualized as well, for $n = 3$ camera views. While this approach outperforms previously proposed ordering structures, it is too complicated for random access applications due to its multiple dependencies.

- a proposed scheme where a central view (*CenterView*) is being used as reference and all other views are P-predicted from that center view,
- and a proposed scheme where inter-dependencies between frames in time and neighboring views in space are fully exploited (*Full*).

In terms of compression efficiency, a scheme which exploits all sorts of redundancies would be considered the best configuration. A scheme which deploys the maximum and most optimal temporal and inter-view dependencies is the scheme which will require the least bit-rate for providing the best video quality. However, when considering free navigation, such a scheme will not allow view random access at any possible time instance due to all the necessary frames which are required for reconstructing a randomly requested picture. To reduce the complexity at the decoder side and make a choice of random access over compression rate, frame dependencies need to be minimized. Optimally, frame independence is achieved by using only HEVC intra-prediction for coding all frames. However, the bandwidth requirements for such a solution are quite demanding. Therefore, depending on the application requirements, the scheme which provides the right trade-off between random access freedom and compression efficiency can be adequately chosen.

3.1 *AllIntra*

To cover the whole range from true random access to best compression efficiency, we start from an all intra configuration, namely *AllIntra*. Every view at every time instance is independently coded such that true random access in time and in space is provided. Consequently, all frames are coded using only intra prediction (I-frames). No temporal predictions are being performed between frames in time nor between views in space (fig. 5). In this case, the presentation frame rate is the same as the decoding frame rate. While an *AllIntra* case enables full random access and completely avoids temporal artifacts, it is however computationally expensive, something which makes it less practical.

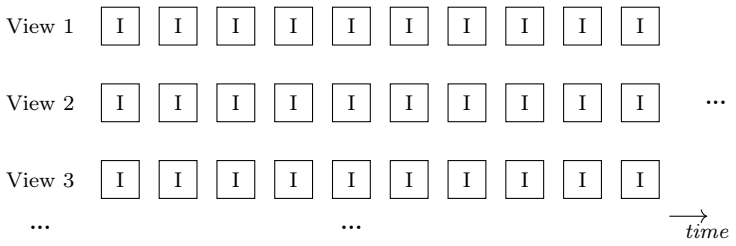

Fig. 5 Fully random accessed case used in the *AllIntra* anchor scheme. Each view/frame is coded independently without a single inter-dependency.

3.2 *InterFrame*

In this scenario, multi-view compression is not fully exploited and every view is encoded as a separate video sequence. Every first frame of each sequence is always an I-frame and the remaining frames are coded using bi-directional predictive frames (B-frames) or uni-directional predictive frames (P-frames). If the group-of-pictures (GOP) size is equal to the intra-period (frequency of the appearance of an I-frame in time), then all the remaining frames are treated as B-frames. This is similar to the typical hierarchical-B coding structure used in HEVC [22]. Since every view is a separate sequence, there are no dependencies between different views. Therefore, while watching a specific view and a switch to another view is necessary, that other view would need to be decoded. This scheme is similar to conventional adaptive streaming techniques [23,24]. However, the absence of inter-view dependencies leaves a large room for improvement in terms of redundancy exploitation, and therefore compression efficiency. Fig. 6 depicts an example of the *InterFrame* scenario for light field video coding used in this work. Note here, that the intra-period is larger than the GOP size (in this case GOP= 8), and therefore the first frames of a GOP

within an intra-period are coded as a P-frame. One could argue that instead of using a new I-Frame for each intra-period of each independent light field view, the corresponding key-frame of a previously decoded neighboring view could be used. In other words, a more bandwidth-efficient simulcast-based scheme at the cost of full view-independence. For this work's experiments, we name this enhanced *InterFrame* scheme *Interframe+* (see sec.4).

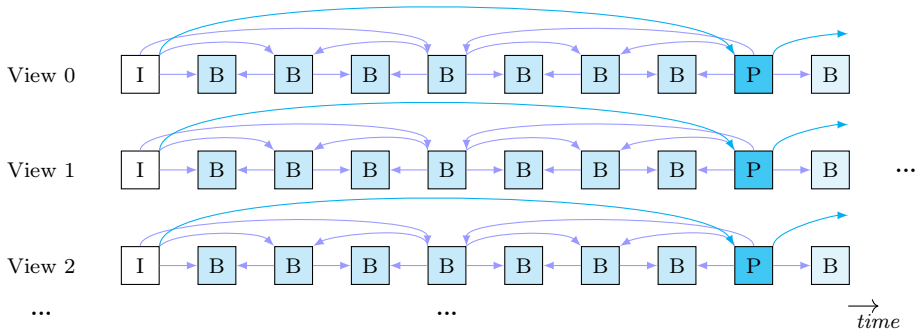

Fig. 6 Temporal HEVC prediction structure (hierarchical-B) used in the *InterFrame* scheme. Each view is considered as a separate HEVC video sequence with no inter-view dependencies as in typical MVC. In the *InterFrame+* case, only one view is using I-frames and all other views are using P-frames instead.

