LIGHT FIELD IMAGE CODING VIA LINEAR APPROXIMATION PRIOR

Shengyang Zhao, Zhibo Chen

CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China, Hefei 230027, China

ABSTRACT

In recent years, the light field (LF) image as a new imaging modality has attracted much interest. While light field camera records both the luminance and direction of the rays in a scene, large amount of data makes it a great challenge for storage and transmission. Thus an adequate compression scheme is desired. In this paper, we propose a new prior, called linear approximation prior that reveals intrinsic property among the LF sub-views. It indicates that we can approximate a certain view with a weighted sum of other views. By fully exploiting this prior we propose a powerful coding scheme. The experiments show the superior performance of our scheme, which achieves as large as 45.51% BD-rate reduction and 37.41% BD-rate reduction on average compared with the High Efficiency Video Coding (HEVC).

Index Terms— Linear approximation prior, LF image coding, view reconstruction, convex optimization

1. INTRODUCTION

The light ray can be represented with two pairs of coordinates of two parallel plates which it passes through. A light field (LF) camera records the amount of the light at every point in space, in every direction. Two major types of light field cameras now exist, i.e. the cameras array [1] and lenslet-based cameras like Lytro [2] and Raytrix [3] which use micro-lenses to capture individual ray of light.

As a new imaging modality, LF images provide new exciting imaging functionalities, e.g. digital refocusing and viewpoint changing. Depth or other geometric information derived from LF images can be extremely beneficial in image processing like segmentation and salient detection[4].

A typical LF image is shown in Fig.1. By extracting pixels in the same position in each macropixels, one can decompose the LF image into an array of views. Obviously, a LF image contains hundreds of times pixels more than a conventional image with the same spatial resolution. Therefore, compressing LF images efficiently becomes an imperative challenge.

Recently there is a wide interest in studying the compressibility of the captured light-field, e.g. a new study action known as JPEG PLENO[5] was started by the JPEG group.

Send correspondence to Zhibo Chen<chenzhibo@ustc.edu.cn>

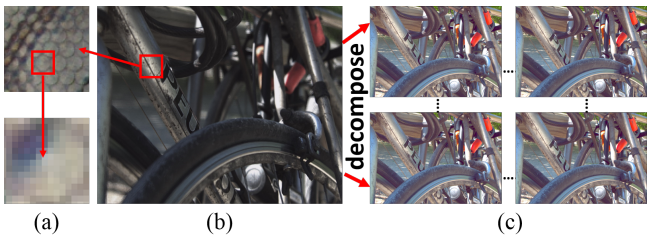

Fig. 1. (a) Macropixels consists of multiple pixels; (b) A LF image in lenslet image format; (c) Decomposed LF image.

An adequate coding scheme should remove redundancy both in spatial domain and angular domain. Since the views share almost the same content, can we compress only a part of the views in the encoder, and reconstruct other views in the decoder? By introducing the image warping technique [6], we explore the correlation within LF views in this paper. We point out the linearity within the angular domain and we call this the linear approximation prior. We believe that the prior reveals intrinsic property of LF data and the experiments verify that it is reliable.

We first divide the views into two subsets S_A and S_B. The set A is compressed and the set B is synthesized with our prior. We formulate the reconstruction as a global optimization problem. Both the bit stream of set A and reconstruction coefficients of set B are transmitted. Experiments show that our coding scheme produces state-of-the-art compression results in terms of rate-distortion performance compared with High Efficiency Video Coding (HEVC): with 37.41% BD-rate[7] reduction and 1.10 dB BD-PSNR improvement achieved on average.

We introduce the related work in section 2. In section 3, we derive the linear approximation prior and propose our coding scheme. Experiments are carried out to evaluate our coding scheme in section 4. Finally we conclude the paper in section 5.

2. RELATED WORK

Major LF compression methods can be divided into 3 categories: pseudo sequence based methods, sparse coding

based methods and lenslet image based methods. The first category[8][9][10] decomposes the LF data into an array of views, as shown in Fig.1(c), and rearranges them into a pseudo sequence. Conventional video coding technologies are then applied to exploit intra-view and inter-view redundancy in spatial domain and angular domain respectively. Discrete wavelet transform [11] and disparity compensated lifting [12] all have been utilized for LF image compression. However these work all ignore the 4D structure of LF data and do not explore the relation among LF views. Notice that Multi-view coding (MVC) is related to this problem, but existing MVC system do not handle hundreds of views in a LF image.

