Optimized Reference Picture Selection for Light Field Image Coding

Ricardo J. S. Monteiro[1,2], Nuno M. M. Rodrigues[1,3], Sérgio M. M. Faria[1,3], Paulo J. L. Nunes[1,2]

[1]Instituto de Telecomunicações; [2]Instituto Universitário de Lisboa (ISCTE-IUL); [3]ESTG, Instituto Politécnico de Leiria; Portugal;

e-mails: {ricardo.monteiro, paulo.nunes}@lx.it.pt, {nuno.rodrigues, sergio.faria}@co.it.pt

***Abstract*—This paper proposes a new reference picture selection method for light field image coding using the pseudo-video sequence (PVS) format. State-of-the-art solutions to encode light field images using the PVS format rely on video coding standards to exploit the inter-view redundancy between each sub-aperture image (SAI) that composes the light field. However, the PVS scanning order is not usually considered by the video codec. The proposed solution signals the PVS scanning order to the decoder, enabling implicit optimized reference picture selection for each specific scanning order. With the proposed method each reference picture is selected by minimizing the Euclidean distance to the current SAI being encoded. Experimental results show that, for the same PVS scanning order, the proposed optimized reference picture selection codec outperforms HEVC video coding standard for light field image coding, up to 50% in terms of bitrate savings.**

Keywords—Light Field Image Coding, Pseudo-video sequence, optimized reference picture selection, HEVC

I. Introduction

Light field (LF) imaging derives from the fundamentals of LF sampling, where not only spatial information about the scene is captured but also angular information. This is possible by using a microlens array (MLA) placed between the camera's main lens and the sensor [1]. By combining both the spatial and angular information about the light being acquired, each microlens creates a micro-image (MI) on the sensor.

As a result, the LF format allows some additional features when compared with traditional 2D images, such as focus and perspective change after the picture has been taken, since the information being captured is 3D [1]. LF imaging has become a promising approach for 3D imaging and sensing, having potential applications in many different areas of research, such as 3D television [2], richer image capturing [1], image recognition and medical imaging [3]. Additionally, the intrinsic scalability of the LF imaging is also an attractive feature, where from one LF image several signals can be generated, e.g., a single 2D view, stereoscopic 3D or Multiview images [1].

Novel standardization initiatives on image and video coding have recently considered the LF technology: the JPEG Pleno [4], targeting richer image capturing, visualization, and manipulation, which includes not only LF but also point-cloud and holographic technologies; and the MPEG-I [5], targeting immersive media such as 360 video and virtual navigation.

Due to the large amount of data involved in LF imaging, providing efficient transmission of LF content over limited bandwidth networks requires efficient coding tools. In this sense, several LF image coding schemes have been proposed in the literature. These schemes exploit the additional redundancy of this type of content, which is usually referred to as non-local spatial redundancy, i.e., the redundancy between neighboring MIs, which is a consequence of the used optical system for capturing LF images.

The proposed schemes in the literature, which aim to exploit the non-local spatial redundancy, are normally focused on transform- or search-based methods. In [6]–[8] the discrete cosine transform (DCT) is used to this end. In [7], a 3D-DCT is applied to a stack of MIs, to exploit the existing redundancy between the several MIs and within the same MI. In [8], a 4D-DCT is used in combination with hexadeca-trees to divide the LF into grouped 4D blocks to generate a stream, which is then encoded using adaptive arithmetic coding. Regarding search-based methods, several authors proposed different approaches [9]–[13]. In [9], a self-similarity (SS) compensated prediction is introduced into HEVC to exploit the redundancy between neighboring MI in the LF. This SS prediction was extended as a bidirectional prediction method in [10] and [14] demonstrating a significant increase in terms of coding efficiency over the unidirectional approach. These approaches are considered low order prediction because they are limited to 2 degrees of freedom (DoF). In order to further exploit the 3D nature of the LF images, the authors in [12] proposed a new approach including an unidirectional prediction mode that uses up to 8 DoF. In [11], another prediction method was proposed to improve the prediction of the non-local spatial redundancy based on locally linear embedding (LLE) that allows a linear combination of 1 to 8 unidirectional searches.

