LIGHT FIELD IMAGE COMPRESSION BASED ON CONVOLUTIONAL NEURAL NETWORKS AND LINEAR APPROXIMATION

N. Bakir, W. Hamidouche and O. Déforges

Univ Rennes, INSA Rennes,
CNRS, IETR - UMR 6164, France
E-mail: whamidou@insa-rennes.fr

K. Samrouth

LaSTRe, EDST
Lebanese University
Tripoli, Lebanon

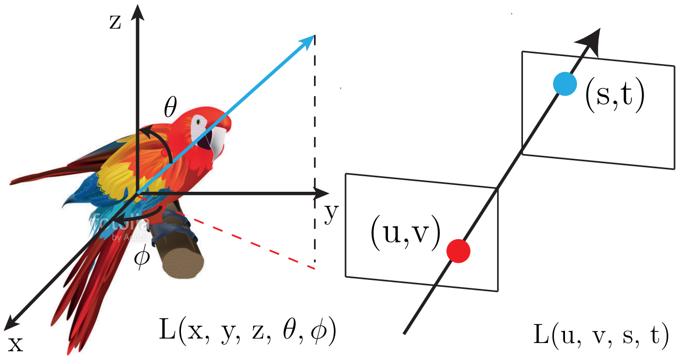
ABSTRACT

Computer vision applications such as refocusing, segmentation and classification become one of the most advanced imaging services. Light Field (LF) imaging systems provide a rich semantic information of the scene. Using a dense set of cameras and microlens arrays (Plenoptic camera), the direction of each ray coming from the scene toward the LF capture system can be extracted and represented by spatial and angular coordinates. However, such imaging system induces many drawbacks including the large amount of data produced and complexity increase for scene representation. In this paper, we propose an efficient LF image coding scheme. This scheme first encodes a sparse set of views using the latest hybrid video encoder (JEM). Then, it estimates a second sparse set of views using a linear approximation. At the decoder side, we use a Deep Learning (DL) approach to estimate the whole LF image from the reconstructed sparse sets of views. Experimental results show that the proposed scheme provides higher visual quality and overcomes the state of the art LF image compression solution by 30 % bitrate gain.

Index Terms— Light Field, Machine Learning, Linear approximation, CNN, future video coding

1. INTRODUCTION

Recently, Light Field (LF) imaging is becoming one of the most trending multimedia technology. It consists in representing the scene from different point of views. Mathematically, a plenoptic image can be described by a 7 dimensional function $L(\lambda,\, t,\, x,\, y,\, z,\, \theta, \phi)$ assigning to every point in free space and to every direction a corresponding radiance for specific wavelength λ and time. For static scenes, the dimension of this function is reduced to 5 dimensions without considering time and wavelength. On the other hand, the plenoptic function can be represented by 4 dimensions with 2 parallel plans s, t and u, v denoted by $L(u,\, v,\, s,\, t)$ [1, 2] as illustrated in **Fig. 1**. Raytrix [3] and Lytro [4] companies provide commercially LF cameras with array of microlens (plenoptic camera). Different from conventional cameras, a plenoptic camera can capture multiple views of a scene from a single shoot. There are several representations of the LF image. For instance, there are the micro-image, sub-aperture and epipolar image representations [5]. For plenoptic images, the baseline between two sub-aperture views is very narrow. The disparity range between the adjacent sub-aperture views is smaller than 1 pixel [6].



(a) 5D-LF representation

(b) 4D-LF representation

Fig. 1: The spatial parametrization of 5D-LF and 4D-LF representations, the s, t plane is closer to the camera, and the u, v plane is closer to the scene [5]

Several studies [2, 7, 8, 9, 10, 11, 12] have investigated the compression of LF images in its different representations. The first works used vector quantization and LempelZiv entropy coding [13] tools to remove statistical redundancies within the LF data [2]. Subsequently, the pseudo sequence coding approach [7], [9] consists in rearranging LF elements (usually sub-aperture images) as a pseudo sequence in a specific order (ie. spiral, horizontal, zigzag, raster scan order). This pseudo-sequence is then coded using classical hybrid (Intra and Inter predictions) video encoders. The High Efficiency Video Coding (HEVC) standard through its reference software model HEVC reference Model (HM) and Joint Exploration Model (JEM) codec developed for the future video coding standard [14] are used in [7, 9] and [15], respectively. The Joint Video Exploration Team (JVET) has been recently investigating several new coding solutions [16] to show the evidence of developing a new standard with coding capability beyond HEVC. These new tools enable to increase the coding efficiency by up to 30% with respect to HEVC [14]. Zhu et al. [7] propose a new method that reduces the number of reference frames by dividing the 4D LF sub-aperture image into 4 quadrants which are separately encoded. The inter dependency is then built to minimize the distance between reference views.