The sparse coding based methods are motivated by dictionary-based dense light field acquisition technique [13] [14], which can capture LF images from a sparse camera array. [15] proposes a disparity-guided method, in which some key views are selected and one dictionary is trained to recover the whole LF image. Although their algorithm shows amazing performance, it is not robust enough, and the dictionary training phases both in encoder and decoder bring high complexity. Jiang et al. [16] exploit the inter-view correlation with homograph transformation and significant gains are achieved. Our work is more related to this compression category but we do not have any training phases, which reduces the algorithm complexity.

The lenslet image based methods compress the lenslet image (like Fig.1(b)) directly. Li et al. [17] use a sparse image set and its associated disparities for predicting the macropixels. Their codec produces impressive results but it is limited for sparsely sampled LF data, which is not practical for LF applications. Monteiro et al. [18] exploits the linearity of macropixels with locally linear embedding. But their algorithm actually assumes linearity in spatial domain (i.e. within in each sub-view), which is only satisfied in homogeneous regions in one image, while our prior mathematically points out the linearity in angular domain (i.e. among different views) which should be a general rule for LF images.

3. PROPOSED METHOD

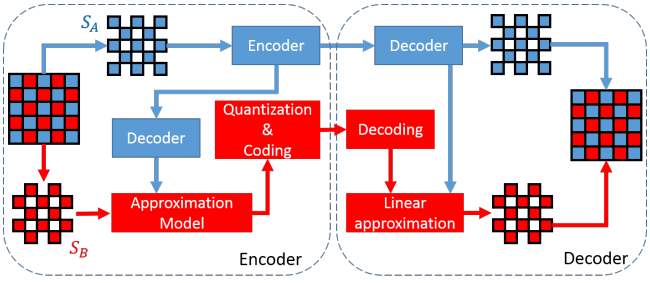

Fig. 2. The proposed linear approximation prior based (LAP) coding scheme.

In this section, we first give a brief overview of our proposed linear approximation prior based (LAP) coding scheme. Then we focus on the view approximation model and formulate it as an optimization problem.

3.1. Overview of LAP scheme

As shown in Fig.2, first the encoder divides views into two sets, the selected views set S_A (the blue ones) and the dropped views set S_B (the red ones). The selected views are then rearranged into a pseudo sequence and a video encoder is used to compress the sequence.

For each dropped view in set S_B, we linearly approximate it with the decoded views in set S_A. An approximation model is used to optimize the reconstruction coefficients, which will be described in details later. Subsequently we quantize and compress the coefficients as an image in JPEG format. Both the compressed sequence and coefficients will be transmitted to the decoder.

The decoder works in an inverse order. The selected views will be decoded with a decoder and the dropped views are approximated with the weighted sum of the selected views. Finally the whole LF image is recovered.

Notice that we here simply take a heuristic way to select views as illustrated in Fig.2, i.e. half of the views are dropped.

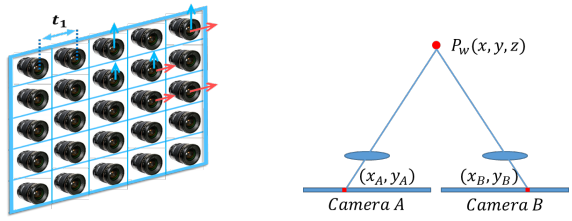

Fig. 3. (a) We treat the lenslet LF camera as a virtual camera array. (b) Stereo view condition.

3.2. View approximation model

To obtain the relation among LF views, we first prove the linear disparity property and derive our new prior for LF image. And then we propose our linear approximation model.

3.2.1. Linear disparity property

The linear disparity property tells that the disparity value of the same scene point between adjacent views should be the same. Mathematically, the light field is a 4D function:

$$L : \Omega \times \Pi \to \mathbb{R},\ (\boldsymbol{p}, \boldsymbol{\varphi}) \mapsto L(\boldsymbol{p}, \boldsymbol{\varphi}) \qquad (1)$$

where $\boldsymbol{p} := (x, y)^T$ denotes a point in the image plane $\Omega := \{(x, y) \in \mathbb{N}^2 | x < n, y < m\}$ and $\boldsymbol{\varphi} := (u, v)^T$ denotes the offset of one view w.r.t. the center view in lens plane.

