LIGHT FIELD HEVC-BASED IMAGE CODING USING LOCALLY LINEAR EMBEDDING AND SELF-SIMILARITY COMPENSATED PREDICTION

Ricardo Monteiro[1,2], Luís Lucas[1,5], Caroline Conti[1,2], Paulo Nunes[1,2], Nuno Rodrigues[1,3], Sérgio Faria[1,3]
Carla Pagliari[4], Eduardo da Silva[5], Luís Soares[1,2]

[1]Instituto de Telecomunicações; [2]ISCTE, Inst. Univ. de Lisboa; [3]ESTG, Inst. Pol. De Leiria, Portugal;
[4]DEE, Inst. Militar de Engenharia; [5]PEE/COPPE/DEK/Poli, Univ. Federal do Rio de Janeiro, Brazil;
e-mails: ricardo.monteiro,caroline.conti,paulo.nunes,lds@lx.it.pt, luis.lucas,eduardo@smt.ufrj.br, carla@ime.eb.br, nuno.rodrigues,sergio.faria@co.it.pt

ABSTRACT

Light field imaging is a promising new technology that allows the user not only to change the focus and perspective after taking a picture, as well as to generate 3D content, among other applications. However, light field images are characterized by large amounts of data and there is a lack of coding tools to efficiently encode this type of content. Therefore, this paper proposes the addition of two new prediction tools to the HEVC framework, to improve its coding efficiency. The first tool is based on the local linear embedding-based prediction and the second one is based on the self-similarity compensated prediction. Experimental results show improvements over JPEG and HEVC in terms of average bitrate savings of -71.44% and -31.87%, and average PSNR gains of 4.73dB and 0.89dB, respectively.

***Index Terms*—** light field image coding, self-similarity, image prediction, locally linear embedding, HEVC

1. INTRODUCTION

Light field imaging, also known as – holoscopic, integral and plenoptic imaging – is an imaging technology that is able to capture a 4D light field by means of multiplexing the light field data in the camera's 2D conventional sensor resolution [1]. This multiplexing is done through an array of microlens placed between the main lens and the camera sensor. Each microlens creates a micro-image (MI), which is the microlens scene perspective being captured through the main lens. As a result, a light field image tends to be similar to the output of an array of very small cameras.

This image acquisition approach supports new image manipulation features not straightforwardly possible with traditional 2D image acquisition, like refocusing and changing the perspective after a picture has been taken [1]. Additionally, several applications exist that would benefit from these features, *e.g.*, richer image capturing [1], 3D Television [2], image recognition and medical imaging [3].

The authors acknowledge the support of Fundação para a Ciência e Tecnologia, under the project UID/EEA/50008/2013, and the A/BIM/Nº37/2015 grant.

Recognizing the potential of this technology, the JPEG Committee has launched a new standardization activity (the JPEG Pleno [4]), which is targeting both representation and compression of light field, point-cloud and holographic content. New compression tools for light field images are of paramount importance because of the high amounts of data involved (*e.g.*, the sensor resolution of a Lytro Illum light field camera is 40 megapixel [5]). Available image/video coding standards like JPEG [6] and HEVC [7] have sub-optimal performances, since they are not optimized for this kind of content.

Some light field coding schemes described in the literature are based on a discrete cosine transform (DCT) [8,9] or a discrete wavelet transform (DWT) [10]. In [8], a 3D-DCT is applied to a stack of micro-images, to exploit the existing spatial redundancy within each micro-image, as well as the redundancy between adjacent micro-images. In [10], a light field image is decomposed into viewpoint images and a 3D-DWT is applied to a stack of these viewpoint images. The lower frequency bands are transformed using a two-dimensional discrete wavelet transform (2D-DWT), while the remaining high frequency coefficients are simply quantized and arithmetic encoded. These coding schemes are more efficient than JPEG but not as efficient as HEVC still picture coding.