Another important group of techniques converts the LF image into a group of sub-aperture images (SAIs), which can be interpreted as a 2D or Multiview video signal and therefore encoded using standard codecs with their highly efficient prediction tools [15]–[17]. A video signal generated from a set of SAIs of a LF image is usually referred to a pseudo-video sequence (PVS). A PVS is generated by organizing the SAIs into a video sequence using a specific scanning order. The inter-view redundancy between the SAIs of the PVS is exploited by the inter prediction tools of the used video codec (e.g., HEVC). These techniques are very efficient especially on Lenslet-based LF images because the disparity between neighboring SAIs is very low. However, the chosen PVS scanning order may affect the coding efficiency. Several scanning orders have been studied, namely, raster and spiral [15]. Alternatively to the PVS approach, the authors in [18], [19] interpreted the LF as a multiview signal. By using MV-HEVC, a two-dimensional reference picture prediction is available, i.e., temporal and inter-view prediction, which allows encoding a LF image coding as a Multiview signal.

978-9-0827-9703-9/19/$31.00

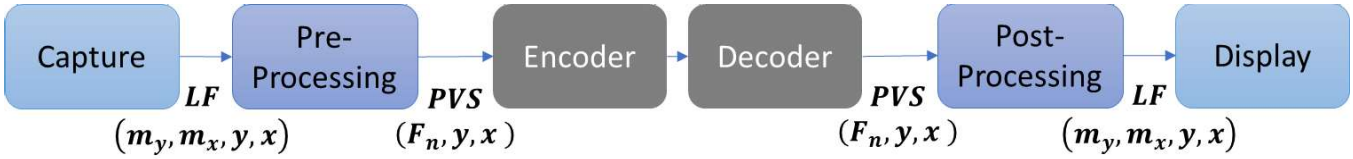

Fig. 1. Processing chain for Lenslet LF including pre- and post-processing that allows the conversion to the PVS format

Additionally, other approaches using SAI-based coding techniques include transmitting only part of the of the SAIs, i.e., transmitting structural key views (SKVs) and then using additional information and the SKVs to reconstruct the remaining SAIs at the decoder. Several authors also proposed techniques based on this approach [20]–[22], differing on the process to additional information used to generate the remaining SAIs. This information may be obtained by using: a convolutional neural network (CNN) based on an angular super resolution algorithm [20]; weighting coefficients that are generated through linear approximation [21]; or depth image based rendering methods [22].

Although the coding efficiency of PVS-based LF coding approaches comes from the ability to use the efficient prediction tools of standard video codecs [15], it can be further improved by exploiting the knowledge of the scanning order that was used to generate the PVS. Therefore, this paper proposes to signal the PVS scanning order to the decoder enabling the generation of an optimized implicit reference picture selection (RPS). Although this optimized RPS can be applied to any scanning order, the spiral scan was used since it is more efficient than the raster scan [15]. Experimental results show that the proposed approach can outperform state-of-the-art LF image codecs for both MI- and SAI-based approaches.

The remainder of this paper is organized as follows: Section 2 presents the proposed optimized RPS, Section 3 presents and discusses the experimental results, and, finally, Section 4 concludes the paper.

II. Proposed Optimized Reference Picture Selection

Encoding a LF image using the PVS format requires additional pre- and post-processing steps before and after the encoding step, respectively. The pre-processing step organizes SAIs using a specific scanning order and the post-processing step reconstructs the original LF image. The complete processing chain for Lenslet LF images including the pre- and post-processing steps is shown in Fig. 1.

If the decoder is aware of the PVS scanning order in the pre-processing step, the original spatial position of the encoded SAI can be determined based on the frame number. In order to make the decoder aware of the specific scanning order that is being used, this information can be transmitted as an index in the supplemental enhancement information (SEI) message [23]. Therefore, since the decoder is aware of the PVS scanning order some features can be implemented to increase the compression efficiency, namely: 1) optimized RPS based on a specific scanning order that is being used; 2) reduction of the amount of information transmitted in the bitstream to signal the RPS structure (detailed in the following subsections). Additionally, since the index that is transmitted as a SEI message identifies the PVS scanning order, the proposed codec can be used for any list of PVS scanning orders that are established between the encoder and the decoder.