3.3 *CenterView*

This is the proposed solution where only one view, typically a center view, is encoded as a regular video. The other views are predictively coded (P-pictures) from that center view as depicted in fig 7. While the center view uses hierarchical-B coding structure for inter-prediction, the frames of the remaining views are P-predicted from the corresponding frame within the center view as illustrated in fig. 8. Randomly watching any position in the light field video would require the decoder to be able to decode at twice the frame rate i.e., the center view and the chosen view. In this way, MVC capabilities are exploited due to the inter-view prediction, while temporal dependencies are completely discarded from all the surrounding views for the sake of view random access. Therefore, random access can be partially claimed since only two views need to be decoded for offering a frame rate equal to the actual decoding frame rate. However, a minor disadvantage of this scheme is that in future light field applications where light fields consist of several hundreds or thousands of views with a large base-line between them, it will then be harder (more expensive) for a corner view to be encoded using a center view as a reference. This is due to the large error between the two views because of the significant distance between the corresponding cameras.

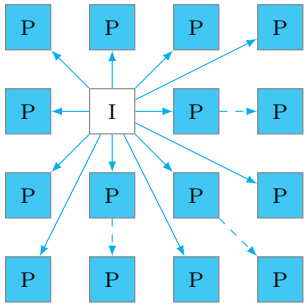

Fig. 7 Coding structure for the *CenterView* scheme for an example of a 4×4 view light field video. All views are P-predicted from one single center view (dashed arrows also begin from the center tile).

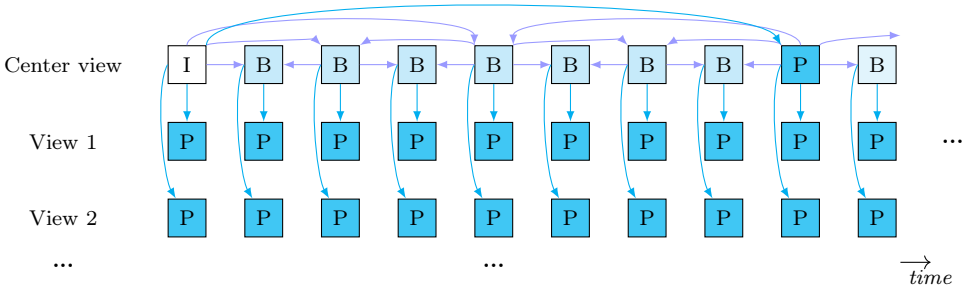

Fig. 8 Proposed prediction structure used in the *CenterView* scheme for inter-view prediction. Every frame from every view is predicted by its corresponding frame originated from a center view. The center view uses hierarchical-B coding structure for temporal prediction.

3.4 *Full*

This structure is the most compression efficient setting presented in this paper. Within MV-HEVC, an IPBBBB coding structure is deployed (see fig. 9-b) for exploiting dependencies between the closest neighboring views. To result in this coding structure, the views need to be indexed in 2D space. There are unlimited scanning topologies which can be used for indexing the different sub-aperture views. For reaching optimal coding efficiency, we used a modified spiral scanning technique as shown in fig. 9-a for a light field with 4×4 views, similar to our previous work [10, 20]. Starting from the corners of the grid, the 2D space is scanned by revolving in a clock-wise manner until the center view points are reached. For an optimal encoding structure, the maximum number of possible bi-directional dependencies must be used for a significant reduction in bit-rate. Therefore, vertical correlations (and not only horizontal) between views need to be considered as well [17]. Consequently, for predicting an actual view, the two closest horizontal, vertical or diagonal views are used similarly to the illustration in fig. 9-b. A slight preference for the closest horizontal view

is given, since the human visual system is more sensitive to horizontal correlations [25]. An example of the basic MVC bi-directional prediction principle is illustrated in fig. 10. The views are reordered by following the proposed indexing topology, and the dependencies are organized accordingly. While this is an efficient solution for coding light field video at minimum bit-rate ranges, the temporal inter-dependencies and the inter-dependencies between views complicate random access to a great extent. More specifically, to have random access capabilities and be able to see any random view at any point in time when using this scheme, all views need to be decoded at once. In other words, the decoding hardware needs to decode at n times (n is the number of views) the frame rate to provide the full light field experience.

