Graph-based Transforms for Predictive Light Field Compression based on Super-Pixels

Mira Rizkallah, Xin Su, Thomas Maugey, Christine Guillemot

Graph-based Transforms for Predictive Light Field Compression based on Super-Pixels

Mira Rizkallah Mira.Rizkallah@inria.fr
Xin Su Xin.Su@inria.fr
Thomas Maugey Thomas.Maugey@inria.fr
Christine Guillemot Christine.Guillemot@inria.fr

***Abstract*—In this paper, we explore the use of graph-based transforms to capture correlation in light fields. We consider a scheme in which view synthesis is used as a first step to exploit inter-view correlation. Local graph-based transforms (GT) are then considered for energy compaction of the residue signals. The structure of the local graphs is derived from a coherent super-pixel over-segmentation of the different views. The GT is computed and applied in a separable manner with a first spatial unweighted transform followed by an inter-view GT. For the inter-view GT, both unweighted and weighted GT have been considered. The use of separable instead of non separable transforms allows us to limit the complexity inherent to the computation of the basis functions. A dedicated simple coding scheme is then described for the proposed GT based light field decomposition. Experimental results show a significant improvement with our method compared to the CNN view synthesis method and to the HEVC direct coding of the light field views.**

***Index Terms*—Light Fields, Compression, Graph Transforms, Super-pixels**

I. Introduction

Recently, there has been a growing interest in light field imaging. By sampling the radiance of light rays emitted by the scene along several directions, light fields enable a variety of post-capture processing techniques such as refocusing, changing perspectives and viewpoints, depth estimation, simulating captures with different depth of fields and 3D reconstructions [21], [15], [6]. This comes at the expense of collecting large volumes of high-dimensional data, which appears to be the key downside of light fields. To overcome the problem, research effort has been dedicated to light field compression.

Many methods have been proposed so far to adapt standardized solutions (in particular HEVC) to light field data [3], [4], [13], [16], [14], [12]. In [10], the authors propose an homography based low rank approximation method to reduce the angular dimensionality prior encoding. A compression scheme based on view synthesis is described in [9]. The view synthesis predicts all the views from a subset of views.

Here we focus on the problem of residue coding with graph transforms in a light field compression scheme based on view synthesis. We consider the view synthesis method proposed in [11] based on a learning architecture using two consecutive convolutional neural networks (CNN). From features extracted from 4 corner views of the light field, the first CNN predicts depth maps which are then used to produce by warping 4 estimates of each synthesized view. A second CNN then reconstructs each light field view from these 4 estimates. The compression scheme assumes the transmission of a subset of views (the four corner views to have the maximum information of the scene including dis-occlusions). These four views are encoded using HEVC-Inter. The synthesis of the whole light field from these 4 views already gives a reconstructed light field of good quality. However, to further enhance the light field quality, the transmission of the residue between the synthesized and original views is needed.

While the authors in [9] use HEVC-inter to code the residues of all the synthesized views as a pseudo sequence, in this paper we explore the use of graph-based transforms to better de-correlate the residue signals within and across views. The large number of transforms proposed in the literature can be classified into non-localized transforms derived from the Graph Laplacian and into localized wavelet-like transforms. Graph transforms, *e.g.,* graph Fourier transforms [18] and graph wavelet transform [7], have been used to compress piecewise smooth images (*e.g.,* depth images) [8], [19], dis-occluded pixels in light fields [20], 3D point clouds [2].

In this paper, we consider transforms computed from the Graph Laplacian in order to best de-correlate the residue signals to be encoded. In order to construct a graph which would best capture pixel dependencies, super-pixels are computed on a reference view using the SLIC algorithm [1]. The super-pixels group pixels located in a local region having similar color values. Super-pixels in other views are assumed to be co-located to cope with a behaviour of the graph transforms which will be discussed in the paper.