Imagine one pinhole camera (like Fig.3) and a scene point $\boldsymbol{P}_w = (x, y, z)^T$ in world coordinate system. Let $\tilde{\boldsymbol{P}}_w$ denote the homogeneous coordinate vector of $\boldsymbol{P_w}$. We adopt $\tilde{\boldsymbol{p}}_A = (x_A, y_A, 1)^T$ for the projection point of $\boldsymbol{P}_w$ in the image plane and $\tilde{\boldsymbol{p}}_B$ in a similar fashion. We then get the perspective projection equation:

$$\tilde{\boldsymbol{p}}_A = \frac{1}{z_A} K_A [R_A \;\; \boldsymbol{t}_A] \tilde{\boldsymbol{P}}_w \tag{2}$$

where the K denotes the camera intrinsic matrix. z_A is the depth of $\boldsymbol{P}_w$ from camera A. The change of coordinate between image plane system and world reference frame is represented by rotation matrix R_A and translation vector $\boldsymbol{t}_A$.

We take similar notations in [19], the world reference frame is attached to the camera A and it shares the same orientation with camera B. Thus the rotation matrix between A and B is an identity matrix. We assume that they have the same intrinsic parameters K, the disparity between A and B is:

$$d(B, A) = \frac{1}{z_A} K \boldsymbol{t}_{BA} \tag{3}$$

We now treat the LF camera as a virtual camera array, in which all the cameras are arranged in a square grid. As shown in Fig.3, red and blue arrows represent relative perspective disparity in the horizontal and vertical directions for each view. The distance t_1 between adjacent camera is a constant scalar. We denote with $D(\boldsymbol{p}, \boldsymbol{\varphi}_i) : \Omega \times \Pi \to \mathbb{R}$ a function which defines for each scene point, captured at the position $\boldsymbol{p} \in \Omega$ in the center view V_0, the corresponding point's disparity in view V_i:

$$D(\boldsymbol{p}, \boldsymbol{\varphi}_i) = \frac{1}{z_p} K t_1 \boldsymbol{\varphi_i} \tag{4}$$

As the Eqn.4 shows, for a certain $\boldsymbol{p} \in \Omega$, the disparity is proportional to $\boldsymbol{\varphi}_i$.

3.2.2. Linear approximation prior

We introduce image warping technique [6] to explore the relation between LF views. First we warp the V_i to center view, i.e. we represent V_i with the center view and the corresponding disparity. According to Eqn.4, we define the unit disparity D_u:

$$D_u(\boldsymbol{p}) = \frac{1}{z_p} K t_1 \tag{5}$$

so we rewrite the view V_i as follows:

$$V_i(\boldsymbol{p}) = V_0(\boldsymbol{p} - D_u(\boldsymbol{p}) \boldsymbol{\varphi_i})_{\boldsymbol{p} \in \Omega} \tag{6}$$

The Eqn.6 works only when two images share a certain amount of overlapping, which is obviously met for LF images. In order to represent V_i with V_j, we use the first-order Taylor approximation for Eqn.6 at $\boldsymbol{\varphi}_j$:

$$\begin{aligned} V_i(\boldsymbol{p}) \approx\; & V_0(\boldsymbol{p} - D_u(\boldsymbol{p}) \boldsymbol{\varphi}_j) - \\ & D_u(\boldsymbol{p}) \left\| \boldsymbol{\varphi}_{ij} \right\| \nabla_{\frac{\boldsymbol{\varphi}_{ij}}{\|\boldsymbol{\varphi}_{ij}\|}} V_0(\boldsymbol{p} - D_u(\boldsymbol{p}) \boldsymbol{\varphi}_j) \end{aligned} \tag{7}$$

where $\boldsymbol{\varphi}_{ij} = \boldsymbol{\varphi}_i - \boldsymbol{\varphi}_j$ and $\nabla_{\boldsymbol{\varphi}}$ denotes the directional derivatives of V_j with direction $\boldsymbol{\varphi}$. Taking the weighted sum of V_m, $1 \le m \le N$ and N is the total view number. We then have:

$$\begin{aligned} \sum_{m \neq j}^{M} x_m V_m \approx\; & V_j \sum x_m - \\ & D_u \sum x_m \|\boldsymbol{\varphi}_{mj}\| \nabla_{\frac{\boldsymbol{\varphi}_{mj}}{\|\boldsymbol{\varphi}_{mj}\|}} V_j \end{aligned} \tag{8}$$

If the weight coefficients meet:

$$\sum_{m \neq j}^{M} x_m \boldsymbol{\varphi}_{mj} = \mathbf{0}, [x_1, x_2, ..., x_m] \neq \mathbf{0} \tag{9}$$

then we can remove the first-order term, i.e

$$V_j \approx \frac{1}{\sum x_m} \sum_{m \neq j}^{M} x_m V_m \quad 2 \le M \le N \tag{10}$$

where M is the number of selected views and N is the total view number. The Eqn.10 shows the linearity among LF views. We call this **the linear approximation prior**. Following this we can approximate one view with acceptable quality, which is demonstrated by our experiment.