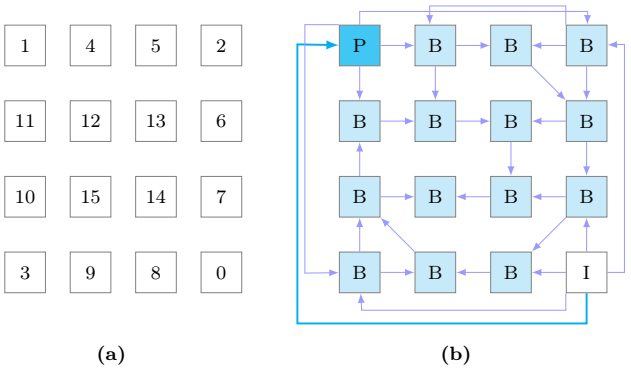

Fig. 9 Grid view of the 8×8 light field view ordering structure (a) and inter-view prediction structure for a 4×4 light field view case (b) used in the proposed *Full* prediction scheme. Each view is predicted by using two other (horizontal, vertical or diagonal) neighboring views.

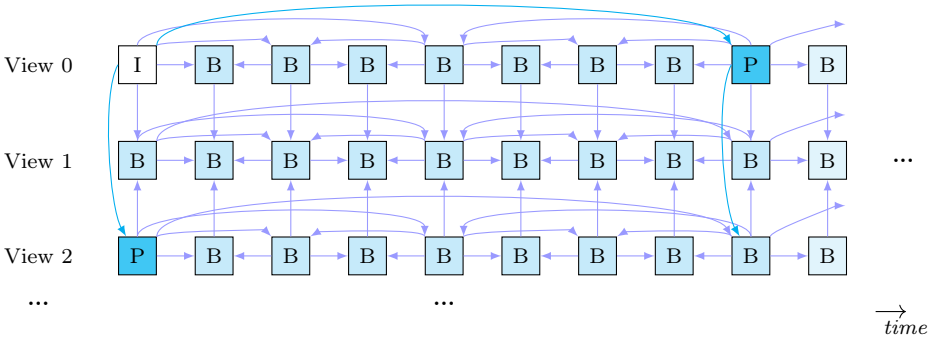

Fig. 10 Typical MVC coding structure for inter-view and temporal prediction. The same principle has been used in the *Full* coding scheme with a modified view ordering structure as shown in fig. 9.

4 Experiments

In this section, we compare the performance of the coding structure methodologies presented in this work with regard to compression gain and random access. The dataset used for this comparison is the light field videos provided by researchers from the University of California [26]. It contains three light field sequences namely *cats*, *train1* and *train2*. In fig. 11, a thumbnail picture of each dataset is shown. The dataset consists of three camera-captured light field video sequences in total, two with a resolution of 512×352 pixels and one of 544×320 pixels. All test sequences were captured at 30 frames per second for approximately 100 frames and $8 \times 8 = 64$ views. In terms of content, the scene is relatively static with the motion of the moving object(s) gradually increasing from *cats* to *train1* to *train2*. The reference MV-HEVC software *HTM-16.5* [27] was used for encoding and decoding the three light field video sequences in four different coding structure scenarios (*AllIntra, InterFrame, CenterView, Full* - see sec. 3). The GOP size and intra period used in all cases with temporal dependencies were 8 and 24 respectively (random access at roughly each second) as recommended in [28]. Fig 12 shows the view indexing for the proposed *Full* prediction scheme.

Fig. 11 The three light field video sequences used in this work. From left to right, "cats", "train1" and "train2" [26].

1	4	5	20	21	6	7	2
19	28	29	36	37	30	31	8
18	47	48	49	50	51	38	9
27	46	59	60	61	52	39	22
26	45	58	63	62	53	40	23
17	44	57	56	55	54	41	10
16	35	34	43	42	33	32	11
3	15	14	25	24	13	12	0

Fig. 12 Grid view of the 8×8 light field view ordering. This is an adaptation from the 4×4 example presented in fig. 9, and the inter-view dependencies can be drawn accordingly.

For assessing the light field video quality we used conventional video quality assessment (VQA) techniques [29]. For accelerating the process and not calculating quality metrics for each individual view of our 8×8 view dataset, a shortcut method has been utilized, similar to previous works [19,20]. Five views taken from different areas of the 2D space were selected and a single 2D video sequence is created by iterating through those selected views as visualized in fig. 13. It is still not clear which quality metric is the best for measuring light field video quality. Therefore, we calculated two conventional VQA metrics, peak signal to noise ratio (PSNR) and structural similarity index (SSIM). Since both metrics correlated well for the tested content we only present PSNR as the distortion unit. Typical values for the PSNR in lossy image and video compression are between 30 and 50 dB (bit depth of 8 bits), where higher is better.

