Overview of MV-HEVC prediction structures for light field video

Vasileios Avramelos, Glenn Van Wallendael, and Peter Lambert

Ghent University - imec, IDLab, Dept. of Electronics and Information Systems,
Technologiepark-Zwijnaarde 122, 9050 Ghent, Belgium

ABSTRACT

Light field video is a promising technology for delivering the required six-degrees-of-freedom for natural content in virtual reality. Already existing multi-view coding (MVC) and multi-view plus depth (MVD) formats, such as MV-HEVC and 3D-HEVC, are the most conventional light field video coding solutions since they can compress video sequences captured simultaneously from multiple camera angles. 3D-HEVC treats a single view as a video sequence and the other sub-aperture views as gray-scale disparity (depth) maps. On the other hand, MV-HEVC treats each view as a separate video sequence, which allows the use of motion compensated algorithms similar to HEVC. While MV-HEVC and 3D-HEVC provide similar results, MV-HEVC does not require any disparity maps to be readily available, and it has a more straightforward implementation since it only uses syntax elements rather than additional prediction tools for inter-view prediction. However, there are many degrees of freedom in choosing an appropriate structure and it is currently still unknown which one is optimal for a given set of application requirements. In this work, various prediction structures for MV-HEVC are implemented and tested. The findings reveal the trade-off between compression gains, distortion and random access capabilities in MV-HEVC light field video coding. The results give an overview of the most optimal solutions developed in the context of this work, and prediction structure algorithms proposed in state-of-the-art literature. This overview provides a useful benchmark for future development of light field video coding solutions.

Keywords: Light field video coding, multi-view video coding, MV-HEVC, prediction structures

1. INTRODUCTION

Virtual reality (VR) is undoubtedly emerging in the field of entertainment and edutainment systems. Currently, computer generated scenes (e.g., computer games) are significantly outpacing camera capture content (e.g., 360° video). In 360° video, also known as panorama video, only rotational head movements around the three perpendicular axes are allowed, while any translational movement in the same 3D space is disregarded. To obtain a more realistic sense of free navigation, it is necessary to account all six-degrees-of-freedom (6DoF), i.e, three translational combined with three rotational movements.[1,2]

Light field technology is the technology which rose to the occasion by showing great potential to provide the prerequisites to advance towards the realization of 6DoF in VR. A light field is defined as a higher-dimensional image modality which captures the light intensity (similar to regular imaging), as well as the light direction traveling into the 3D space. 4D light field images and 5D light field video can be captured by either plenoptic cameras or multi-array cameras.[3] Due to the added spatial (view) dimension the resulted amount of data significantly increase in comparison to conventional image modalities such as image and video. Therefore, new coding solutions specifically designed for compressing immersive data have been announced by standardization bodies.

A 5D (2 spatial dimensions + 2 view dimensions + 1 time dimension) light field video can be interpreted as multi-view video, which consists of a single video sequence for each viewpoint. Therefore, multi-view coding (MVC) standards are widely used for compressing such data. MVC coding solutions can efficiently compress video sequences simultaneously captured from multiple camera angles by exploiting not only redundancies in space and in time, but also redundancies between different views. In practice, in addition to the motion compensation

Further author information: vasileios.avramelos@ugent.be

algorithms for temporal prediction, MVC utilizes inter-view motion compensation for inter-view prediction as well.

The latest and most popular representative of MVC, is the multi-view high efficiency video coding (MV-HEVC). MV-HEVC is an extension of the high efficiency video coding standard (HEVC), which was standardized in 2013 and it is now ready to be succeed by the new versatile video coding standard (VVC) by the year 2021.[4] It treats each sub-aperture view as a different HEVC sequence, and with the deployment of inter-view motion compensation performs significantly well in terms of compression efficiency.[5,6] A similar format, initially developed for compressing 3D video, can compress light light field data by treating only one given view (typically the center view) or a set of views as HEVC sequences, and the rest sub-aperture views as depth maps. This methodology is called multi-view plus depth (MVD) and it is preferred over MVC because it is not limited to a given number of views, i.e., it can synthesize additional views at the decoder side. However, MVC and specifically MV-HEVC solutions are being favoured for more straight forward implementations since they do not require additional depth coding, depth estimation and view synthesizing tools. Additionally, MV-HEVC simply uses HEVC syntax elements, something which facilitates at a great extent the whole encoding configuration for the standard user.