3.2.3. Optimization model

The Eqn.10 shows that we can approximate one view linearly when the coefficients meet Eqn.9. But there are two **problems**. First, when M is larger than 3, the solution of Eqn.9 is not unique. Second, the video encoder will introduce distortion to the selected views in S_A. Therefore we use a convex model instead Eqn.9 to obtain the optimal coefficients:

$$\begin{aligned} & min \left\| \boldsymbol{x} \right\|_1 \\ & s.t. \left\| A\boldsymbol{x} - \boldsymbol{b} \right\|_2 \le \varepsilon \end{aligned} \tag{11}$$

The l_1 norm guarantees the sparsity of coefficient vector $\boldsymbol{x} \in \mathbb{R}^M$. $A \in \mathbb{R}^{mn \times M}$ consists of all vectorized views in S_A and $\boldsymbol{b}$ is the target view to be recovered. Greedy Orthogonal Matching Pursuit (OMP) algorithm is applied to solve the problem (11) and we optimize $\boldsymbol{x}$ for each view that needs recovering.

4. EXPERIMENT RESULTS

In this section we carry out experiments to evaluate the proposed coding scheme. As the grand challenge requires, we select 6 lenslet light field images from EPFL dataset[20], which are all captured by the Lytro camera. The Fig.4 below shows all the thumbnails of selected images. Each image is decoded with Matlab Toolbox [21] into a 4D light field structure of dimensions $13 \times 13 \times 434 \times 625$. According to the challenge's format, the images are then converted into YUV422 color space with 10 bits precision.

Fig. 4. Thumbnails of the test LF images.

We implement our coding scheme with the x265 HEVC encoder which works in the "main422-10" mode. The bitrate parameter is adjusted to achieve final bit per pixel (bpp) 0.75, 0.10, 0.02 and 0.005, as required by the challenge. The 85 selected views are rearranged in horizontal zigzag scan order and then compressed. We synthesize the views in set B by solving our optimization problem with the Spectral Projected Gradient for L1 (SPGL1) Matlab toolbox [22]. The coefficients are quantized and compressed as one lossless JPEG image. The sum of bytes of bitstream and the coefficients are taken as the final rate for the light field image.

To evaluate the compression efficiency, both the origin and compressed light field images are converted into the same light field structure. The PSNR is calculated per view and we report the average YUV PSNR for all views. The bpp is taken as final bit rates for the light field images.

Comparisons are carried out between our linear approximation prior based scheme(LAP) and a x265 HEVC codec working in the same manner in our encoder (x256-HZ), but for all the views. The LAP is also compared with [8] under the defined processing and over the database required by ICME 2016 grand challenge. On the average a bit rate reduction 25.98% is achieved . We do not list those results because of the different pre-processing and data format requirements of the two grand challenges.

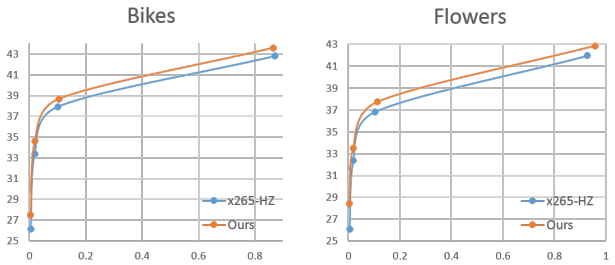

Fig. 5. Rate-distortion curves of LAP and x265-HZ anchor.

The rate-distortion curves are partially presented in Fig.5. It is obvious that our coding scheme outperforms the x265-HZ method. Table 1 summarizes the BD-PSNR and BD-BR [7] comparisons for LAP and x265-HZ. As can be seen, for image *Friends 1* the highest PSNR gain 1.52 dB is achieved. For all the LF images, bit rate reduction of at least 33% could be achieved. It indicates that our method is powerful and robust for different scenes. The average PSNR gain of the 6 test images reaches 1.32 dB. we observe that the PSNR gain at lower bit rate is more significant than at higher bit rate.