More recently, other authors proposed new prediction tools to HEVC to improve its coding efficiency for light field images. Such tools include locally linear embedding (LLE)-based prediction [11,12] and the self-similarity (SS) compensated prediction [13]. Both methods improve the efficiency of HEVC standard to encode light field images, by exploiting their inherent non-local spatial redundancy.

This paper proposes a light field image coding solution based on HEVC coding architecture combined with the LLE and the SS prediction methods [12,13]. Since both approaches are conceptually different, it is expected that there may be compression ratios, content types and optical setups where one method performs better than the other. By joining both approaches improved coding efficiency is expected.

The remainder of this paper is organized as follows: Section 2 briefly describes the advantages of using HEVC as a basis environment to implement other coding techniques, as

978-1-4799-7082-7/15/$31.00 ©2015 IEEE

well as LLE and SS prediction methods. Section 3 accesses the performance of the proposed solution in comparison with relevant benchmarks and, finally, Section 4 concludes the paper.

2. PROPOSED PREDICTION MODES FOR HEVC

This proposal incorporates the LLE and SS prediction modes in a HEVC codec. The use of HEVC as the basis environment has several advantages:

1. HEVC has very flexible prediction modes and partition patterns. HEVC intra prediction uses planar, DC and 33 directional modes to exploit spatial redundancy. Coding blocks (CB) can have a size between 64×64 and 8×8 pixels. The prediction blocks (PB) can be 2N×2N, 2N×N, N×2N, N×N, 2N×nU, 2N×nD, nL×2N and nU×2N, where N can be 32, 16 and 8.
2. The utilized HEVC test module [14] chooses the best mode for each CB based on a rate-distortion optimization (RDO) criterion. The selected CB is encoded using the mode that achieves the lowest cost. The HEVC framework is therefore flexible to host new coding tools, which can improve the performance of the conventional HEVC tools.
3. The proposed tools can be applied as an extension of the HEVC algorithm.

The following subsections describe how both proposed tools are integrated in HEVC to improve its efficiency for light field image coding.

2.1. Locally Linear Embedding-Based Prediction

LLE is a mathematical tool used to map nonlinear high dimensional data into a low dimensional coordinate system. Its working principle has been used in literature as the basis of a new intra image prediction method [11]. The main idea behind LLE-based prediction method is to estimate the current CB through a linear combination of the k-nearest neighbor (k-NN) patches. To find the linear coefficients, LLE solves a least-squares optimization problem with a constraint on the sum of the coefficients that has to be 1. This estimation procedure is based on a previously coded and reconstructed area of the image, so that it can be performed in both encoder and decoder sides.

When applying LLE as an image prediction method it is necessary to define the search window, a template format and the block that is going to be predicted. Fig. 1 illustrates the mentioned elements. The causal window, $\boldsymbol{W}$, is used for searching the k-NN template patches that present the lowest matching error with template $\boldsymbol{C}$. The k best template patches obtained in the search procedure are then linearly combined to approximate the template $\boldsymbol{C}$, using optimally estimated coefficients. Finally, the block $\boldsymbol{P}$, is predicted using the same linear coefficients estimated for template patches, but used to combine the square blocks associated to each template patch. A formal description is given as follows.

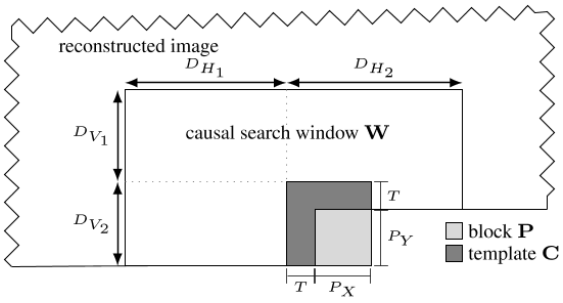

Fig. 1. Search window used by LLE-based prediction method

Consider the region $\boldsymbol{S}$, the N-pixel area that corresponds to both block $\boldsymbol{P}$ (N_p-pixels) that is being predicted, and the known template $\boldsymbol{C}$ (N_c-pixels). A vector $\vec{b}$ can be defined by stacking all the values from region $\boldsymbol{S}$ into a single column (the pixels from the unknown block $\boldsymbol{P}$ are assumed to be zero). Additionally, an $N \times M$ matrix $\boldsymbol{A}$ is defined as the basis dictionary, by stacking all the patches from the search window, similar to region $\boldsymbol{S}$.

