SCALABLE LIGHT FIELD COMPRESSION SCHEME USING SPARSE RECONSTRUCTION AND RESTORATION

Fatma HAWARY [†] [*], *Christine GUILLEMOT*[*], *Dominique THOREAU*[†] *and Guillaume BOISSON*[†]

[*]Inria Rennes Bretagne-Atlantique, Campus de Beaulieu, Rennes, France
[†]Technicolor, 975 Avenue des Champs Blancs, 35576 Cesson Sévigné, France

ABSTRACT

This paper describes a light field scalable compression scheme based on the sparsity of the angular Fourier transform of the light field. A subset of sub-aperture images (or views) is compressed using HEVC as a base layer and transmitted to the decoder. An entire light field is reconstructed from this view subset using a method exploiting the sparsity of the light field in the continuous Fourier domain. The reconstructed light field is enhanced using a patch-based restoration method. Then, restored samples are used to predict original ones, in a SHVC-based SNR-scalable scheme. Experiments with different datasets show a significant bit rate reduction of up to 24% in favor of the proposed compression method compared with a direct encoding of all the views with HEVC. The impact of the compression on the quality of the all-in-focus images is also analyzed showing the advantage of the proposed scheme.

Index Terms— Light field, compression, HEVC, SHVC, sparsity.

1. INTRODUCTION

During the last two decades, there has been a growing interest in light field imaging. By capturing the radiance of light rays emitted by a scene along various directions, light fields enable a variety of post-capture image processing applications such as digital refocusing, viewpoint and perspective changing, depth estimation, 3D scene reconstruction, or creating images with an extended focus (where all pixels are in focus). Light fields can be captured by either an array of cameras [1] (resulting in a wide baseline) or by single cameras mounted on moving gantries, or by plenoptic cameras using arrays of micro-lenses placed in front of the photosensor, leading to the light fields with narrow baselines [2, 3]. The acquired light field data exhibits large amount of information, which poses challenging problems in terms of storage capacity, hence the need for efficient compression schemes.

Effort has already been dedicated to light field compression in the community, using either simple tools such as vector quantization followed by Lempel-Ziv (LZ) entropy coding [4] or wavelet coding schemes [5, 6], which nevertheless yield a limited compression efficiency of the order of 20:1. In addition, 4D wavelet transforms and PCA were applied for the progressive coding [7, 8]. In the recent years, the state-of-the-art mono and multi-view video compression schemes, considering the light field as an array of multiple views, have also been applied. Inspired by H.264 and MVC, [9] and [10] proposed new schemes, where specific prediction modes are added, with or without disparity compensation.

Among the most recent proposals, the latest HEVC video compression standard was considered for light fields: the array of sub-aperture images can be taken as an input [11], or, in other cases, direct encoding is performed over the lenslet images captured by plenoptic cameras [12]. Prediction modes have also been proposed and integrated into an HEVC encoder to compress 3D holoscopic content [13]. In [14], a sparse set of micro-lens images (also called elemental images) is encoded in a base layer. The other elemental images are reconstructed at the decoder using disparity-based interpolation and inpainting. The reconstructed images are then used to predict the entire lenslet image and a prediction residue is transmitted yielding a 3-layer scheme. Another version of the multi-view HEVC extension for light field video was introduced in [15], where the inter-view prediction is represented in a two-directional parallel structure. A homography-based low rank approximation is proposed in [16], to exploit correlation between the sub-aperture images, using HEVC intra coding of the low rank representation. An alternative way to compress light field images was also proposed in [17], where the focal stack (which consists of photographs focused at different depth values) is encoded with a 3D wavelet and SPIHT scheme. The entire light field is then reconstructed from the compressed focus stack using a combination of dimension reduction and a 2-D filtering. The paper in [18] presents a layered approach that segments objects in the ray space and applies wavelets for the compression of each segment.