Local non-separable graphs could then be constructed connecting pixels of all the views to jointly capture correlation spatially and across views. However, the Laplacian of such graph, despite the locality, remains of high dimension leading to a high transform computational cost. For this reason, we consider instead separable local transforms. A local spatial graph is constructed per super-pixel for each view leading to a first spatial GT. A second set of graphs connecting coefficients of same bands computed by the first spatial transforms is then constructed to capture inter-view correlations. For this inter-view GT, we consider both an unweighted and a weighted GT. A dedicated coding scheme of the transformed coefficients is then described. The method is compared against two HEVC-based schemes, one scheme (called HEVC-Lozenge) which directly encodes the set of views as a pseudo video sequence with HEVC-inter coding [17], and a second scheme which, after view synthesis based on the CNN architecture, encodes the residues with HEVC-inter coding.

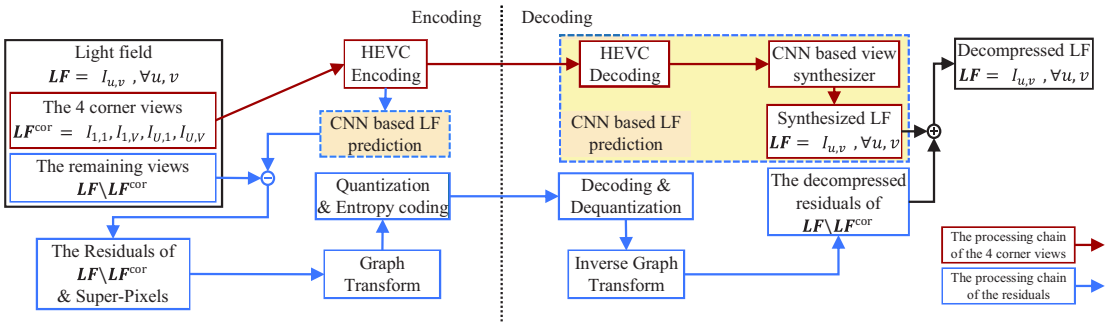

Fig. 1: Overview of proposed coding scheme.

II. Light Field Predictive Coding Scheme

A. Scheme Overview

Fig.1 depicts the proposed coding scheme. Let $\mathbf{LF} = \{I_{u,v}\}$ denote a light field, where $u = 1, \ldots, U$ and $v = 1, \ldots, V$ are the view indices. Four views at the corners $\mathbf{LF}^{\text{cor}} = \{I_{1,1}, I_{1,V}, I_{U,1}, I_{U,V}\}$ are encoded using HEVC-Inter and used to synthesize the whole light field with the CNN based synthesis method [11], as shown in Fig.1 (red arrows). To improve the quality of the synthesized light field, the residuals between the synthesized and original views are encoded by graph transforms, (see Fig.1, blue arrows). The residuals of all the views but the 4 corner views $\mathbf{LF} \setminus \mathbf{LF}^{\text{cor}}$ are considered here. These residual signals are grouped into super-pixels using the SLIC algorithm [1], then graph transforms are applied on each super-pixel followed by quantization and entropy coding. At the decoder, the decompressed residuals are added to the synthesized light field to obtain the final decompressed light field.

B. CNN based View Synthesis

Machine learning methods have been recently considered for view synthesis. In [11], the authors only use the four corner sub-aperture views to synthesize the whole light field with high quality by two convolutional neural networks (CNN). One of the CNNs is trained to model the disparity in the given light field, while the other one is used to estimate the color of the synthesized views.

In this paper, we use this architecture to predict the light field views from the four corner views, as shown by the yellow parts in Fig.1. Four corner views $\mathbf{LF}^{\text{cor}}$ are encoded using HEVC-Inter for transmission. At the decoder, the whole light field can be synthesized using these four decompressed views $\widehat{\mathbf{LF}}^{\text{cor}}$. The synthesis quality depends on the QP value of the HEVC-inter coder.