The predictive-coding approach selects a set of views from the array of sub-aperture views or Elemental Images (EI) to be encoded as reference for coding the remaining no-reference LF views. X. Jiang et al. [10] propose an approximation method called Homography Low Rank Approximation (HLRA) based joint optimization of multiple homographies and low rank approximation. HLRA method uses homography in order to reduce the error produced by low rank approximation between the sub-aperture views. Le Pendu et al. [11] propose a coding method that selects a sparse set of sub-aperture views (4 corners) in order to synthesize the whole LF image using the depth image and low rank matrix completion. Olsson et al. [17] propose to encode a lenslet image in HEVC intra-prediction by introducing inter-prediction within the LF image. Shengyang et al. [18] divide LF image into 2 sets of views where the first set of reference views (S_A) are encoded with HEVC and the second set of views (S_B) are estimated from the first set by Linear approximation (LA). Supervised learning [12], [19] with Convolutional Neural Networks (CNN) has been widely adopted in computer vision applications. A super-resolution algorithm for LF based on CNNs, which are used to reconstruct all sub-aperture views at higher resolution (increasing the spatial resolution) has been presented in [20].

In this paper we propose a compression scheme based on a joint LA with CNN to estimate the LF views. The first set of selected reference views (S_R) are coded with the JEM video codec. Then, a second set of views (S_E) are linearly estimated from the first decoded set of views. The bitstream of the first reference set with the coefficients of LA are transmitted to the decoder. The decoder decodes the reference set of views, estimates the second set with the received coefficients and then synthesizes the third set of views with the trained CNN.

The remainder of this paper is organized as follows. Section 2 describes two efficient solutions for LF views coding and estimation. Then, in Section 3 we investigate the proposed LF compression solution. Section 4 presents and discusses the experimental results. Finally, Section 5 concludes this paper.

2. RELATED WORKS

As mentioned in Section 1, the proposed compression solution is based on LA and Deep Learning (DL). In this section, we briefly introduce these two concepts.

2.1. Linear Approximation

In [18], *Shengyang et al.* propose a powerful LF coding scheme. The distance between adjacent cameras is a constant scalar. Mathematically, the LF image is modelled by a 4D function:

$$L : \Omega \times \Pi\{\Longrightarrow \mathbb{R}\}, (\{\rho, \varphi\}) = \mathrm{L}(\{\rho, \varphi\}), \{\rho \in \Omega\} \quad (1)$$

where ρ is a scene point, Ω represents the image plane and $\varphi :=$ (u,v)^T denotes the offset of one view w.r.t. the center view in lens plane. As shown in **Fig. 2**, this scheme consists in coding a sparse set of LF views (S_A) using HEVC and then linearly approximating the other views (S_B) and sending only the approximation coefficients to the decoder. The LA prior of the dropped vectorized view j is given as follows:

$$V_j \approx \frac{1}{\Sigma x_m} \sum_{m \neq j}^{M} x_m V_m, \quad 2 \leq M \leq N \quad (2)$$

where M is the number of selected views and N is the total view number, 1 $\leq$ m $\leq$ M and x_m are the weight coefficients. This coding scheme enables between 37.41% and 45.51% Bjøntegaard Delta Bit Rate (BD-BR) reduction on average compared to the HEVC encoding all views (HM-All). This gain is achieved when half of views are encoded transmitted to the decoder and other half of views are linearly approximated.

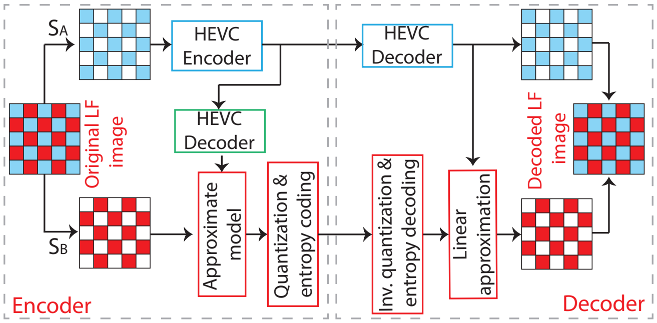


Fig. 2: Linear Approximation Coding Scheme [18]