Notice that both matrix $\boldsymbol{A}$ and $\vec{b}$ contain two vertically concatenated sub-matrices $\boldsymbol{A_c}$ and $\boldsymbol{A_p}$ and $\vec{b}_c$ and $\vec{b}_p$. These sub-matrices contain the pixel values correspondent to the template $\boldsymbol{C}$ and the block $\boldsymbol{P}$, respectively.

In order to reduce the number of dictionary elements, the k-NN method is used to select the k closest patches to template $\boldsymbol{C}$, in terms of Euclidean distance. These k-NN patches are stacked into a new matrix, defined as $\boldsymbol{A_c^k}$.

Given the matrix $\boldsymbol{A_c^k}$ and the vector $\vec{b}_c$, the LLE-based prediction can be defined through the following optimization problem:

$$\arg\min_{x_k} \left\| \overrightarrow{b_c} - A_c^k \overrightarrow{x_k} \right\|_2^2 \quad subject\ to \quad \sum_m \overrightarrow{x_{k_m}} = 1 \qquad (1)$$

where $\overrightarrow{x_k}$ represents the desired solution of k optimal linear coefficients. Note that, this optimization problem can be solved in both encoder and decoder sides since it only depends on causal information, namely the search window ($\boldsymbol{A_c^k}$) and the template $\boldsymbol{C}$ (or $\vec{b}_c$).

Finally, the block prediction is given by $b_p = \boldsymbol{A_p}\, \vec{x}_{opt}$. For improved performance, the encoder tests different k values, from 1 up to 8, and the one that produces the best block prediction result (b_p), is explicitly transmitted to the decoder.

2.2. Self-Similarity Compensated Prediction

The self-similarity compensated prediction (SS) is an intra prediction tool based on a block matching algorithm, applied to a causal area of the image.

In light field images, the repetitive structure of micro-images has a lot of redundant information, which corresponds to the various points of view from each microlens. By applying a SS-estimation using the same search window W from Fig. 1, the best match between the current CB and an

already coded and reconstructed area of the image, is signaled by a shift vector (referred to as the SS vector).

Additionally to the vector, the SS method uses other signaling techniques based on the inter-prediction tools. Using a picture order count (POC) distance, which in the case of SS is always zero, because it corresponds to the current picture, the reconstructed area from the current image is treated as a reference picture.

Alternatively, an SS-skip mode is defined, in which, similarly to HEVC merge technique [7], instead of transmitting the vector difference explicitly, SS vectors can be derived from neighboring PB. The advantage of SS-skip, relatively to SS-estimation, is the fact that a derived vector can be reused by simply signaling its origin PB using an index. The SS-skip mode, as in [7], fills a candidate list with the available spatially neighboring PBs. Their availability depends mostly on which prediction mode was used to encode the neighboring PB (*e.g.*, if an intra mode was used, SS-skip does not add any candidate to the list).

As explained before, a light field image has very different characteristics from a standard image. Due to the repetitive grid of micro-images, the cross-correlation of a light field image is described by several cyclic peaks. The peaks repeat within a distance of one micro-image, in pixels, both vertically and horizontally [13]. Because of this, the SS vector distribution has a statistical behavior that is characteristic to this kind of content. The most likely chosen vectors are centered on multiples of MI sizes, because the cross-correlation is higher on those points.