A. Spiral PVS scanning order

There are several strategies described in the literature to perform the PVS scanning order, being the most popular the raster and the spiral scanning. In [15], the authors showed that the spiral scanning is more efficient than the raster scanning, therefore the spiral scanning was used in the proposed codec.

To determine the spatial position of a SAI relatively to the LF, it is necessary to establish a relationship between the spatial position and the frame number. Fig. 2 shows the spiral scan being applied to an $N \times N$ matrix of SAIs. We define $j = [0, 1, \dots, N-1]$ and $i = [0, 1, \dots, N-1]$ as the vertical and horizontal axis spatial positions, respectively, for each SAI in the $N \times N$ matrix. For every two-dimensional SAI position, P_{ji}, the frame number, F_{ji}, can be calculated using:

$$F_{ji} = \begin{cases} (N - 2 \propto_{ji})^2 - (j - \propto_{ji}) - (i - \propto_{ji}) - 1 & if\ j \le i \\ (N - 2 \propto_{ji} - 2)^2 + (j - \propto_{ji}) + (i - \propto_{ji}) - 1 & if\ j > i \end{cases} \quad (1)$$

where:

$$\propto_{ji} = \min(j, i, N-1-j, N-1-i) \quad (2)$$

Since (1) defines the relationship between the spatial positions, P_{ji}, of every SAI and the correspondent frame number, F_{ji}, this information can be computed in the beginning of the coding and decoding process.

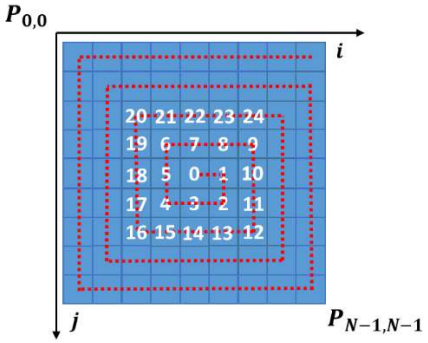

Fig. 2. Spiral PVS scanning order applied to a $N \times N$ matrix of SAIs

B. Optimized RPS

As mentioned before, video coding standards like HEVC have inter-prediction tools are applied to exploit the inter-view redundancy between different SAIs that compose the LF PVS. However, when encoding a LF PVS if the codec is not aware of the scanning order, then the reference picture selection is not going to be optimal.

When using the default HEVC "Low Delay" configuration, the inherent 2D spatial locations of each SAI are not considered. Consequently, the correlation between the selected reference pictures and the current SAI will be lower.

In Fig. 3 the RPS for frames 12 and 22 is shown when using the "Low Delay" configuration using four reference pictures. As it is shown in Fig. 3 the RPS is not optimal for most of the reference pictures. When encoding frame 22, frames 6, 7 and 8 are closer, in terms of spatial position, when compared to frames 20, 16 and 12. This results in lower correlation between the reference pictures and the current SAI and, therefore, lower coding efficiency. Moreover, even when using a more classic "Low Delay" configuration, where the four reference pictures are always the last 4 frames that were encoded, the selection is still not optimal. In this case when encoding frame 22, frames 21, 20, 19 and 18 would be selected as reference frames.

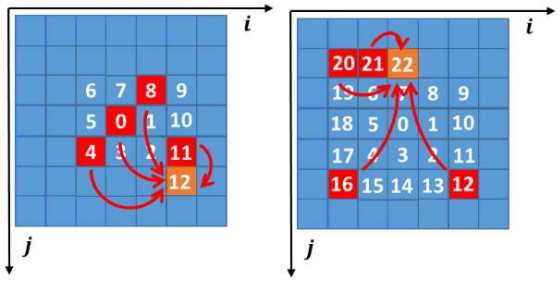

Fig. 3. RPS for frames 12 and 22 when using the "Low Delay" configuration

To create an optimized RPS, the Euclidean distance, d, between the spatial positions of R reference pictures, P_{ji}^{r}, and the current SAI, P_{ji}, should be minimized. The Euclidean distance is computed as:

$$d(P_{ji}, P_{ji}^{r}) = \sqrt{(j - j_r)^2 + (i - i_r)^2}, \quad (3)$$

where $r = [0,1,\dots,R-1]$. The Euclidean distance between the current SAI and the remaining available encoded SAIs is, therefore, calculated. Once the R closest reference pictures, in terms of Euclidean distance, to the current SAI are found, the selected reference pictures are organized in an ascendant order of distance, in the reference picture list of each SAI that is being encoded. When two or more reference pictures have the same Euclidean distance, the one with the lowest F_{ji} is selected first.