For successfully exploiting temporal and inter-view redundancies, prediction structures are deployed to define the order and the inter-dependencies between frames in time and views in space. There is literally a vast amount of possibilities on how to structure the prediction schemes of an HEVC-based encoder. The latest coding standards allow three types of pictures, namely I, P and B, for pictures using intra-prediction, unidirectional prediction and bidirectional prediction respectively. Various research efforts investigate optimal prediction structures for different coding scenarios. While the main interest of those efforts is temporal prediction for regular video, due to the intermittently rise of technologies such as 3D video, stereoscopic video and light field video, specific prediction schemes for inter-view coding have been investigated as well. Merkle et al., presented efficient prediction structures for multi-view video coding with H.264/AVC aiming for compression efficiency.[7] Nasri et al. investigate MVC inter-view prediction to lower the view random access complexity.[8] These prediction structures covered conventional 1D multi-view video, as well as 2D multi-view video captured by multi-camera arrays by using IPPP and IBPBP structures widely used in conventional video coding. Typical hierarchical B prediction schemes were implemented for inter-view coding in MVC by converting 2D to 1D multi-view video with a corresponding scanning topology (e.g., snake, spiral, etc.)[9,10] Moreover, the case of using the center view of a multi-view video as a key point and exploit horizontal and vertical correlations was also investigated in the literature of multi-view and light field video coding.[11,12] Wang et al., implemented a prediction structure for both temporal and inter-view prediction specifically for coding light field video. The authors effectively used two-directional parallel inter-view prediction to outperform conventional MVC structures implemented on light field video.[13] Khoury et al, kept the same temporal prediction structure but modified the view ordering structure to obtain better results in terms of compression efficiency.[12] Lastly, in previous work and within the context of this work, we proposed a three-directional scheme exploiting horizontal, vertical and diagonal correlations as well as a prediction scheme specifically designed to maximize view random access.[14,15]

In this work, we focus on inter-view prediction and we evaluate and compare a selection of the above mentioned prediction schemes for light field video coding. Specifically, we validate the performance of those prediction schemes in terms of rate-distortion, and we examine the necessary trade-offs for gaining random access between different views (view random access). The results can be used as a benchmark for the development of light field video coding solutions with MV-HEVC. In the next section (Sec. 2), a brief overview of prediction structures that are used in the literature and in this work is given, and in Sec. 3 we present the rate-distortion results for enabling comparisons. Finally, Sec. 4 discusses the comparability and the usability of those results.

2. PREDICTION STRUCTURES

In this section we will briefly discuss important prediction structures that have been widely used in video coding, multi-view video coding and light field video coding. Starting from regular video coding, we present how temporal prediction structures are adapted to exploit correlations between neighboring views in multi-view and light field video compression.

2.1 Prediction structures in video coding

Standardized video coding standards typically rely on a hybrid scheme based on a transform step (e.g., discrete wavelet or discrete cosine transform) and an intra-prediction/motion-compensation step. The former step efficiently exploits spatial and temporal redundancies using difference coding and motion compensation. To minimize temporal redundancies, efficient prediction structures have been proposed and standardized over time. These structures define the encoding order between subsequent frames as well as the inter-prediction dependencies. As seen in Fig. 1, popular established structures include:[16]

- *all-Intra* - all frames are I frames, i.e., coded using only intra-frame coding and therefore only spatial dependencies within a frame are being exploited. This scheme has numerous advantages but it is not great in terms of compression efficiency.
- *Regular P prediction* - for every group of pictures (GOP), the first frame is an I frame and the rest (P frames) are using a previous or following frame as a reference.
- *Hierarchical B prediction* - for every GOP, the first frame is an I frame and the rest (P or B frames) are either uni- or bi-directionally predicted by previous or following frames.

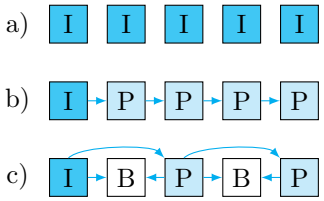

Figure 1. Typical prediction structures in video coding. From top to bottom, *all-Intra*, *regular P prediction* and *hierarchical B prediction*.

There is a vast amount of options on how to structure the prediction scheme in an encoding system. From the GOP size to the frame type (I, P or B) and from the intra-period (frequency of the appearance of an I frame) to the actual dependencies, the desired trade-off between compression efficiency, random access and computational complexity can be reached.