With this prior knowledge it was possible to increase the efficiency of SS vector prediction. The SS vectors that were used to encode spatially neighboring PB are used to fill a list of candidates for SS vector prediction, similarly to HEVC advanced motion vector prediction (AMVP). AMVP is used originally by HEVC to allow motion vectors to be transmitted relatively to vectors applied in nearby PB. Since it is very likely that the current PB is going to have similar motion, the vector difference is transmitted, instead of the explicit vector. In the case of SS vectors, this premise also applies. Moreover, three additional candidates are added as part of the MI-based prediction [13], the AMVP and merge candidate list, that are calculated based on the PB and MI sizes. These vectors will point to the left, above and above-left micro-images.

2.3. Integration of proposed prediction modes in HEVC

As mentioned before, HEVC framework is flexible to host new coding tools. Therefore, both prediction methods were added to this framework so they can compete, in terms of RD cost, with the HEVC conventional tools to exploit spatial redundancy. This means that additionally to the HEVC intra modes, LLE and SS are tested as well. The mode with the lowest RD cost is selected to encode the CB.

Whenever new coding tools are added to the HEVC framework, it is necessary to define a way to signal the additional information to the decoder. In order to facilitate this step, for the implementation of LLE, 8 directional modes from HEVC where substituted for LLE to signal the value of k. Since using all the 8 NN patches is not always the most efficient option [12], each k value is explicitly transmitted by using the already available signaling methods for 8 directional modes. The correspondence between the mode that was substituted and the value of k follows the rule: $k = (mode + 1)/4$. Since SS-estimation and SS-skip modes work similarly to HEVC inter modes, only high level syntax was necessary for signaling the necessary information to the decoder. This high level syntax includes a modified I slice that uses the tools from P slices.

The configuration of both techniques took into account a good compromise between efficiency and computational complexity. Most of the computational complexity of the proposed techniques comes from the search algorithms. Both techniques use a full search algorithm with a search range of 128 pixels. More specifically, in Fig. 1 the causal search area used for both LLE and SS corresponds to $D_{H_1} = 128$, $D_{V_1} = 128 - T$ and $DV_2 = T + P_Y$. For LLE, the region length defined for the template C was $T = 4$. Additionally $D_{H_2} = 64$ was used for LLE and $D_{H_2} = 128$ was used for SS. Moreover, each method accounts for a computational complexity similar to what HEVC would need to encode an inter-coded frame using the same search algorithm and search range.

3. EXPERIMENTAL RESULTS

The performance of the proposed LLE-based and SS prediction methods for light field image coding combined with HEVC (referred as HEVC+LLE+SS) was evaluated against JPEG and HEVC standards, as well as SS [13] individually (referred to as HEVC+SS). The reference HEVC software version HM-14.0 was used for the proposed schemes HEVC+LLE+SS and HEVC+SS, and it was also used as a benchmark.

The experimental setup was defined using the ICME 2016 Grand Challenge criteria, the Light Field Image Compression document [15]. This document specifies the usage of the EPFL light field image dataset [16] and how to evaluate the efficiency of the proposed techniques. The raw YCbCr 4:2:0 light field was encoded by all the referred encoders for target compression ratios of 10, 20, 40, and 100, which corresponds to 5185488, 2592744, 1296372, 518549 bytes, respectively. Since no bitrate control algorithm was developed, because it was not the focus of this work, several QPs (quantization parameters) were tested. QPs achieving the closest target compression ratios were chosen for the final results. The reconstructed light field image was then converted to a LF data structure, which is a stack of 2D low-resolution RGB images in addition to a weighting image. The generated LF data structure is then compared to the provided reference LF data structure, using the average PSNR-YUV of all 2D low resolution images. In order to compare the several benchmarks with the proposed HEVC+LLE+SS solution, Bjontegard delta metrics (BJM) [17] were used. The comparative results for the proposed solution, in relation to each of these benchmarks, are presented in Table 1.