Fig. 4 shows the optimized RPS for $R = 4$, after minimizing the Euclidean distance as in (3), for the examples of frame 12 and 22 previously shown in Fig. 3.

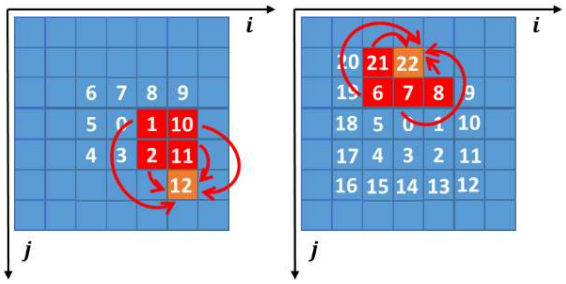

Fig. 4. Optimized RPS for frames 12 and 22

C. *Implicit RPS signaling*

The RPS signaling allows the HEVC decoder to configure the reference picture structure. Indicating for each frame that is being decoded, what reference pictures are necessary for the decoding process to take place normally. HEVC allows for the RPS to be transmitted for every single frame [23]. Additionally, depending if RPS prediction is being used or not, several flags are transmitted that allow RPS to be described. However, since the proposed decoder is aware of the scanning order and the RPS is done in an optimized implicit way, it is not necessary to transmit any of the flags that are responsible for the RPS signaling and configuration. Additionally, the extra index that is transmitted through a SEI message, by the proposed codec, can also indicate that the bitstream is using the regular RPS signaling. This allows the proposed codec to be able to decode any established scanning order between the encoder and the decoder, and also, decode regular HEVC bitstreams.

III. Experimental Results

In this section the experimental results of the proposed PVS-based coding solution for LF image coding are evaluated and compared against state-of-the-art PVS- and MI-based LF image coding methods. The PVS-based approaches are referred to as HEVC-PVS-LD1 [15] and HEVC-PVS-LD2, i.e., PVS-based methods using the spiral scan with two different HEVC "Low Delay" configurations. The proposed optimized RPS uses HEVC-PVS-LD1 as the basis approach and it is referred to as HEVC-OPT. The MI-based approaches are: HEVC-SS [9], HEVC-LLE [11] and HEVC-HOP [12]. The basis codec for the implementation of both PVS- and MI-based LF image is an HEVC implementation referred to as HM-15.0.

To evaluate the performance of the proposed optimized RPS solution, a subset of the EPFL dataset is used composing 12 LF images. These images were captured using a Lytro Illum. The "RAW" Lenslet LF images are converted to the 4D LF format using the LF Toolbox [24] as suggested by JPEG Pleno Common Test Conditions document [25]. For HEVC-PVS-LD1, HEVC-PVS-LD2 and HEVC-OPT the used processing chain is shown in Fig. 1. The LF images are converted to the PVS representation using the spiral scanning method, prior to being encoded. After the pre-processing step, the PVS-based solutions encode 13×13 SAIs with a resolution of 625×434 pixels. On the other hand, the MI-based solutions encode one LF image with a resolution of 8125×5642 pixels, composed by MIs with 13×13 pixels. For both PVS and MI formats, after the decoding step, 13×13 SAIs are generated with a resolution of 625×434 pixels, using the YUV 4:2:0 8-bit color format. The reconstructed SAIs after the post-processing step are compared with the reference SAIs. The reference SAIs are generated through the same process as the decoded SAIs but without the encoding and decoding steps.