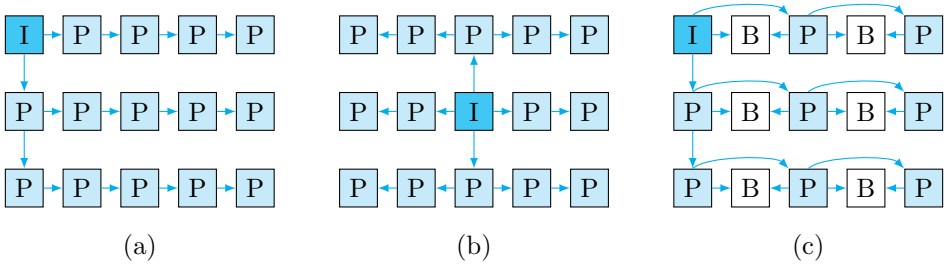

Figure 2. Prediction structures in multi-view video coding. From left to right, an adapted *regular P prediction*, a *regular P prediction* starting from the center tile/view, and an adapted *hierarchical B prediction* structure.

2.2 Prediction structures in multi-view video coding

In MVC, besides the inter-frame dependencies in time, inter-view dependencies can be similarly set in order to exploit redundancies between the neighboring views of the multi-view vector. When the different viewpoints are two-dimensional, e.g., captured by a 2D camera array, then the corresponding 2D multi-view array can be easily converted to a 1D video sequence vector, and the prediction structures of Sec. 2.1 can be implemented accordingly. Fig. 2, shows an example of two *regular P* and one *hierarchical B prediction* structure adapted for 2D multi-view video sequences.[7]

A typical example of an MVC prediction structure widely used for two-dimensional multi-view sequences can be seen in Fig. 3. The prediction flow is following a "snake" scan (row-by-row) from the top left to the bottom right. This is the most straight forward 2D prediction structure and it has been also adopted in light field video coding.[9,13] Choosing the most appropriate scanning topology for a corresponding application is a crucial point for reaching the desired performance.

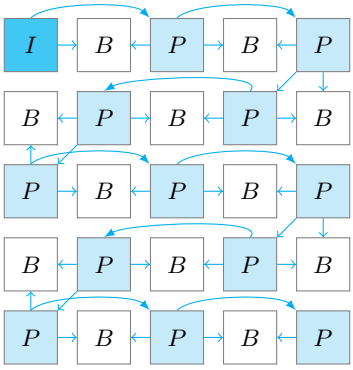

Figure 3. Typical MVC prediction structure using a "snake" scan (row-by-row).

2.3 Prediction structures for light field video coding

A light field video can be either captured using a light field camera or a multi-camera array. Therefore, a light field video can be directly interpreted as a 2D set of multi-view video sequences. In this case, MV-HEVC is a straight forward solution for compressing such data. Typical scanning topologies, such as row-by-row "snake" scanning (see Fig. 3), can be used in combination with a P or B prediction scheme for exploiting redundancies and deliver acceptable compression rates. Other scanning topologies may be raster, zig-zag or spiral.

Wang et al. highlighted the limitations of conventional prediction schemes and proposed a prediction structure specifically tailored to exploit vertical and horizontal correlation between neighboring views.[13] Moreover, the scheme proposed by Wang et al., increases the number of B frames. In theory, increasing the number of B frames results to compression gains. The authors proved that with such an hierarchical B inter-view prediction scheme, the state-of-the-art by that time could be outperformed in terms of compression efficiency. The proposed scheme can be seen in Fig. 4 (left).

Khoury et al., modified the hierarchical B prediction scheme of Wang et al. by only moving the reference I view at the center of the grid. As such, due to the fact that the center tile is better correlated with the surrounding tiles/views, the current scheme achieves an approximate of 25% - 35% of bitrate savings for the same reconstruction quality (tested on two data sets).[12] The current scheme can be seen in Fig. 4 (right).

In our previous work, we needed a robust HEVC-based light field video coding scheme for stress-testing a new unconventional rendering method.[14] The proposed *Full* scheme is a fully-referenced reverse spiral scanning technique which starts from a corner of the grid and revolves in a clock-wise manner until reaching the center tiles/views. The B views were maximized by only using one I view (the first one), one P view (the second one) and all the rest are B views using two neighboring viewpoints for prediction. Those two views can be either the closest horizontal, vertical or diagonal view with a slight preference to the horizontal one since the human visual system is more sensitive to horizontal correlations.[17] The *Full* scheme can be seen in Fig. 5 (right).