2.2. Deep Learning

In [21], authors proposed a learning-based approach to synthesize new views from a sparse set of input views. The LF synthesis scheme is composed of disparity and color estimation components (**Fig. 3**). Authors use two sequential CNNs to model these two components and train both networks simultaneously by minimizing the error between the synthesized and ground truth images. They used only four corner sub-aperture views from the LF captured by the Lytro Illum camera to synthesize high-quality images that are superior to the state-of-the-art techniques. As shown in **Fig. 3**, a set of features (mean and standard deviation) of a sparse set of views are fed to the first CNN that estimates the disparity at an intermediate view using Equation 3.

$$D_q = g_d(K), \quad (3)$$

This equation models how the estimated disparity D_q at the novel view at position q is generated from the set of features K including the mean and standard deviation. Finally, the second CNN generates the final intermediate view using Equation 4.

$$F_q = g_c(H), \quad (4)$$

where F$_q$ represents the image at the intermediate view, H the feature set and g_c defines the relationship between these features and the final intermediate image.

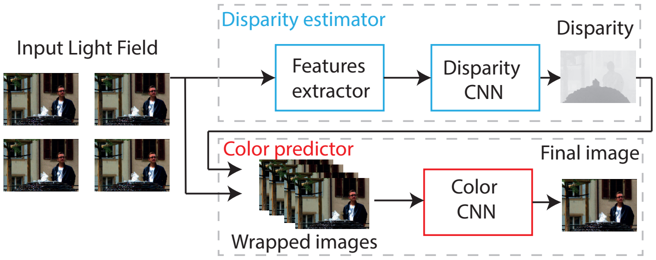


Fig. 3: Deep Learning Views Synthesis [21]

3. PROPOSED COMPRESSION SOLUTION

Kalantari et al. have not taken into consideration the bitrate criterion and *Zhao et al.* have shown a high visual quality for reconstructed LF images when a large number of views (half of views) are used as reference. Therefore, we propose in this paper to smoothly merge these two concepts in order to build an efficient LF compression solution

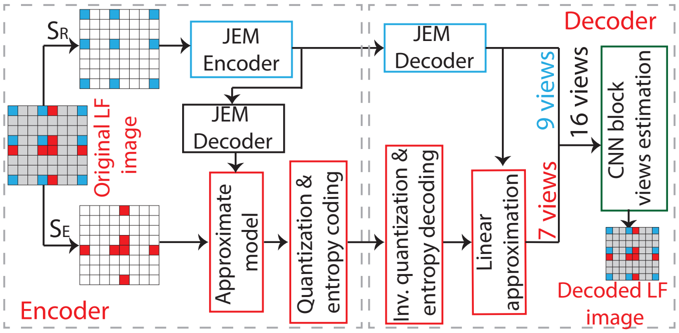


Fig. 4: CNN and LA based coding scheme

for a better rate-distortion optimization. In the following section, we explain in details the encoder and decoder of the proposed coding scheme.

3.1. Coding Scheme

In the proposed scheme, we consider the sub-aperture based representation of LF image of 8 × 8 views. It consists in dividing the plenoptic image into four Groups Of Views (GOV) (4 × 4 images each). For each GOV, we take the 4 corners as reference in order to synthesize the novels views. In total, the number of references views is 16 for the whole LF image. As first step, we select a sparse set of sub-aperture views (S_R in blue) with specific position that give the best result after testing all possible combinations. Then, we rearrange the nine S_R views into a pseudo sequence (zigzag order scan) and encode it with a simple JEM encoder with chrominance downscale, e.g yuv 420. In the second step, we estimate the 7 adjacent views set (S_E in red) using linear approximation explained in Section 2.1.
For each frame in the dropped views set S_E, we linearly approximate the views with the decoded views in S_R set. An approximation model is used to optimize the reconstruction of the weight coefficients X, by using the Spectral Projected Gradient for L1 (SPGL1) function. This one generates the coefficients for one target view at each time and for each channel color separately (i.e. rgb, 3 channels).
As this vector X contains floating point values, we quantize X at 16 bits before encoding it with entropy coding. The JEM bitstream encoding the S_R set of views with the quantized and entropy coded linear coefficients are sent to the decoder.