C. Super-Pixel Segmentation

The next step is to encode the residual signals for each predicted view. We consider graph transforms adapted to the local signal characteristics. In order to define these local transforms, super-pixels are computed on a reference view (we take here the central view of the synthesized light field) using the SLIC algorithm [1] which groups pixels having similar color values and that are close spatially, as shown in Fig.2. The segmentation in the *central* view is propagated to other views without changing the position and size of the segmentation mask. Note that the super-pixels are computed on the synthesized view, since the synthesized views are available at both the encoder and decoder. In this case, we can recover the super-pixel segmentation from the four corner views $\widehat{\mathbf{LF}}^{\text{cor}}$ and do not need to transmit it.

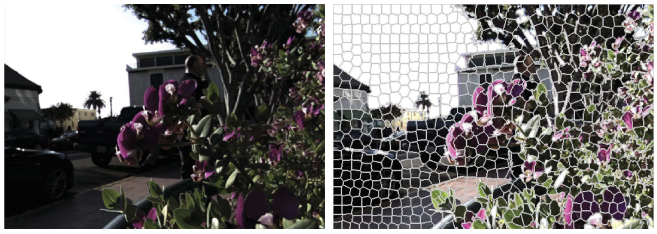

Fig. 2: The original view $I_{4,4}$ of *Flower1* dataset in [11] (left) and the corresponding super-pixel segmentation (right).

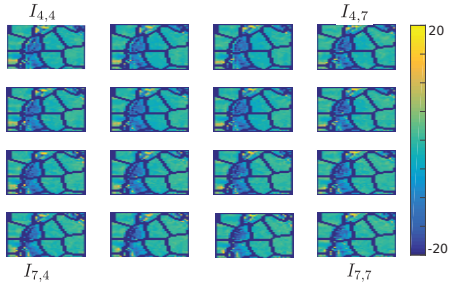

Fig. 3: Illustration of coherent residual signals in superpixels for a subset of views of Flower 1 (luminance).

III. Graph-based Transform and Coding

Thanks to the superpixel ability to adhere to image borders, the sub-aperture residual images are subdivided into uniform regions where the residual signal is supposed to be smooth. Fig. 3 shows the luminance values of a cropped region of the residues for a subset of views of the Flower 1 dataset. Although the disparity is not taken into account, the signals in super-pixels which are co-located across the views are correlated for light fields with narrow baselines. In order to capture these correlations, we use a separable Graph Trans-

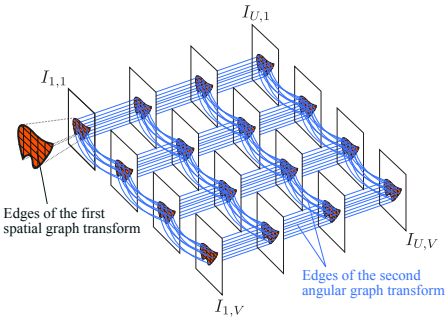

Fig. 4: Illustration of the two graphs used to compute the two local separable graph transforms.

form comprising a local super-pixel based spatial GT followed by a local angular GT.

A. First Spatial Graph Transform

We first construct local spatial graphs inside each superpixel for each view. More precisely, If we consider the residues luminance values in one sub-aperture image v of the light field and a segmentation map M, the k^{th} superpixel SP_k can be represented by a signal $f_k^v \in \mathbb{R}^{N_k}$ defined on an undirected connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$ which consists of a finite number of vertices corresponding to the pixels at positions $\{i_l, j_l\}, l = 1 \dots N_k$ such that $M(i_l, j_l) = k$. Edges are drawn intuitively between each pixel and its 4 neighbors to form the set $\mathcal{E}$. If there is an edge $e = (m, n)$ between two vertices m and n, the entry A_{mn} is equal to 1 otherwise, $A_{mn} = 0$.
The adjacency matrix is used to compute the Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is a diagonal degree matrix whose i^{th} diagonal element D_{ii} is equal to the sum of the weights of all edges incident to node i. The resulting Laplacian matrix $\mathbf{L}$ is symmetric positive semi-definitive and therefore can be diagonalized as:

$$\mathbf{L} = \mathbf{U^T \Lambda U} \tag{1}$$

where $\mathbf{U}$ is the matrix whose rows are the eigenvectors of the graph Laplacian and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. The matrix $\mathbf{U}$ is used to compute the unweighted Graph Fourier Transform (GFT): for the signal f_k^v defined on the vertices of the graph, the transformed coefficients vector $\hat{f}_k^v$ is defined in [18] as:

$$\hat{f}_k^v = \mathbf{U} f_k^v \tag{2}$$

The inverse graph Fourier transform is then given by

$$f_k^v = \mathbf{U}^T \hat{f}_k^v \tag{3}$$

We have observed that when the spatial graph topology in the corresponding super-pixels in different views undergoes a slight change, the basis functions of each spatial GT are not the same, resulting in decreased correlation of the spatial transform coefficients across views, hence in decreased performance of the angular GT. This is the reason why we take the same segmentation map for all the views. Although this does not exploit disparity, this guarantees that we project the signals residing on each super-pixel in all the views onto the same basis functions.

B. Second Angular Graph Transform

In order to capture inter-view dependencies and compact the energy into fewer coefficients, we perform a second graph based transform. We examine two different cases where the weights are either fixed to 1 or learned from a training set of spatial transformed coefficients. Since we have the same number of pixels for a specific superpixel in all the views, we then deal with a graph made of N_v vertices corresponding to the views to be coded. Edges are drawn between each node and its direct four neighbors.

a) Unweighted GT: The Adjacency is used to compute the inter-view unweighted Laplacian as $\mathbf{L_v} = \mathbf{D_v} - \mathbf{A_v}$ with $\mathbf{D_v}$ the degree matrix. $\mathbf{L_v}$ can be diagonalized as:

$$\mathbf{L_v} = \mathbf{U_v^T \Gamma U_v} \tag{4}$$

For a specific band number l and superpixel k, the band signal is defined as $b_k^l = \{\hat{f}_k^v(l), \quad v = 1 : N_v\} \in \mathbb{R}^{N_v}$. The unweighted Graph Transform consists of projecting the signal onto the eigenvectors of $\mathbf{L_v}$ as:

$$\hat{b}_k^l = \mathbf{U_v} b_k^l \tag{5}$$

The inverse unweighted Graph Transform is then given by

$$b_k^l = \mathbf{U_v^T} \hat{b}_k^l \tag{6}$$

A major assumption lying behind the use of the unweighted version of a laplacian is a constant pairwise relationship between neighboring nodes which may not accurately reflect the statistical precisions in our case especially for high frequencies where different patterns of correlations can be observed. To overcome this problem, we propose to find weight matrices $\mathbf{W}$ for different sets of frequency bands.

b) Weighted GT: Instead of applying the same graph transform to all the bands, we divide them into 64 groups, ranging from low to high frequencies. For each group, we compute the sample covariance matrix from a set of training superpixels spatial coefficients. We solve the minimization problem defined in [5] to compute 64 different generalized laplacian matrices, that can be either computed separately for each dataset and sent as additional information or learned for a set of training datasets and stored in the decoder side. Due to the high computational cost of the first option, we will learn a fixed set of 64 laplacian matrices to be exploited for all datasets. Let $\mathbf{\Psi}_v^h$ be the matrix whose columns contain the *wGT* basis for a specific group h i.e., the eigenvectors of the corresponding weighted laplacian. The band signals belonging to this group are thus projected onto this basis.