In this paper, we introduce a novel scalable coding method for the light field data based on their sparsity in the angular (view) Fourier domain. A selected set of the light field sub-aperture images is encoded in a base layer as a video sequence using HEVC and transmitted to the decoder. The non-selected views are then reconstructed from the decoded subset of views, by exploiting the light field sparsity in the angular

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

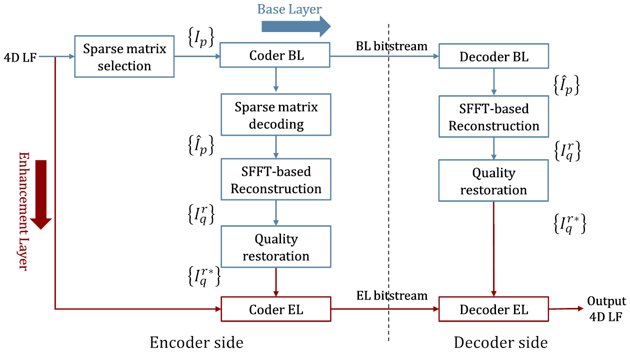

Fig. 1: An overview of the proposed compression scheme.

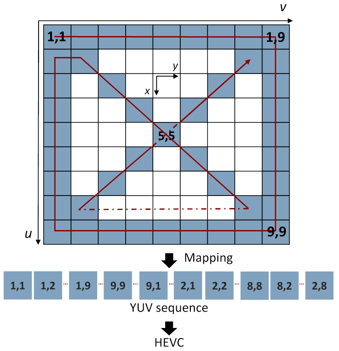

Fig. 2: Selected sub-aperture images $\{I_{\mathbf{p}}\}_{\mathbf{p}\in P}$ sent as video frames following a specific scan order to HEVC encoder.

continuous Fourier domain. The quality of the reconstructed light field is afterwards enhanced to allow its use in predicting the original light field within a scalable coding structure. The prediction residue is encoded in an enhancement layer. The proposed approach has demonstrated a high compression efficiency for both real and synthetic light field datasets. A significant blur reduction has also been observed in the correspondent all-in-focus images.

2. COMPRESSION SCHEME

Let $L(x, y, u, v)$ denote the 4D representation of a light field, describing the radiance of a light ray parameterized by its intersection with two parallel planes [19]. The angular (view) coordinates are denoted with (u, v), where $u = 1 \dots U$ and $v = 1 \dots V$. The spatial (pixel) coordinates are denoted with (x, y), where $x = 1 \dots N_x$ and $y = 1 \dots N_y$). The view at the angular position (u, v) is defined as $I_{u,v}$.

The main steps of the proposed compression scheme are depicted in Fig.1. First, a sparse set $\{I_{\mathbf{p}}\}_{\mathbf{p}\in P}$ of light field views at pre-defined positions in P is selected (the selection approach is shown in Fig.2), and encoded as a video sequence using HEVC [20] in a base layer (BL). Then, the set of non-selected views $\{I_{\mathbf{q}}\}_{\mathbf{q}\in Q}$ is reconstructed from the decoded views $\{\hat{I}_{\mathbf{p}}\}_{\mathbf{p}\in P}$ ($Q \cup P = \Omega$, Ω denoting the entire set of view positions). The reconstruction is performed using a method which exploits light field sparsity in the angular Fourier domain [21]: SFFT (Sparse Fast Fourier Transform). Finally, the restored light field is used as an inter-layer predictor of the original light field, in an enhancement layer (EL) using SHVC [22], leading to a two-layer SNR-scalable scheme.

2.1. Sparse Reconstruction

A subset of light field views $\{I_{\mathbf{p}}\}_{\mathbf{p}\in P}$ is first selected and then encoded in HEVC as a YUV sequence (Fig.2) in the base layer. The corresponding decoded views $\{\hat{I}_{\mathbf{p}}\}_{\mathbf{p}\in P}$ are used to reconstruct the remaining views. For this purpose, a reconstruction method, inspired by [21], is used. Assuming the light field is k-sparse in the angular continuous Fourier domain, it can be represented as a linear combination of k non-zero continuous angular frequency coefficients

$$\mathscr{L}_{w_x,w_y}(u,v) = \sum_{i=0}^{k} \frac{a_{u,v}(i)}{N} exp(2j\pi \frac{uw_u(i) + vw_v(i)}{N}), \tag{1}$$

where $\{w_u, w_v\}$ are the continuous frequency positions and $\{a_{u,v}\}$ are the corresponding coefficients. The objective is to recover the set $F = \{w_u, w_v, a_{u,v}\}$ of the sparse spectrum from the decoded views.