3.2. Decoding Scheme

At the decoder side, the JEM bitstream is first decoded to reconstruct the reference views of the set S_R. Based on these views and by using the disparity correlation and linear approximation between the views, we search the optimal vector of coefficient X. Then, from the decoded vector X and decoded views S_R, we estimate the other sparse set of views S_E. The two decoded sets S_E and S_R together form the 16 reference views used to feed the DL block in order to synthesize the remaining 48 views.

3.3. CNN Training Phase

For training the CNN, we run the training of DL that uses the disparity and color estimation components in two sequential CNNs. These CNNs are used to synthesize the novel views for each GOV separately with 7 layers (4 convolutions and 3 ReLUs), angular resolution 4 × 4 and the numerical evaluation and the final image has index (2,2). We take the 4 corner source views as input. For this training, we used 100 images, 28 from Stanford Lytro LF archive [22], and 72 from California Lytro LF archive [21] captured by Lytro camera.

4. RESULTS AND DISCUSSION

4.1. Experimental setup

The proposed scheme described in the previous section (DL-LA-9-7) encodes 9 views with JEM and linearly estimates 7 views to construct the 16 views used as an input of the trained CNN block. We compare the proposed scheme to four other solutions (we also implemented): JEM-All that encodes all views with the JEM software in Random Access (RA) coding configuration, HM-All that encodes all views with the HM [23] (HEVC) reference software in RA configuration, LA-32 solution which encodes half of views with JEM and other half views are linearly approximated at the decoder [18] and DL-16 scheme that encodes 16 views with the JEM and estimates the rest of views at the decoder by the trained CNN block with the 16 JEM decoded views as input. Nine LF images illustrated in **Fig. 5** have been selected from Ecole Polytechnique Federale de Lausanne (EPLF) LF images dataset [24] composed of 8x8 views.
The BD-BR [25, 26] is a Peak Signal to Noise Ratio (PSNR) based metric. It is used in this paper to assess the gain of the solutions compared to the anchor solution. A negative BD-BR value refers to a bitrate reduction compared to the anchor while a positive value expresses a bitrate overhead.

Fig. 5: Thumbnails of the considered nine LF images from EPFL data set [24] 1) *ParcDuLuxembourg* 2) *FountainVincent2* 3) *Friends1* 4) *University* 5) *Pillars* 6) *Friends4* 7) *Yan&KriosStanding* 8) *RustyFence* 9) *Stairs*

4.2. Results

Table 1 gives the BD-BR performance of four solutions JEM-All, LA-32, DL-16, DL-LA-9-7 with respect to the anchor encoding the 64 views with the HM reference software (HM-All). We can notice from this table that the proposed scheme DL-LA-9-7 enables in average the highest coding performance of 51.16 % bitrate reduction followed by DL-16 solution with 48.38 % and then LA-32 and JEM-All with 32.74 % and 12.51 % bitrate reductions, respectively. The last column of **Table 1** gives the BD-BR performance of the

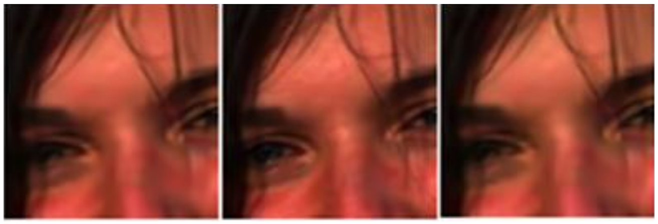


Fig. 6: Visual quality of *Friends1* LF image, view (2,2) in the array sub-aperture 8 × 8, QP= 32. From left to right: encoded, estimated and synthesized views respectively with JEM-All (bitrate 7.11×10^{-4} bits per pixel (bpp), WPSNR = 37 dB), LA-32 (5.03×10^{-4} bpp, 36.51 dB), DL-LA-9-7 (3.06×10^{-4} bpp, 36.3 dB)