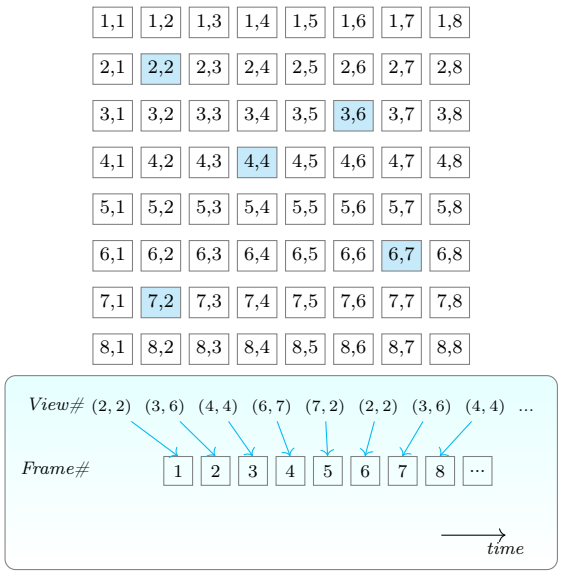

Fig. 13 Frame structure of video sequence generated out of different light field views for quicker quality assessment. Five views generated a single video sequence which used for evaluation purposes respecting the standardized video quality assessment methods.

As seen in fig. 14, in terms of rate-distortion, the proposed *CenterView* solution performs better than the *InterFrame* (no inter-view dependencies), yet worse than the *Full* (fully-referenced) scenario. More specifically, for achieving 37 dB in the PSNR measurement, network capabilities of approx. 1, 3, and 9 Mbps would need to be deployed respectively for *Full*, *CenterView* and *InterFrame* (see fig. 15). However, the main advantage of *CenterView* is its freedom in terms of random access, since it requires the decoding of at most two views at any time to select any random view. Additionally, by using *CenterView*

instead of *InterFrame*, significant bit-rate savings can be achieved. Similarly, by using the *Full* instead of the *CenterView* solution, a remarkable bit-rate reduction can be achieved as well. Nevertheless, using the *Full* solution comes at an important cost of losing any freedom in terms of random access (table 1). For the best-performing *Full* case, a comparison with the work proposed by Wang et al. [17] is given in fig. 16. Lastly, fig. 17, shows where the *CenterView* stands between the Simulcast *InterFrame* case and an enhanced version of *InterFrame* (*InterFrame+*) where full view independence is sacrificed for bandwidth efficiency (see sec.3.2).

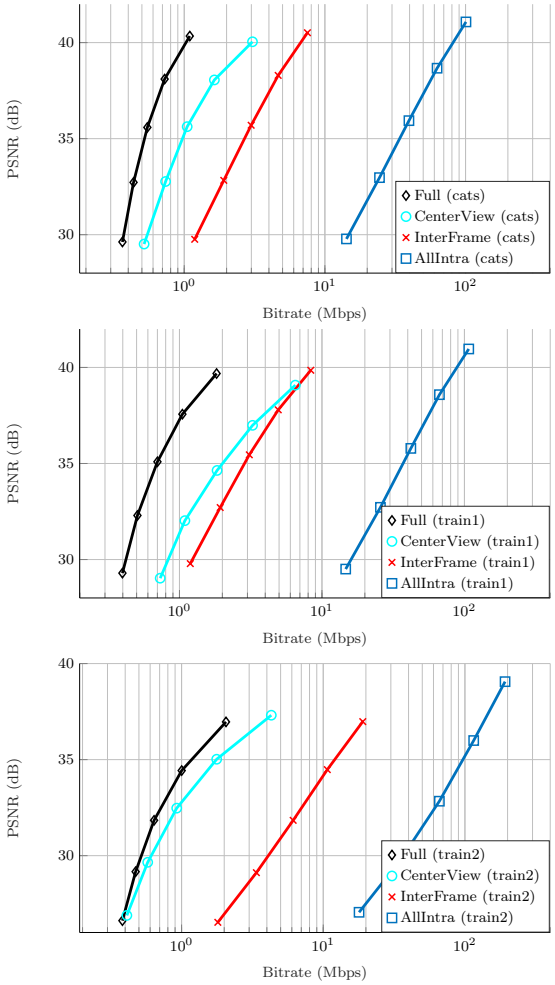

Fig. 14 Rate-distortion comparison between the four different random access scenarios tested in this work, for each one of the three datasets. Logarithmic scale is used for better visualization.