For HEVC-OPT, HEVC-PVS-LD1 and HEVC-PVS-LD2, QPs 17, 22, 27, 32 and 37 were used and for HEVC-SS. HEVC-LLE and HEVC-HOP, QPs 22, 27, 32, 37 and 42 were used. The different QPs allow the use of a common bitrate range for every tested codec, allowing a direct comparison. The base configuration for both HEVC-OPT and HEVC-PVS-LD1 is the "Low Delay" with B slices configuration. For HEVC-PVS-LD2, the default HEVC "Low Delay" with B slices configuration was changed so the reference pictures are the last 4 encoded frames. For HEVC-SS, HEVC-LLE and HEVC-HOP the intra main configuration was used. The RD analysis is done by comparing the size of the bitstream (rate) and the average PSNR-YUV of the 13×13 SAIs (distortion) generated at the decoder side for each codec. The average PSNR-YUV of the 13×13 SAIs is calculated by comparing the decoded SAIs encoded by different codecs, and the reference 13×13 SAIs.

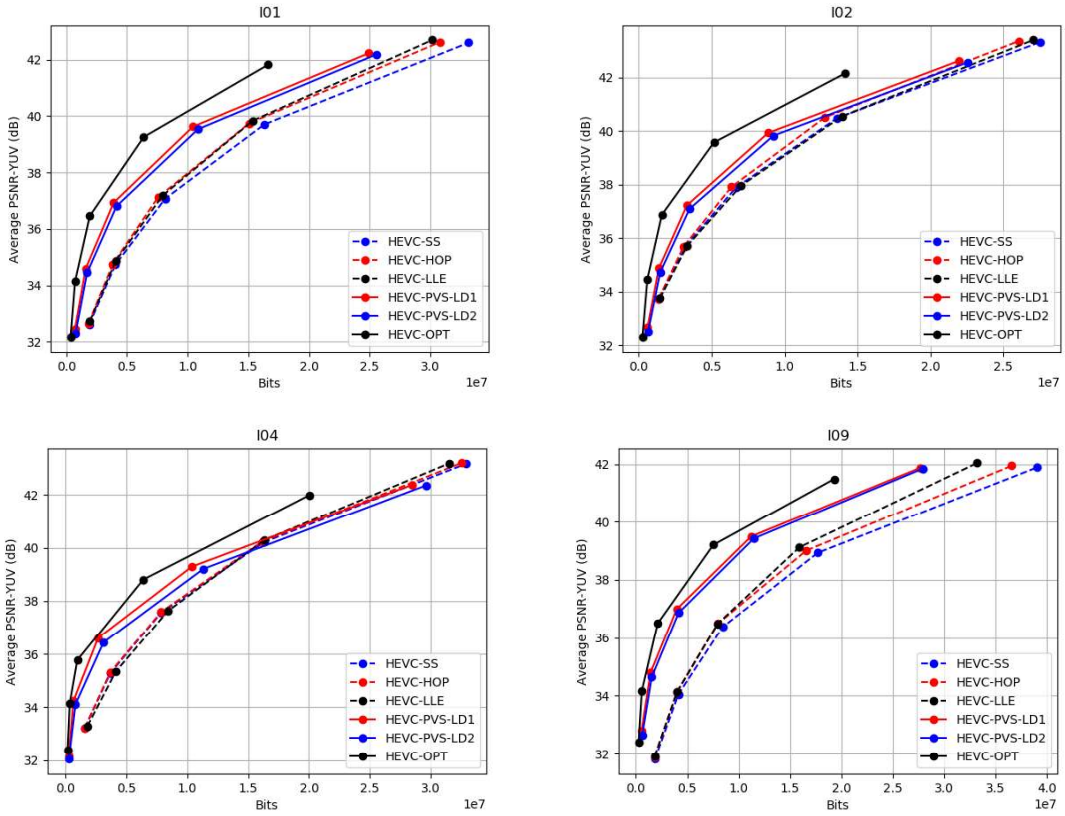

Fig. 5. RD curves comparing PVS- and MI-based LF coding methods for images I01, I02, I04 and I09