No.	LF Image	BD-PSNR (dB)	BD-BR
1	Bikes	1.03	-33.50%
2	Danger de mort	1.33	-33.33%
3	Flowers	1.33	-32.28%
4	Stone Pillars Outside	1.25	-35.68%
5	Fountain & Vincent 2	1.49	-45.51%
6	Friends 1	1.52	-44.16%
Average		1.32	-37.41%

Table 1. Comparisons of our LAP scheme with x265-HZ.

Additionally, the proposed scheme has 0.026 SSIM better than the anchor on the average. Especially, our scheme will achieve a 0.068 SSIM gain compared with the anchor at 0.005 bpp. We also provide a visual comparison for LAP with the x265-HZ at the same bit rate 0.004 bpp. Views at the same position are extracted. As shown in Fig.6, our coding scheme (the left one) preserves more details (e.g. the hair before forehead and grass in the background) and offers much better visual experience. It is noticeable that the left view is **synthesized** by our optimization model.

Fig. 6. Visual comparison for LAP (left) and x265-HZ.

5. CONCLUSION

In this paper we propose a new prior, i.e. the linear approximation prior, which reveals the linearity for light field images in angular domain. Based on this new prior, we propose a coding scheme. The experiments demonstrate that our scheme is efficient and robust for different scenes.

6. REFERENCES

[1] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance

imaging using large camera arrays," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3, pp. 765–776, 2005.

[2] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[3] Christian Perwass and Lennart Wietzke, "Single lens 3d-camera with extended depth-of-field," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 829108–829108.

[4] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu, "Saliency detection on light field," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2806–2813.

[5] Touradj Ebrahimi, Siegfried Foessel, Fernando Pereira, and Peter Schelkens, "Jpeg pleno: Toward an efficient representation of visual reality," *Ieee Multimedia*, 2016.

[6] Stefan Heber and Thomas Pock, "Shape from light field meets robust pca," in *European Conference on Computer Vision*. Springer, 2014, pp. 751–767.

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

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

[9] Feng Dai, Jun Zhang, Yike Ma, and Yongdong Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4733–4737.

[10] Shengyang Zhao, Zhibo Chen, Kun Yang, and Hongrui Huang, "Light field image coding with hybrid scan order," in *Visual Communications and Image Processing (VCIP), 2016*. IEEE, 2016, pp. 1–4.

[11] Insung Ihm, Sanghoon Park, and Rae Kyoung Lee, "Rendering of spherical light fields," in *Computer Graphics and Applications, 1997. Proceedings., The Fifth Pacific Conference on*. IEEE, 1997, pp. 59–68.

[12] Bernd Girod, Chuo-Ling Chang, Prashant Ramanathan, and Xiaoqing Zhu, "Light field compression using disparity-compensated lifting," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 1, pp. I–373.

[13] Xuan Cao, Zheng Geng, and Tuotuo Li, "Dictionary-based light field acquisition using sparse camera array," *Optics express*, vol. 22, no. 20, pp. 24081–24095, 2014.

[14] Jie Chen and Lap-Pui Chau, "Light field compressed sensing over a disparity-aware dictionary," *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[15] Jie Chen, Junhui Hou, and Lap-Pui Chau, "Light field compression with disparity guided sparse coding based on structural key views," *arXiv preprint arXiv:1610.03684*, 2016.

[16] Xiaoran Jiang, Mikaël Le Pendu, Reuben A Farrugia, Sheila S Hemami, and Christine Guillemot, "Homography-based low rank approximation of light fields for compression," in *IEEE International Conf. on Acoustics, Speechand Signal Processing (ICASSP), 2017*. IEEE, 2017.

[17] Yun Li, Mårten Sjöström, Roger Olsson, and Ulf Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 80–91, 2016.

[18] Ricardo Monteiro, Luís Lucas, Caroline Conti, Paulo Nunes, Nuno Rodrigues, Sérgio Faria, Carla Pagliari, Eduardo da Silva, and Luís Soares, "Light field hevc-based image coding using locally linear embedding and self-similarity compensated prediction," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.

[19] Richard Szeliski, *Computer vision: algorithms and applications*, Springer Science & Business Media, 2010.

[20] Martin Rerabek, Lin Yuan, Lonard Antoine Authier, and Touradj Ebrahimi, "[ISO/IEC JTC 1/SC 29/WG1 contribution] EPFL Light-Field Image Dataset," 2015.

[21] Donald Dansereau, Oscar Pizarro, and Stefan Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1027–1034.

[22] Ewout Van den Berg and Michael P Friedlander, "Sparse optimization with least-squares constraints," *SIAM Journal on Optimization*, vol. 21, no. 4, pp. 1201–1229, 2011.