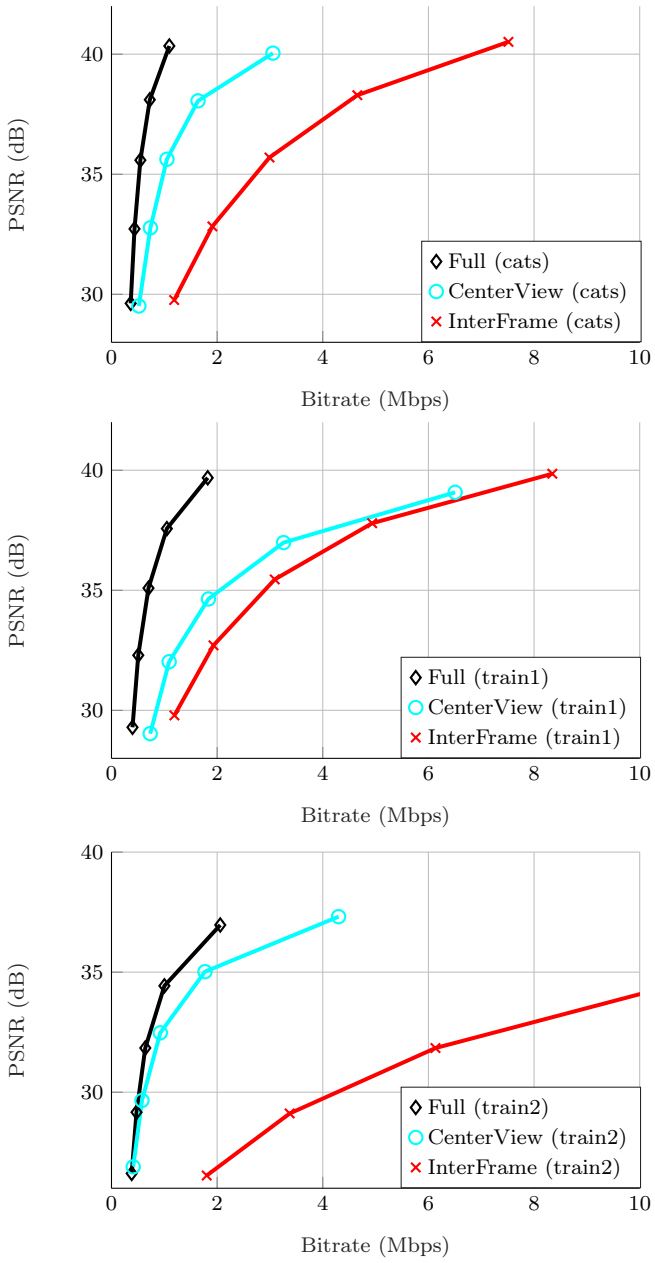

Fig. 15 Comparison between three different tested prediction structures which can be used in realistic scenarios (in terms of bandwidth), for each one of the three datasets. Depending on the content, the proposed *CenterView* structure can closely compete prediction structures where full inter-view dependencies are enabled (*Full*).

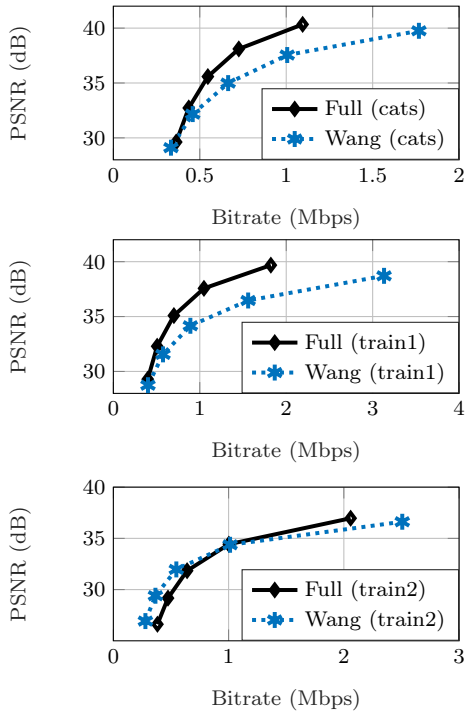

Fig. 16 Comparison between the best-performing in terms of compression efficiency proposed scheme in this work with a proposed scheme from the literature (see fig. 4) [17].

We estimated the average number of necessary decoded pictures required to be able to access the video at a random time instance and then switch to another random view in space (table 2). Fig. 18 shows an example of the derived estimation of the results shown in table 2. *AllIntra* offers full frame independence, although it is inefficient in terms of required bit-rate. The most efficient in terms of bit-rate *Full* scheme needs all the frames to be decoded, due to its multiple inter-dependencies. *Interframe* (temporal prediction only) needs 10 frames on average for switching views at any time instance, and the *CenterView* (proposed random access scheme) requires an average of approx. 5.88, 5.93 and 5.99 decoded frames (less than 25% of total frames) while offering acceptable rates. The most compression efficient methods proposed in the literature either require to decode all frames [17], or need an average of approx. 38% more frames [15], for the number of tested views (3×3, 4×4, and 8×8).