First, the 2D Fourier transform of each input view (u, v) from the decoded set $\{\hat{I}_{\mathbf{p}}\}_{\mathbf{p}\in P}$ is computed as

$$\mathscr{L}_{u,v}(w_x, w_y) = FFT(L_{u,v}(x, y)), \tag{2}$$

which gives the spatial frequencies (w_x, w_y) at the positions set P, representing the input 1D discrete lines. We will refer to this data as $\mathscr{L}_{w_x,w_y}(u, v)_{|P}$, from which the 2D angular spectrum $\mathscr{L}_{w_x,w_y}(w_u, w_v)$ for each spatial frequency (w_x, w_y) will be recovered.

The algorithm proceeds by first estimating integer frequency positions $\{w_u, w_v\}$ using a voting approach from the input sparse set. Then, the corresponding coefficients $\{a_{u,v}\}$ are estimated and the frequency positions are refined to non-integer values, using a two-step iterative approach. The method is detailed in Fig.3.

2.1.1. Initial frequency positions estimation

The goal of this step is to initialize the set of positions $\{(w_u, w_v)_i\}_{i=0}^{k}$ of the non-zero frequency positions. Per the slicing theorem [23], the Fourier transform of a discrete line of a signal gives the projection of its spectrum onto that line. Therefore, we calculate the Fourier transform in the angular domain (u, v) of each line segment of the input data $\mathscr{L}_{w_x,w_y}(u, v)_{|P}$. This yields projections of the 2D spectrum $\mathscr{L}_{w_x,w_y}(w_u, w_v)$ on these lines. Then, the discrete frequency positions that receive a vote from each projection will construct the initial estimation for angular frequencies' positions. Based on the assumption that the spectrum is sparse, only

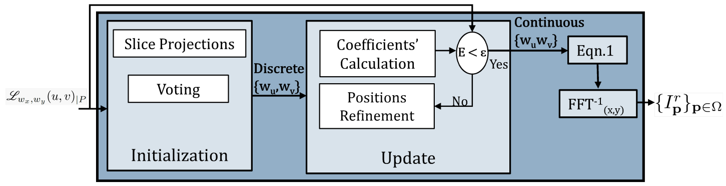

Fig. 3: An overview of the sparse reconstruction scheme [21].

few frequencies that receive a vote each time will finally be preserved.

2.1.2. *Coefficients estimation and frequency refinement*

The algorithm proceeds afterwards by iteratively refining the coefficients and their frequency positions as follows:

- Given a set of positions $\{w_u, w_v\}$, the corresponding coefficients $\{a_{u,v}\}$ are recovered by solving the linear system of Eqn.1 over the set P of known (u, v) positions;
- Given the coefficients $\{a_{u,v}\}$, the frequency positions $\{w_u, w_v\}$ are refined to minimize the residual error $E(P)$, using a gradient descent algorithm based on finite differences

$$E(P) = \sum_{\mathbf{p}\in P} ||\hat{I}_{\mathbf{p}} - I^*_{\mathbf{p}}||^2, \tag{3}$$

where $\hat{I}_{\mathbf{p}}$ denotes the BL-decoded view at position $\mathbf{p}$ and $I^*_{\mathbf{p}}$ is the view at the same position, obtained by the sparse reconstruction algorithm. The error threshold value is transmitted in the BL coder. Based on the final frequency positions and corresponding coefficients, all the light field views are reconstructed using Eqn.1, followed by the inverse Fourier transform in (x, y) coordinates. This gives the reconstructed light field data $\{I^r_{\mathbf{p}}\}_{\mathbf{p}\in\Omega}$.