TABLE I. BD-PSNR-YUV AND BD-RATE RESULTS AGAINST HEVC-PVS-LD1 USING HEVC-OPT CODEC

Image	BD-PSNR-YUV	BD-RATE
I01	1.21 dB	-37.60 %
I02	1.37 dB	-41.85 %
I03	1.39 dB	-43.38 %
I04	0.80 dB	-32.73 %
I05	0.99 dB	-38.71 %
I06	0.91 dB	-49.88 %
I07	0.82 dB	-34.25 %
I08	0.54 dB	-33.93 %
I09	0.86 dB	-32.77 %
I10	1.05 dB	-39.58 %
I11	0.51 dB	-28.86 %
I12	0.51 dB	-25.65 %
AVG.	**0.91 dB**	**-36.60 %**

The RD curves comparing all PVS- and MI-based LF image codecs for images I01, I02, I04 and I09 are shown in Fig. 5. From the RD curves in Fig. 5 it is possible to conclude that, in general, the PVS-based LF image codecs are more efficient than the MI-based LF images codecs across the tested bitrates. This is the case for all the 12 images from the EPFL dataset. When analyzing the RD curves of both HEVC-PVS-LD1, HEVC-PVS-LD2 and the proposed HEVC-OPT it is possible to observe that HEVC-OPT achieves higher coding efficiency for every single point of the RD curve. The difference in results from HEVC-PVS-LD1 to HEVC-PVS-LD2 is the used rate allocation. In HEVC the "Low Delay" configuration has an associated QP offset by frame, which varies from 1 to 3 every 4 frames. In HEVC-PVS-LD1 the frames with the lowest QP are chosen to be reference pictures much more often than frames with higher QPs, allowing therefore frames with lower distortion to be the reference pictures. In HEVC-PVS-LD2 the last 4 frames that were encoded are the reference pictures regardless of the QP that was used to encode them, which lowers the coding efficiency.

To further analyze the coding efficiency of HEVC-OPT relatively to HEVC-PVS-LD1, the comparison using the Bjøntegaard Delta Metric is shown in Table I. From these results it is possible to conclude that HEVC-OPT achieves an average of 36.60% (up to 49.88% for image I06) bitrate savings over HEVC-PVS-LD1 for the 12 EPFL dataset LF images. Thus, it is possible to conclude that optimized RPS highly increases the coding efficiency over HEVC-PVS-LD1 using the same spiral scanning order. It is expected that for different scanning orders, similar increases in the coding efficiency would be observed. However, this assumption requires further testing with different scanning orders, which is going to done as future work.

The computational complexity is similar for both HEVC-PVS-LD1 and HEVC-OPT as the only additional computation of HEVC-OPT is to solve (1), which is performed only once in the encoding process and once during the decoding process. Moreover, HEVC-OPT can not only be used with any list of scanning orders that are established between the encoder and the decoder but also decode a regular HEVC bitstream.

Although HEVC-OPT achieves very significant bitrate savings when compared to HEVC-PVS-LD1 and consequently the remaining MI-based benchmarks, it also presents some disadvantages in relation to HEVC-PVS-LD1. In terms of memory requirements, the HEVC-PVS-LD1 decoder only requires enough memory for the current SAI that is being decoded and the 4 reference SAIs that are used for the decoding process. In the case of HEVC-OPT, no SAI is discarded from the memory until the full LF image is decoded. Additionally, HEVC-PVS-LD1 does not require any changes

to the HEVC decoder to be implemented. HEVC-OPT, however, it requires an extra processing step to recognize the extra index that signals the type of scanning order and then calculates (1) and performs proposed optimized RPS for the specific scanning order.

IV. Conclusions

This paper proposes a PVS-based LF image coding approach that allows the video codec to be aware of the scanning order that was used to convert the LF image to the PVS format. Since the codec is aware of the original spatial position of each SAI it is able to implicitly optimize the RPS for each processed SAI. Experimental results show that the proposed RPS optimized implementation is able to outperform both PVS- and MI-based benchmarks for every LF image that was tested. On average the proposed RPS optimized codec is able to achieve 36.60% bitrate savings over HEVC-PVS-LD1 using the same scanning order. Although the computational complexity increase is negligible, the proposed RPS optimized codec requires a decoder with enough memory for the full LF image at once. In comparison HEVC-PVS-LD1 only requires enough memory for the current SAI and 4 reference SAIs.