The *Full* scheme was also used for comparison reasons in a previous work which was focusing on random access prediction schemes for light field video coding.[15] In that research work, a comparison between an all-intra scheme (full temporal and view random access), a fully-referenced scheme (increased number of inter-dependencies for improving compression rates but losing in terms of random access), a simulcast scheme (no inter-view prediction), and a scheme specifically tailored for applications where view random access is important. For the former scheme, namely *Center*, every frame of every view is predicted by the corresponding frame of the center tile/view. In that way when the user requests a change of view at a random point in time, only two views (the center and the requested) need to be decoded. This scheme can be seen in Fig.5 (left).

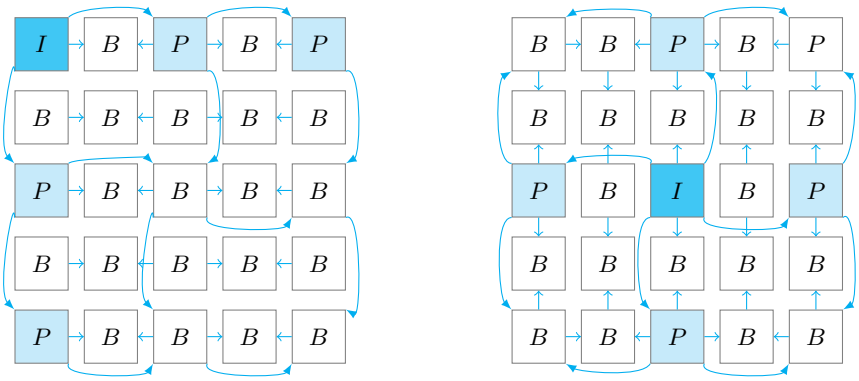

Figure 4. Grid view of a 5×5 light field view ordering structure as presented by Wang et al.[13] (left) and by Khoury et al.[12] (right). A similar referencing structure has been derived for the 8×8 case of this work.

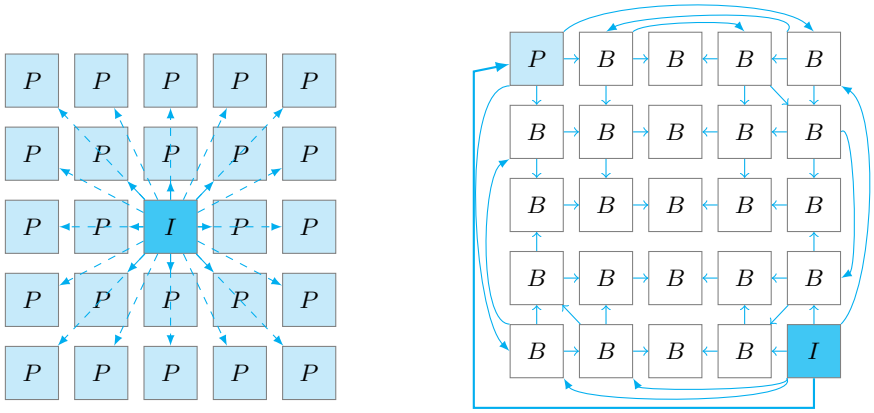

Figure 5. Grid view of a 5×5 light field view ordering structure as presented within the context of this project.[14,15] On the left, a prediction scheme aiming towards view random access, and on the right, a scheme aiming towards compression efficiency. A similar referencing structure has been derived for the 8×8 case of this work

2.4 Requirements for multi-view video coding

Experts in the field of video coding, and more specifically MVC, defined a number of requirements that every MVC algorithm must or should satisfy.[18] The same implications stand for compressing light field video using MVC techniques. These requirements fall in different categories from compression to transmission. In this work we only focus on compression related requirements. Those prerequisites cover, among others, the following important aspects:

- *Compression efficiency* - the best trade-off between bitrate (number of bits of the compressed bitstream) and distortion (the objective reconstruction quality)
- *View scalability* - minimum decoding effort for accessing selected views
- *Low delay* - low latency encoding/decoding suitable for real-time applications
- *Picture quality among views* - consistent or flexible quality allocation between views
- *Temporal random access* - random access in the time dimension
- *View random access* - random access in the view dimension

In this overview, a comparison of 2D inter-view prediction structures which can be used in MVC for compressing light field video sequences is presented. Our main focus is compression efficiency and view random

access. In the next sections, Sec. 3 and 4, the experimental results of this overview are being presented and discussed respectively.