Table 3 shows the Bjøntegaard-Delta rate (BD-rate) [30], i.e., the bit-rate overhead for the same quality, between the different tested scenarios. Negative values represent improvement of the *CenterView* over the scheme in comparison. For instance, we achieved 61.3% of bit-rate savings when we switched from *InterFrame* to *CenterView* (for the *cats* sequence). On the other hand, a

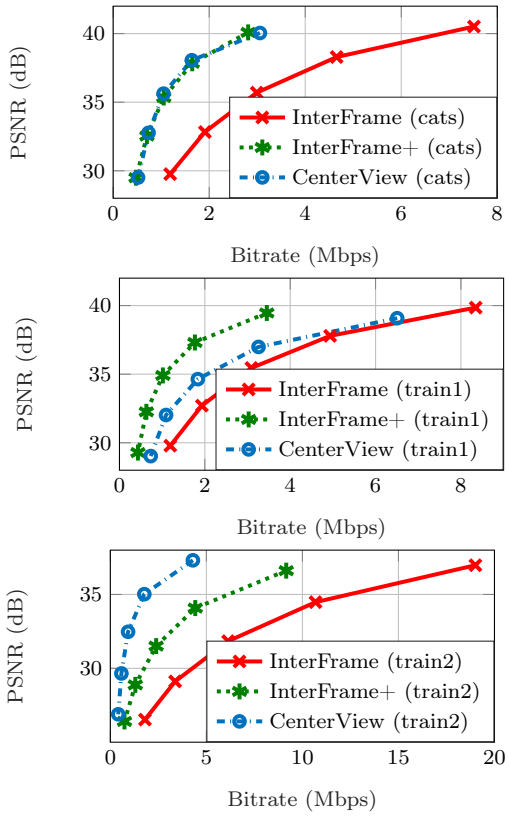

Fig. 17 Comparison of the proposed *CenterView* scheme with the *InterFrame* simulcast scheme and an enanced version of *InterFrame* (*InterFrame+*) for increased compression performance at the cost of view independence.

90.7% bit-rate overhead was observed when a switch from *Full* to *CenterView* was made for the same dataset.

Table 1 Summary of the results for the four MV-HEVC coding structure scenarios in terms of bit-rate (Mbps) and view random access efficiency.

Prediction structure	Bitrate (Mbps)	View random access efficiency
AllIntra	$\times 10^{1} - \times 10^{2}$	Yes
InterFrame	$\times 10^{0} - \times 10^{1}$	No
CenterView	$\times 10^{-1} - \times 10^{0}$	Yes
Full	$\times 10^{-1} - \times 10^{0}$	No

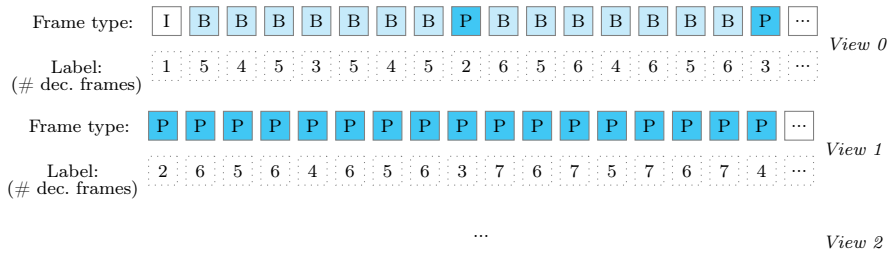

Fig. 18 Example of frames labeling based on the number of decoded pictures needed for visualizing each frame for the proposed *CenterView* structure. Using the above labeling, we estimate the average number of decoded frames necessary for accessing a random frame in time of a random view in space, and instantaneously switch to another view. Dependencies between frames can be seen in fig. 8.

Table 2 Average number of decoded pictures needed for switching from a frame at a random time instance to another view in space. The results apply for a sample of 24 frames.

	Number of decoded frames		
Setting	3×3 *views*	4×4 *views*	64×64 *views*
AllIntra	2	2	2
InterFrame	10	10	10
CenterView	5.88	5.93	5.99
Full	24	24	24
Central2D [15]	10.25	13.76	21.66
Wang et al. [17]	24	24	24

Table 3 Bjontegaard-Delta Rate (BD-Rate) percentages in terms of bit-rate overhead of the proposed *CenterView* setting when compared to the tested scenarios.