As can be observed, the proposed solution is able to outperform all the other tested solutions for every image,

except for I10 encoded with HEVC+SS. As expected, the gains relative to JPEG are very high, with an average of 71.44% bitrate savings and 4.73dB of PSNR gains. Although HEVC+SS solution already presents interesting gains when compared to HEVC, its combination with LLE is able to further increase the RD efficiency, with bitrate savings of up to 42.63% and PSNR gains up to 0.72dB. Both LLE and SS prediction methods complement each other in the HEVC framework, as LLE is based on implicit predictors and SS is based on explicit predictors. This means that for LLE the decoder needs to repeat almost the same procedure of the encoder, with the exception of determining the optimal *k* value, while for SS the decoder applies the prediction vectors computed in the encoder.

Table 1 - BJM results of proposed HEVC+LLE+SS solution for several benchmarks

Benchmark	**JPEG**		**HEVC**		**HEVC+SS**	
BD	**Rate**	**PSNR**	**Rate**	**PSNR**	**Rate**	**PSNR**
I01	-63.97	4.76	-26.98	0.96	-9.56	0.28
I02	-69.78	5.14	-19.5	0.62	-7.25	0.21
I03	-57.93	4.15	-8.16	0.27	-4.39	0.14
I04	-61.96	4.61	-8.20	0.23	-1.01	0.03
I05	-71.46	3.69	-31.59	0.66	-16.72	0.26
I06	-78.63	5.03	-60.15	1.56	-42.63	0.66
I07	-67.11	4.19	-16.93	0.42	-8.84	0.19
I08	-69.89	3.82	-37.77	0.71	-30.15	0.48
I09	-79.49	5.50	-47.13	1.63	-17.52	0.44
I10	-70.09	3.56	-5.60	0.11	0.40	0.03
I11	-86.61	5.87	-68.34	1.86	-40.68	0.72
I12	-80.41	6.42	-52.05	1.60	-24.03	0.46
Average	**-71.44**	**4.73**	**-31.87**	**0.89**	**-16.87**	**0.33**

4. CONCLUSIONS

This paper proposes to improve the coding efficiency of HEVC codec for light field images, by means of enriching its prediction framework with two new tools: LLE and SS.
Experimental results show that by combining these prediction tools, higher rate-distortion gains can be achieved when comparing to various benchmarks, namely, JPEG, HEVC and HEVC+SS. Average bitrate savings of 71.44%, 31.86% and 16.87% are achieved, when comparing to, JPEG, HEVC and HEVC+SS, respectively. Future work will include the study of a unified method that combines the advantages of the presented techniques, exploiting both the implicit and explicit prediction characteristics.

5. REFERENCES

[1] T. Georgiev and A. Lumsdaine, "Rich Image Capture with Plenoptic Cameras," in *IEEE ICCP 2010*, Romania, 2010, pp. 1–8.

[2] J. Arai, "Integral three-dimensional television (FTV Seminar)," Sapporo, 2014.

[3] X. Xiao, et al., "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, 2-6 Feb.

[5] "Lytro Illum." [Online]. Available: https://illum.lytro.com/illum/specs. [Accessed: 21-Mar-2016].

[6] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[7] G. J. Sullivan, et al., "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

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

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

[10] A. Aggoun, "Compression of 3D Integral Images Using 3D Wavelet Transform," *Journal of Display Technology*, vol. 7, no. 11, pp. 586–592, Nov. 2011.

[11] M. Turkan, et al., "Image Prediction Based on Neighbor-Embedding Methods," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1885–1898, Apr. 2012.

[12] L. F. R. Lucas, et al., "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, Lisbon, Portugal, 2014, pp. 11–15.

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

[14] "HEVC Test Module." [Online]. Available: https://hevc.hhi.fraunhofer.de/. [Accessed: 13-May-2016].

[15] "ICME 2016 Grand Challenge: Light-Field Image Compression." [Online]. Available: http://mmspg.epfl.ch/ICME2016GrandChallenge. [Accessed: 02-Apr-2016].

[16] "EPFL Light-field image dataset." [Online]. Available: http://mmspg.epfl.ch/EPFL-light-field-image-dataset. [Accessed: 02-Apr-2016].

[17] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," *VCEG-M33*, Apr. 2001.