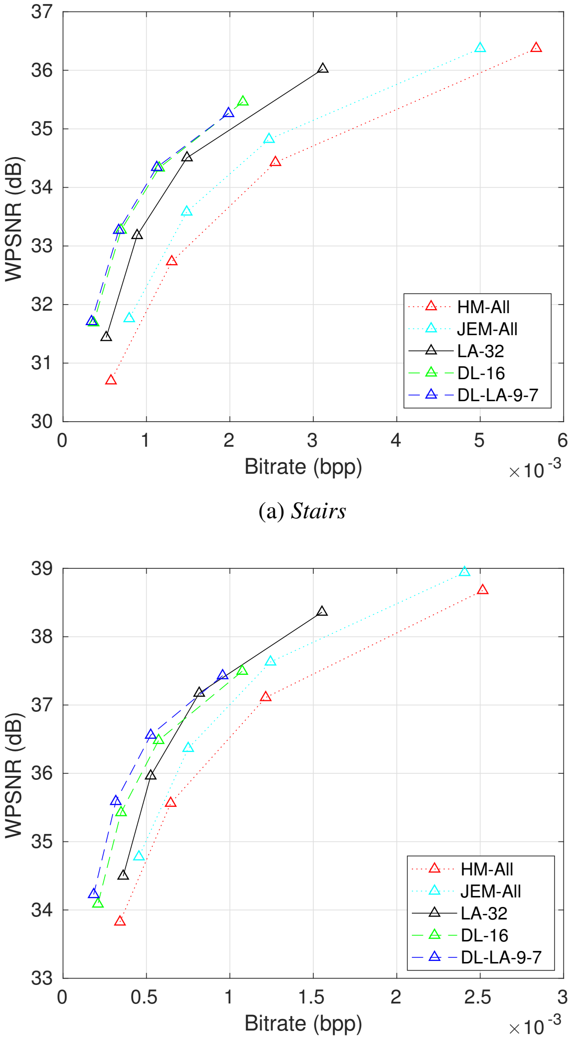
proposed coding scheme with respect to the LA-32 solution. The proposed scheme outperforms the state of the art solution by 30 % in average with achieving a significant coding gain for all considered LF images. It should be noted that the proposed solution shows an inconsistent performance for 4) *University* image. This is mainly caused by the interpolation interval of the BD-BR metric which considers wider rate interval in the case of comparison with the HM-All scheme than with the LA-32 scheme.

Fig. 6 gives a visual illustration of a zoom on the *Friends1* estimated LF view decoded with the JEM-All, LA-32 and DL-LA-9-7 solutions. We can notice that the three images look similar while the bitrate is highly decreased by the proposed DL-LA-9-7 solution. This one enhances the compression ratio by factors of 2.3 and 1.64 compared to JEM-All and LA-32 compression solutions, respectively.

Fig. 7 shows the weighted PSNR (WPSNR) performance of the five solutions versus the bitrate for two LF images *Stairs* and *Friends4*. We can notice that the proposed scheme enables the highest PSNR performance at all considered rates for both images.

Table 1: Coding gains in terms of BD-BR of four solutions in comparison with HM-All anchor. 1) *ParcDuLuxembourg* 2) *FountainVincent2* 3) *Friends1* 4) *University* 5) *Pillars* 6) *Friends4* 7) *Yan&KriosStanding* 8) *RustyFence* 9) *Stairs*

Im.	BD-BR versus HM-All				vs. LA-32
	JEM-All	LA-32	DL-16	DL-LA-9-7	DL-LA-9-7
1)	-16.60 %	-31.39 %	-39.00 %	**-42.68 %**	-15.96 %
2)	-19.28 %	-34.57 %	-53.55 %	**-55.40 %**	-31.79 %
3)	-10.80 %	-22.54 %	-39.42 %	**-47.40 %**	-31.63 %
4)	-20.03 %	**-61.21** %	-59.67 %	-44.52 %	-28.33 %
5)	-4.56 %	-23.73 %	-47.36 %	**-52.07 %**	-40.54 %
6)	-9.82 %	-23.10 %	-35.30 %	**-44.20 %**	-26.48 %
7)	-0.29 %	-20.27 %	-58.83 %	**-65.25 %**	-52.75 %
8)	-15.19 %	-36.92 %	-47.84 %	**-52.52 %**	-24.39 %
9)	-16.10 %	-40.96 %	-54.52 %	**-56.48 %**	-25.15 %
Av.	-12.51%	-32.74 %	-48.38 %	**-51.16%**	- 30.77 %



(a) *Stairs*

(b) *Friends4*

Fig. 7: wPSNR-based comparison of the five considered solutions

5. CONCLUSION

We have proposed in this paper a coding solution based on CNN and linear approximation for 4D-LF images. By selecting a sparse set of views, we fully exploit the correlation between adjacent views to synthesize the novel views and reduce the number of reference views. Experimental results showed that the proposed solution reduces by 51% the bit-rate compared to the HM encoder and increases the visual quality of the novel views. Moreover, the proposed solution enables a bitrate reduction of 30% in average compared to the state of the art LF image compression solution (LA-32).