3. EXPERIMENTAL RESULTS

In this section, we enable a comparison between selected prediction structures from the literature, mainly in terms of compression efficiency. Rate-distortion curves have been used for visualizing compression rates and objective reconstruction quality. In the following, the dataset (light field video sequences) is presented. Finally, the results and the observations of the comparison are discussed.

3.1 Dataset

Three light field video sequences were used to evaluate the performance of MV-HEVC compression with different inter-view prediction structures. The dataset consists of two light field video sequences of resolution 512×352 and one of 544×320 at 30 frames per second for approximately 100 frames and 8×8 views. Respectively, the video sequences are named *cats*, *train1* and *train2* and a thumbnail of those sequences can be seen in Fig. 6. It should be noted here that the sequences originate from interpolating light fields from a plenoptic camera and some artifacts are caused by this process.[19] However, since we objectively compare the reconstruction quality, those artifacts do not affect the results.

In order to assess the reconstruction quality we compared the signal-to-noise ratio of the pixel intensities of the light field video sequences. To be 100% correct, all views, all frames, all pixels should be evaluated. However, to speed up the measurement process, we subsampled the views and only chose five random views for evaluation. More specifically, view22, view36, view44, view67 and view72 were used, where for instance view22 corresponds to viewXY (X and Y being the spatial coordinates of the view on the view grid).

Figure 6. The three different light field video sequences used in this work. From left to right: *cats* - 512×352 109 frames, *train1* - 512×352 84 frames, and *train2* - 544×320 97 frames.[19]

3.2 Results

The goal of this work was to enable a comparison between different 2D inter-view prediction structures for compressing light field video sequences. The mean for this comparison is rate-distortion curves. In terms of rate, the amount of compressed bits per second were measured in kilo-bit per second (kbps), while in terms of distortion, the peak signal-to-noise ratio (PSNR) was measured in decibel (dB).

The four different prediction structures are named as *Full*, *Khoury*, *Wang* and *Center*, they can be found in the literature[12–15] and they are briefly described in Sec. 2. The software used for implementing these schemes is the MV-HEVC reference coder HTM-16.2.[20] It should be noted here that this work focused on providing a benchmark on inter-view prediction structures aiming for both view random access and compression efficiency. Therefore, the temporal prediction was modified and consistently maintained for all schemes as follows. The first view (I view), used a default IBPBP scheme with a GOP size of 8 and an intra-period of 24. Every other P or B view used the corresponding frame in time from the first I view as a reference. In that way, the necessary number of decoded frames in the case of a view switch by the user (view random access complexity), is significantly decreased. However, we kept the nomenclature from the literature since the presented inter-view schemes were directly adopted from those publications.