C. Transform coefficient coding

At the end of those two transform stages, coefficients are grouped into a three-dimensional array $\mathbf{R}$ where $\mathbf{R}(i_{SP}, i_{bd}, v)$ is the v^{th} transformed coefficient of the band i_{bd} for the superpixel i_{SP}. Using the observations on all the superpixels in a training dataset (*Flower1* with QP 40), we can find the best ordering for quantization. We first sort the variances of coefficients with enough observations in

decreasing order. We then split them into 64 classes assigning to each class a quantization index in the range 1 to 64. All the remaining coefficients with less observations will be considered in the last group. We use the zigzag ordering of the *JPEG* quantization matrix to assign the quantization step size for each. The quantized coefficients are further coded using an arithmetic coder. Note that to construct the spatial and angular graphs on the decoder side, we do not need extra information. Since the decoder already received the four corner images, the CNN method is used to predict the whole Light field. With the SLIC algorithm, the decoder can deduce the segmentation map which is fixed for all the views. Also, for the weighted angular graph transform, the laplacians are learned on a training set of light fields and then fixed.

IV. Experimental results

We test our coding scheme on four real LF with 8×8 sub-aperture images of size ($X = 536, Y = 376$) from the dataset used in [11], called *Flower1*, *Flower2*, *Cars* and *Rock*. We first evaluate the energy compaction of the transformed coefficients for the three transforms (only spatial GT, spatial + unweighted angular GT, spatial + weighted angular GT) to show the utility of exploring inter-view correlation.

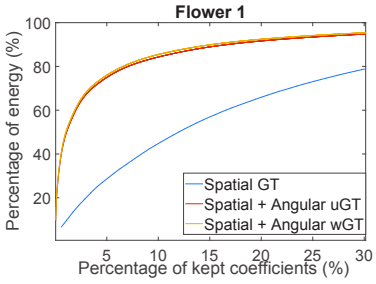

Fig. 5: Energy Compaction of the transformed residues for Flower 1 corner images compressed with HEVC QP=40.

Energy compaction is measured by ordering all coefficients (for the luminance component) according to their decreasing variances. The total energy in the transform coefficients is the same as that in the Light Field residual signal, due to the orthogonality of the transforms. Fig. 5 shows the fraction of the total energy captured by α % of transform coefficients as a function of α for the residuals of Flower 1. Higher energy compaction is observed with the second angular transform compared with only applying the spatial transform, with a slight improvement for the wGT. This shows the utility of exploring the inter-view correlations between residues in different views and adapting the graph weights for that purpose compared to only performing local spatial transforms.

In Fig. 6, our results are generated by selecting the best pairs of parameters (Q, QP) where Q is the quality parameter used to control the quantization of the transformed residuals and QP is used in the HEVC inter-coding of the four corners. Such selection can be automatically predicted after training a model represented by a function of light field features and target bitrate as in [10]. The observed bit allocation

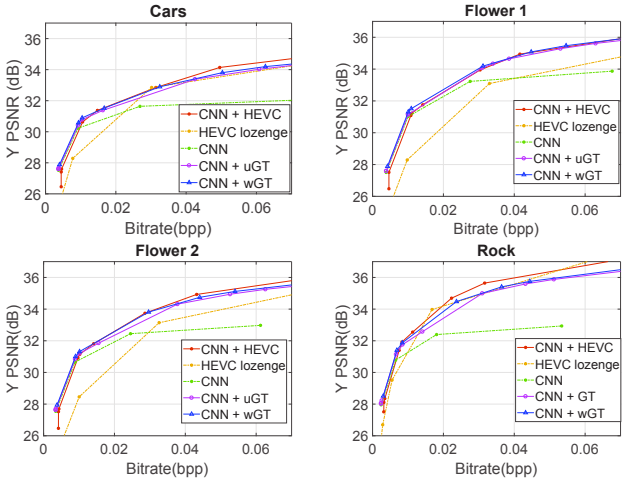

Fig. 6: Rate-distortion comparison.