Sequence	*CenterView* vs. *Full*	*CenterView* vs. *InterFrame*	*CenterView* vs. *AllIntra*
cats	90.7	-61.3	-96.9
train1	171.7	-28.6	-94.4
train2	31.1	-84.7	-98.1

5 Conclusions

In this work the goal is to seek for a prediction structure scheme within multi-view coding which is not efficient only in terms of compression efficiency, but in terms of random accessibility as well. The target application is light field video compression aiming at VR applications. Consequently, we propose a prediction structure (*CenterView*) which offers an acceptable trade-off between bandwidth requirements and ease of use in free navigation applications. We additionally propose a prediction scheme (*Full*) which aims at the minimization of bit-rate requirements by efficiently exploiting spatial redundancies between different views. For comparison reasons we implemented two anchor schemes and a widely-used prediction scheme from the literature of light field video coding. The anchors, namely *AllIntra* and *InterFrame*, use only intra-prediction and only temporal prediction, respectively.

The best configuration in terms of random access, the *AllIntra* scheme, encodes all views and all frames independently using only I-frames. However, the large bitrates needed for encoding makes it inefficient. The proposed solution (*CenterView*) has the best trade-off between random access and compression efficiency, providing remarkable improvement when compared to simulcast (*InterFrame*) and enhanced simulcast solutions (*InterFrame+*). On the other hand, when compared to the (*Full*) solution, which is optimal in terms of compression gain but without random access capabilities, the *CenterView* solution needs at most two times more bit-rate for the same quality to offer full view random access only by decoding a maximum of two views at a time. That makes the *CenterView* a realistic solution for coding light field video data while retaining random access features necessary for free navigation in VR.

Finally, the performance is content-dependent. In cases where less motion is present in a video, one can assume that the *CenterView* will offer less gain (due to the preference of inter-view over inter-frame dependencies) than the one presented in the current results since less motion typically causes better inter-frame (temporal) predictions and therefore more compression gain. However, since temporal prediction is also crucial for coding the reference central view, the proposed *CenterView* structure will benefit from it as well in terms of bit-rate savings. The same reasoning holds for inter-view or inter-tile differences. When objects are further from the camera, or the array of cameras, then inter-view predictions are expected to improve performance.

The findings of this work indicate that it is possible to enable random access capabilities in light field video coding, a crucial requirement for freely navigating through different views in VR applications. In the past, random access has been discussed in the literature of multi-view video coding. However, no work investigated both compression efficiency and random access at the same time. The proposed *CenterView* scheme can be readily applied in light field video coding solutions in MV-HEVC for significantly reducing decoding delays while maintaining the ability of switching between views. Future work consists of investigating random access in multi-view plus depth techniques, which appear to be the basis of the future immersive video coding standard of MPEG.

Acknowledgements The research activities described in this article were funded by IDLab (Ghent University - imec), Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union. We would also like to share our gratitude to John Carmack (currently CTO of Oculus VR) for sharing his ideas via social media (Twitter), something which triggered the further investigation of specific parts in this work.