Future work will include testing the proposed optimized RPS with other scanning orders.

Acknowledgment

The authors acknowledge the support of Fundação para a Ciência e Tecnologia, under the grant SFRH/BD/136953/2018 and the project UID/EEA/50008/2019.

References

[1] T. Georgiev and A. Lumsdaine, "Rich Image Capture with Plenoptic Cameras," in *IEEE International Conference on Computational Photography*, Cluj-Napoca, Romania, 2010, pp. 1–8.

[2] J. Arai, "Integral three-dimensional television (FTV Seminar)," Sapporo, Japan, ISO/IEC JTC1/SC29/WG11 MPEG2014/N14552, Jul. 2014.

[3] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, "Advances in three-dimensional integral imaging: sensing, display, and applications," *Appl Opt*, vol. 52, no. 4, pp. 546–560, Feb. 2013.

[4] "JPEG PLENO Abstract and Executive Summary," Sydney, ISO/IEC JTC 1/SC 29/WG1 N6922, Feb. 2015.

[5] "MPEG-I Technical Report on Immersive Media," Torino, Italy, ISO/IEC JTC1/SC29/WG11 N17069, Jul. 2017.

[6] A. Aggoun, "A 3D DCT Compression Algorithm For Omnidirectional Integral Images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, vol. 2, pp. II–II.

[7] M. C. Forman and A. Aggoun, "Quantisation strategies for 3D-DCT-based compression of full parallax 3D images," in *International Conference on Image Processing and Its Applications*, 1997, vol. 1, pp. 32–35.

[8] M. B. de Carvalho *et al.*, "A 4D DCT-Based Lenslet Light Field Codec," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 435–439.

[9] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process. Image Commun.*, vol. 42, pp. 59–78, Mar. 2016.

[10] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *IEEE International Conference on Multimedia Expo Workshops*, 2016, pp. 1–4.

[11] L. F. R. Lucas *et al.*, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *European Signal Processing Conference*, Lisbon, Portugal, 2014, pp. 11–15.

[12] R. J. Monteiro, P. Nunes, N. Rodrigues, and S. M. M. de Faria, "Light Field Image Coding using High Order Intra Block Prediction," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 7, pp. 1120–1131, Oct. 2017.

[13] R. Monteiro *et al.*, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *IEEE International Conference on Multimedia Expo Workshops*, 2016, pp. 1–4.

[14] C. Conti, P. Nunes, and L. D. Soares, "Light field image coding with jointly estimated self-similarity bi-prediction," *Signal Process. Image Commun.*, vol. 60, pp. 144–159, Feb. 2018.

[15] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of Lytro-Illum light fields," in *International Conference on Image Processing Theory, Tools and Applications*, 2015, pp. 494–497.

[16] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *IEEE International Conference on Multimedia Expo Workshops*, 2016, pp. 1–4.

[17] C. Jia *et al.*, "Optimized inter-view prediction based light field image compression with adaptive reconstruction," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4572–4576.

[18] W. Ahmad, R. Olsson, and M. Sjöström, "Interpreting plenoptic images as multi-view sequences for improved compression," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4557–4561.

[19] W. Ahmad, R. Olsson, and M. Sjostrom, "Towards a Generic Compression Solution for Densely and Sparsely Sampled Light Field Data," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 654–658.

[20] J. Hou, J. Chen, and L. Chau, "Light Field Image Compression Based on Bi-Level View Compensation with Rate-Distortion Optimization," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2018.

[21] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 4562–4566.

[22] X. Jiang, M. L. Pendu, and C. Guillemot, "Light field compression using depth image based view synthesis," in *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2017, pp. 19–24.

[23] R. Sjoberg *et al.*, "Overview of HEVC High-Level Syntax and Reference Picture Management," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[24] *Light Field Toolbox v0.4*. http://dgd.vision/Tools/LFToolbox/.

[25] "JPEG Pleno Light Field Coding Common Test Conditions," Germany, ISO /IEC JTC 1/SC 29 /WG 1 M80008, Jul. 2018.