TABLE I: Bjontegaard comparison (ΔPSNR (dB)) at low bitrate (< 0.04 bpp)

	CNN+uGT *vs*			CNN+wGT *vs*
	CNN	HEVC lozenge	CNN+HEVC	CNN+uGT
Car	0.6	0.9	0.3	0.1
Flower 1	0.3	1.7	0.2	0.1
Flower 2	0.4	1.6	0.3	0.2
Rock	-0.1	0.7	-0.1	0.3

(Low bitrate: GT 7% HEVC 93%, High bitrate: GT 40% HEVC 60%) shows the strength of the prediction at low bitrate. Moreover, we observe that for the four datasets, our Graph based transform approaches defined by CNN+uGT and CNN+wGT slightly outperform CNN learning based scheme at low bitrate and bring a small improvement to the HEVC based coding of the residues (Table I). For higher bitrates, the compression performance is further enhanced compared to CNN, and almost reaching CNN+HEVC performance. At low to middle bitrates, both graph-based transform schemes outperform direct use of HEVC inter coding as we can also observe after computing the bjontegaard metric in Table I.

V. Conclusion

In this paper, we have explored the use of graph-based transforms to capture correlation in the spatial and angular dimensions of light fields. Once a prediction using a CNN architecture is performed, local graph based transforms are applied in a separable manner to compact the energy of the residues. Experimental results show that we can enhance the quality of the reconstructed light field while maintaining a small coding bitrate. Also, we have shown a high gain compared to HEVC inter-coding of the light field as a video sequence.

Acknowledgment

This work has been supported in part by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

REFERENCES

[1] R. Achanta, A. Shaji, Kevin K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.*

[2] Aamir Anis, Philip A Chou, and Antonio Ortega. Compression of dynamic 3D point clouds using subdivisional meshes and graph wavelet transforms. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6360–6364. IEEE, 2016.

[3] C. Conti, P. Nunes, and L. D. Soares. HEVC-based light field image coding with bi-predicted self-similarity compensation. In *IEEE Int. Conf. on Multimedia Expo Workshops (ICMEW)*, Jul. 2016.

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

[5] H. E. Egilmez, E. Pavez, and A. Ortega. Graph learning from data under laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, Sept 2017.

[6] T. Georgiev and A. Lumsdaine. Focused plenoptic camera and rendering. *J. of Electronic Imaging*, 19(2), Apr. 2010.

[7] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[8] Wei Hu, Gene Cheung, Antonio Ortega, and Oscar C Au. Multiresolution graph fourier transform for compression of piecewise smooth images. *IEEE Transactions on Image Processing*, 24(1):419–433, 2015.

[9] X. Jiang, M. Le Pendu, and Guillemot C. Light field compression using depth image based view synthesis. In *Hot3D workshop held jointly with ICME*, July 2017.

[10] X. Jiang, M. Le Pendu, R. Farrugia, and Guillemot C. Light field compression with homography-based low-rank approximation. *IEEE J. on Selected Topics in Signal Processing, special issue on light field image processing*, Oct. 2017.

[11] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6), 2016.

[12] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag. Efficient intra prediction scheme for light field image compression. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 539–543, Florence, Italy, May 2014.

[13] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng. Pseudo-sequence-based light field image compression. In *IEEE Int. Conf. on Multimedia Expo Workshops (ICMEW)*, Jul. 2016.

[14] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. da Silva, and L. Soares. Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction. In *IEEE Int. Conf. on Multimedia Expo Workshops (ICMEW)*, Jul. 2016.

[15] R. Ng. *Light Field Photography*. PhD thesis, Stanford University, 2006.

[16] C. Perra and P. Assuncao. High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement. In *IEEE Int. Conf. on Multimedia Expo Workshops (ICMEW)*, Jul. 2016.

[17] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot. Impact of light field compression on focus stack and extended focus images. In *24th European Signal Processing Conf. (EUSIPCO)*, pages 898–902, Aug. 2016.

[18] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

[19] Weng-Tai Su, Gene Cheung, and Chia-Wen Lin. Graph Fourier Transform with Negative Edges for Depth Image Coding. In *IEEE ICIP*, 2017.

[20] Xin Su, Mira Rizkallah, Thomas Maugey, and Christine Guillemot. Graph-based light fields representation and coding using geometry information. In *IEEE International Conference on Image Processing (ICIP)*, 2017.

[21] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. *ACM Trans. Graph.*, 24(3):765–776, July 2005.