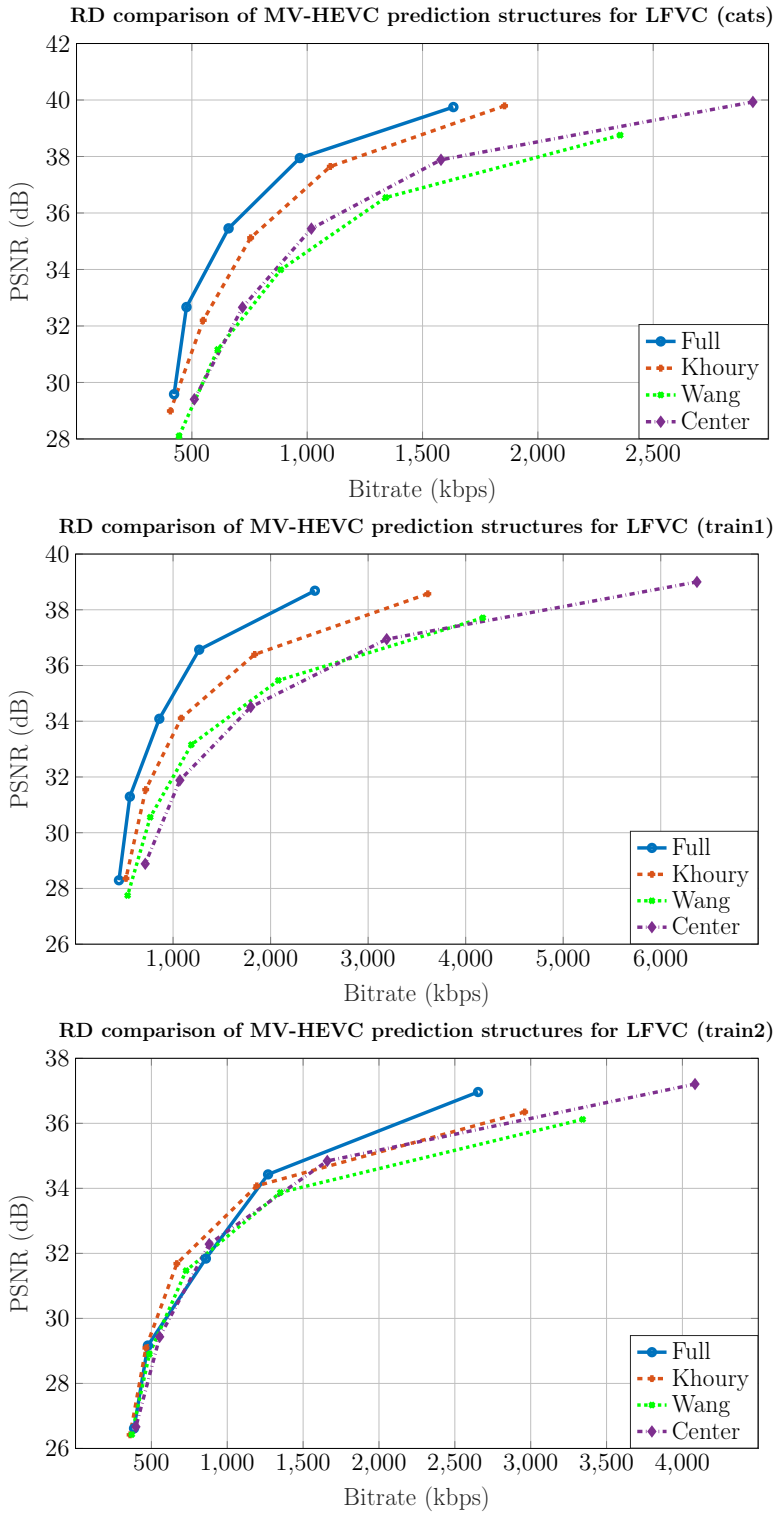

Figure 7. Rate-distortion curves for prediction structures evaluated in this work. From top to bottom, the results for the three light field video sequences (*cats, train1, train2*) are presented.

As seen in Fig. 7, *Wang* and *Khoury* validate the initial assumption as seen in the literature: *Khoury* performs noticeably better by exploiting the better correlation between the center tile/view and the neighboring views. The *Full* scheme maximizes the number of B views and utilizes a reverse spiral scanning topology. The B views are using the closest horizontal, vertical or diagonal neighbor to provide a significant improvement in terms of compression efficiency. Not unexpectedly, the resulted compression gains are content-dependent. However, the multiple dependencies between views, can slightly complicate view random access (which is clearly not content-dependent). With the *Center* prediction scheme, we can avoid the former issue and minimize the required dependencies, and consequently the number of decoded views/frames when for example a view switch is requested. The *Center* prediction scheme shows the necessary trade-off in terms of rate-distortion in the case of applications with crucial view random access requirements.

4. CONCLUSIONS

The goal of this work was to provide a useful benchmark for light field video coding applications. More specifically, a selection of inter-view prediction structures from the literature and from our previous work was implemented for enabling a useful comparison in terms of compression rate and distortion. The two best performing inter-view prediction schemes were adopted from the literature and compared with an alternative scheme (in terms of topological space), as well as with a scheme enabling view random access capabilities. Results showed that all schemes provide an acceptable performance, and they also show that the necessary trade-off for view random access is rather acceptable in most cases. In future work, MVD techniques will be also implemented and compared to the MVC techniques. Ideally, by that time, a preliminary comparison with early versions of the new MPEG-I standard (upcoming standard for specifically compressing immersive data) will be possible.