As future works, we have identified several research ideas to further enhance the coding performance of the proposed scheme with using more advanced DL systems and enhancing the CNN parameters with reinforcement learning. Moreover, a complexity evaluation of the proposed solution will be investigated at both encoder and decoder sides. Finally, the optimal coding of residuals at the encoder side will be investigated to further increase the quality of the estimated views.

6. REFERENCES

[1] Edward H. Adelson and James R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. 1991, pp. 3–20, MIT Press.

[2] Marc Levoy and Pat Hanrahan, "Light field rendering," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 1996, SIGGRAPH '96, pp. 31–42, ACM.

[3] Christian Perwa and Lennart Wietzke, "Single lens 3d-camera with extended depth-of-field," in *Human Vision and Electronic Imaging XVII, 829108*, 2012.

[4] Georgiev Todor, Yu Zhan, Lumsdaine Andrew, and Sergio Goma, "Lytro camera technology: theory, algorithms, performance analysis," 2013.

[5] Donald Gilbert Dansereau, "Plenoptic signal processing for robust vision in field robotics," *Ph.D. dissertation, University of Sydney Graduate School of Engineering and IT School of Aerospace, Mechanical and Mechatronic Engineering*, 2014.

[6] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, Washington, DC, USA, 2013, ICCV '13, pp. 673–680, IEEE Computer Society.

[7] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo sequence based 2-d hierarchical coding structure for light-field image compression," in *2017 Data Compression Conference (DCC)*, April 2017, pp. 131–140.

[8] D. Liu, L. Wang, L. Li, Zhiwei Xiong, Feng Wu, and Wenjun Zeng, "Pseudo-sequence-based light field image compression," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

[9] Waqas Ahmad, Roger Olsson, and Mrten Sjstrm, "Interpreting Plenoptic Images as Multi-View Sequences for Improved Compression," in *DIVA*, 2017.

[10] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1132–1145, Oct 2017.

[11] Xiaoran Jiang, Mikal Le Pendu, and Christine Guillemot, "Light field compression using depth image based view synthesis," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017.

[12] M. Gupta, A. Jauhari, K. Kulkarni, S. Jayasuriya, A. Molnar, and P. Turaga, "Compressive light field reconstructions using deep learning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1277–1286.

[13] Paul Lalonde and Alain Fournier, "Interactive rendering of wavelet projected light fields," in *Proceedings of the 1999 Conference on Graphics Interface '99*, San Francisco, CA, USA, 1999, pp. 107–114, Morgan Kaufmann Publishers Inc.

[14] Naty Sidaty, Wassim Hamidouche, Olivier Deforges, and Pierrick Philippe, "Compression Efficiency of the Emerging Video Coding Tools," in *IEEE Conference on Image Processing (ICIP)*, September 2017.

[15] S. Zhao, Z. Chen, K. Yang, and H. Huang, "Light field image coding with hybrid scan order," in *2016 Visual Communications and Image Processing (VCIP)*, Nov 2016, pp. 1–4.

[16] J. Chen, E. Alshina, G. J. Sullivan, J. R. Ohm, and J. Boyce, "Algorithm description of joint exploration model 7," in *JVET-G1001*, July 2017.

[17] Y. Li, M. Sjstrm, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 539–543.

[18] Zhao Shengyang and Chen Zhibo, "Light field image coding via linear approximation prior," in *IEEE International Conference on Image Processing, China*, January 2017.

[19] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, June 2017.

[20] R. A. Farrugia, C. Galea, and C. Guillemot, "Super resolution of light field images using linear subspace projection of patch-volumes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1058–1071, Oct 2017.

[21] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 193:1–193:10, Nov. 2016.

[22] Shah Raj, Lowney Michael, and Abhilash Sunder, "Stanford lytro light field archive," in *http://lightfields.stanford.edu/*, 2016.

[23] "Joint Collaborative Team on Video Coding Reference Software, ver. HM," in *https://hevc.hhi.fraunhofer.de/*.

[24] Martin Rerabek and Touradj Ebrahimi, "New light field image dataset," in *https://mmspg.epfl.ch/EPFL-light-field-image-dataset*, 2016.

[25] Gi Bjøntegaard, "Calcuation of average psnr differences between rd-curves," *VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April*, 2001.

[26] Gisle Bjøntegaard, "Improvements of the bd-psnr model," *ITU-T SG16 Q*, vol. 6, pp. 35, 2008.