References

1. G. J. Sullivan, J. R.Ohm, W. J. Han and T. Wiegand: 'Overview of the High Efficiency Video Coding (HEVC) Standard', *IEEE Transactions on Circuits and Systems for Video Technology (vol. 22)*, 2012.
2. G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro: 'Standardized Extensions of High Efficiency Video Coding', *IEEE Journal on Selected Topics in Signal Processing (vol. 7)*, 2013.
3. G. Tech, Y. Chen, K. Mueller, J.-R. Ohm, A. Vetro, and Y.-K. Wang: 'Overview of the Multiview and 3D Extensions of High Efficiency Video Coding', *IEEE Transactions on Circuits and Systems for Video Technology (vol. 26)*, 2015.
4. M. Domanski, O. Stankiewicz, K. Wegner and T. Grajek: 'Immersive visual media, MPEG-I: 360 video, virtual navigation and beyond', *International Conference on Systems, Signals and Image Processing*, 2017
5. ISO/IEC JTC1/SC29 WG11: 'Text of PDTR ISO/IEC 23090-1 Immersive Media Architecture', Doc. N17685, *MPEG 122nd meeting*, San Diego, USA, 2018
6. P. Gao and W. Xiang: 'Rate-distortion optimized mode switching for error-resilient multi-view video plus depth based 3D video coding', *IEEE Transactions on Multimedia, vol. 16, no. 7, pp. 1794-1808*, 2014.
7. F. Zilly, C. Riechert, M. Mueller, P. Eisert, T. Sikora and P. Kauff: 'Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline', *Journal of Visual Communication and Image Representation, vol. 25, pp. 632-648*, 2014.
8. G. Lafruit, M. Domanski, K, Wegner, T. Grajek, T. Senoh, J. Jung, P. T. Kovács, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans, P. Carballeira, S. Garcia and M. Tanimoto: 'New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV', *Electronic Imaging, Stereoscopic Displays and Applications XXVII*, 2016.
9. Á. Huszák: 'Advanced free viewpoint video streaming techniques', *Multimedia Tools and Applications, vol. 76, issue 1, pp. 373-396*, 2017.
10. V. Avramelos, G. Van Wallendael and P. Lambert: 'Overview of MV-HEVC prediction structures for light field video', *SPIE Proceedings 11137*, 2019.
11. D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu and W. Zeng: 'Pseudo-sequence-based light field image compression', *IEEE International Conference in Multimedia & Expo Workshops (ICMEW), pp. 1-4*, 2016.
12. W. Ahmad, M. Sjöström and R. Olsson: 'Compression scheme for sparsely sampled light field data based on pseudo multi-view sequences', *SPIE proceedings (vol. 10679)*, 2018.
13. D. Liu, P. An, R. Ma, C. Yang, L. Shen and K. Li: 'Scalable coding of 3D holoscopic image by using a sparse interlaced view image set and disparity map', *Multimedia Tools and Applications, vol. 77, issue 1, pp. 1261-1283*, 2018.
14. D. Liu, P. An, R. Ma, L. Shen: 'Hybrid linear weighted prediction and intra block copy based light field image coding', *Multimedia Tools and Applications, vol. 77, issue 24, pp. 31929-31951*, 2018.
15. A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, F. Dufaux: 'Full parallax super multi-view video coding', *IEEE International Conference on Image Processing (ICIP)*, 2014.
16. P. Merkle, A. Smolic, K. Muller, T. Wiegand: 'Efficient Prediction Structures for Multiview Video Coding', *IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, issue 11, pp. 1461-1473*, 2007.
17. G. Wang, W. Xiang, M. Pickering and C. W. Chen: 'Light field multi-view video coding with two-directional parallel inter-view prediction', *IEEE Transactions on Image Processing (vol. 25)*, 2016.
18. C. Conti, P. T. Kovacs, T. Balogh, P. Nunes and L. D. Soares: 'Light-field video coding using geometry-based disparity compensation', *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, 2014.
19. R. Verhack, T. Sikora, L. Lange, R. Jongebloed, G. Van Wallendael and P. Lambert: 'Steered mixture-of-experts for light field coding, depth estimation, and processing', *IEEE International Conference on Multimedia and Expo (ICME), pp. 1183-1188*, 2017.
20. V. Avramelos, I. Saenen, R. Verhack, G. Van Wallendael, P. Lambert and T. Sikora: 'Steered Mixture-of-Experts for Light Field Video Coding', *SPIE proceedings (vol. 10752)*, 2018.

21. V. Avramelos, R. Verhack, I. Saenen, G. Van Wallendael, B. Goossens and P. Lambert: 'Highly parallel steered mixture-of-experts rendering at pixel-level for image and light field data', *Journal of Real-Time Image Processing, pp. 1-17*, 2018.
22. V. Sze, M. Budagavi, G. J. Sullivan: 'Hight Efficiency Video Coding - Algorithms and Architectures', *Integrated Circuits and Systems, ISBN 3319068946, 9783319068947, Springer Publishing Company, Incorporated*, 2014
23. C. Ozcinar, E. Ekmekcioglu, J. Ćalić, A. Kondoz: 'Adaptive delivery of immersive 3D multi-view video over the Internet', *Multimedia Tools and Applications, vol. 75, issue 20, pp. 12431-12461*, 2016
24. Y. Sánchez de la Fuente, R. Skupin, T. Schierl: 'Video processing for panoramic streaming using HEVC and its scalable extensions', *Multimedia Tools and Applications, vol. 76, issue 4, pp. 5631-5659*, 2016
25. B. C. Hansen and E. A. Essock: 'A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes', *Journal of vision (vol. 4 12)* 2004.
26. T. C. Wang, J. Y. Zhu, N. Khademi, A. Efro and R. Ramamoorthi: 'Light field video capture using a learning-based hybrid imaging system', *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017), vol. 36, issue 4*, 2017.
27. Fraunhofer HHI: 'Multiview High Efficiency Video Coding (MV-HEVC) - HTM software repository', *https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/*, last accessed on March 11th, 2019.
28. ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio: 'Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation', 2015
29. Multimedia Signal Processing Group (MMSPG): 'VQMT: Video Quality Measurement Tool', *https://mmspg.epfl.ch/downloads/vqmt/*, last accessed on March 23rd, 2019.
30. G. Bjøntegaard: 'Calculation of Average PSNR Differences Between RD-Curves', *ITU-T SG16 Q.6 document VCEG-M33*, 2001.