ACKNOWLEDGMENTS

The research activities in this paper were funded by IDLab (Ghent University-imec), Flanders Innovation and Entrepreneurship (VLAIO), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

REFERENCES

[1] Levoy, M. and Hanrahan, P., "Light field rendering," in [*Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*], *SIGGRAPH '96*, ACM Press (1996).

[2] Anthes, C., García-Hernández, R. J., Wiedemann, M., and Kranzlmüller, D., "State of the art of virtual reality technology," in [*IEEE Aerospace Conference*], (2017).

[3] Adelson, E. and Bergen, J., "The plenoptic function and the elements of early vision," in [*Computational Models of Visual Processing*], 3–20, MIT Press (1991).

[4] Sullivan, G. J., Ohm, J. R., Han, W. J., and Wiegand, T., "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology* **22**, 1649–1668 (2012).

[5] Sullivan, G. J., Boyce, J. M., Chen, Y., Ohm, J. R., Segall, C. A., and Vetro, A., "Standardized Extensions of High Efficiency Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology* **22**, 1649–1668 (2012).

[6] Hannuksela, M. M., Yan, Y., Huang, X., and Li, H., "Overview of the Multiview High Efficiency Video Coding (MV-HEVC) Standard," in [*IEEE International Conference on Image Processing (ICIP)*], 2154–2158 (2017).

[7] Merkle, P., Smolic, A., Muller, K., and Wiegand, T., "Efficient Prediction Structures for Multi-view Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology* **17**, 1461–1473 (2007).

[8] Nasri, S. A. E., Khelil, K., and Doghmane, N., "Enhanced view random access ability for multiview video coding," *Journal of Electronic Imaging (SPIE)* **25** (2016).

[9] Chung, T. Y., Jung, I. L., Song, K., and Kim, C. S., "Multi-view video coding with view interpolation prediction for 2D camera arrays," *Journal of Visual Communication and Image Representation* **21**, 474–476 (2010).

[10] Shi, S., Gioia, P., and Madec, G., "Efficient compression method for integral images using multi-view video coding," in [*Proceedings of the 18th IEEE International Conference on Image Processing (ICIP)*], 137–140 (2011).

[11] Dricot, A., Jung, J., Gagnazzo, M., Pesquet, B., and Dufaux, F., "Full parallax super multi-view video coding," in [*Proceedings of IEEE International Conference on Image Processing (ICIP)*], 135–139 (2011).

[12] Khoury, J., Purazad, M. T., and Nasiopoulos, P., "A New Prediction Structure for Efficient MV-HEVC based Light Field Video Compression," in [*International Conference on Computing, Networking and Communications (ICNC)*], 588–591 (2019).

[13] Wang, G., Xiang, W., Pickering, M., and Chen, C. W., "Light Field Multi-View Video Coding With Two-Directional Parallel Inter-View Prediction," *IEEE Transactions on Image Processing* **25**, 5104–5117 (2016).

[14] Avramelos, V., Saenen, I., Verhack, R., Wallendael, G. V., Lambert, P., and Sikora, T., "Steered Mixture-of-Experts for Light Field Video Coding," in [*Proceedings of SPIE 10752, Applications of Digital Image Processing XLI*], (2018).

[15] Avramelos, V., Praeter, J. D., Wallendael, G. V., and Lambert, P., "Random access prediction structures for light field video coding with MV-HEVC," *Submitted to Multimedia Tools and Applications* (2019).

[16] Sze, V., Budagavi, M., and Sullivan, G. J., "High Efficiency Video Coding - Algorithms and Architectures," in [*Integrated Circuits and Systems*], Springer Publishing Company (2014).

[17] Hansen, B. C. and Essock, E. A., "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes," *Journal of Vision* **4**, 1044–1060 (2004).

[18] Ho, Y. S. and Oh, K. J., "Overview of Multi-view Video Coding," in [*14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*], (2007).

[19] Wang, T. C., Zhu, J. Y., Kalantari, N. K., Efros, A. A., and Ramamoorthi, R., "Light field video capture using a learning-based hybrid imaging system," *ACM Transactions on Graphics (Proceedings of SIGGRAPH '17)* **36** (2017).

[20] Fraunhofer-HHI, "Multiview High Efficiency Video Coding (MV-HEVC) - HTM software repository." https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/. (Accessed: 16 July 2019).