2.2. Quality Restoration

The quality of the reconstructed views $\{I^r_{\mathbf{q}}\}_{\mathbf{q}\in Q}$ may not be sufficient for some post-processing applications, because of the introduced reconstruction noise. We therefore consider these images, along with the images $\{\hat{I}_{\mathbf{p}}\}_{\mathbf{p}\in P}$, as inter-layer predictors of the original views in a scalable scheme which further encodes a prediction residue. However, to ensure an efficient inter-layer prediction, the quality of the reconstructed images $\{I^r_{\mathbf{q}}\}_{\mathbf{q}\in Q}$ is first improved using a patch-based restoration method. The proposed restoration technique searches with the PatchMatch algorithm [24] for the best matches between patches in each reconstructed view $I^r_{\mathbf{q}}$ and the spatially closest BL-decoded image $\hat{I}_{\mathbf{p}}$, and vice versa. Once the matching process is over, the pixel values of the restored image are computed as a weighted average of overlapping patches, which minimizes the bidirectional similarity distance [25] applied between the reference image and the restored one. To illustrate the pixel value calculation, let R_1,..., R_m denote all the patches in $I^r_{\mathbf{q}}$ that contain a pixel s. D_1,...,D_m indicate the corresponding best matches in $\hat{I}_{\mathbf{p}}$, where d_1,...,d_m are the co-located pixels (see Fig.4). Also, let $\hat{R}_1$,...,$\hat{R}_n$ denote all the patches in $I^r_{\mathbf{q}}$ that contain pixel s and serve as the best matches to $\hat{D}_1$,...,$\hat{D}_n$ in $\hat{I}_{\mathbf{p}}$, and $\hat{d}_1$,...,$\hat{d}_n$ be the co-located pixels in patches $\hat{D}_1$,...,$\hat{D}_n$ (see Fig.4). Hence, the value of s in the final restored image $I^{r*}_{\mathbf{q}}$ is expressed as

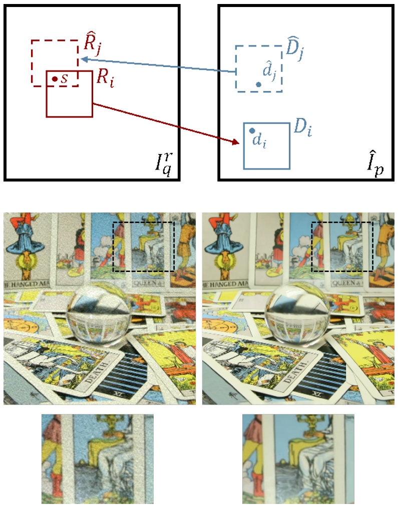

Fig. 4: Quality Restoration. First row: pixel restoration based on bidirectional matching. Second row: an example of a reconstructed image $I^r_{\mathbf{q}}$ (left) and its corresponding restored image $I^{r*}_{\mathbf{q}}$ (right) from *Crystal* dataset: note that the presence of heavy noise in the reconstructed image does not allow its use as an inter-layer predictor to enhance the compression efficiency of the EL.

$$I^{r*}_{\mathbf{q}}(s) = \frac{\sum_{i=1}^{m} \hat{I}_{\mathbf{p}}(d_i) + \sum_{j=1}^{n} \hat{I}_{\mathbf{p}}(\hat{d}_j)}{m+n}. \tag{4}$$

2.3. Enhancement Layer

The restored views $\{I^{r*}_{\mathbf{q}}\}_{\mathbf{q}\in Q}$ and the BL-decoded $\{\hat{I}_{\mathbf{p}}\}_{\mathbf{p}\in P}$ are loaded into the inter-layer reference picture list in a SNR-scalable coder in SHVC, for the prediction of the original light field. Both inter-layer and intra/inter predictions are performed during encoding, and the best coding mode is chosen (with Rate Distortion Optimization) for each block. Residues from prediction are then quantized, transformed and encoded, delivering the enhancement layer (EL) bitstream. A final full light field data is delivered by the decoder, using both BL and EL bitstreams.

3. EXPERIMENTAL RESULTS

The sub-aperture images in the BL are encoded using the HEVC Test Model (HM 16.9), with a GOP size of 8 in a hierarchical structure. Only the upper-left view is intra-encoded.

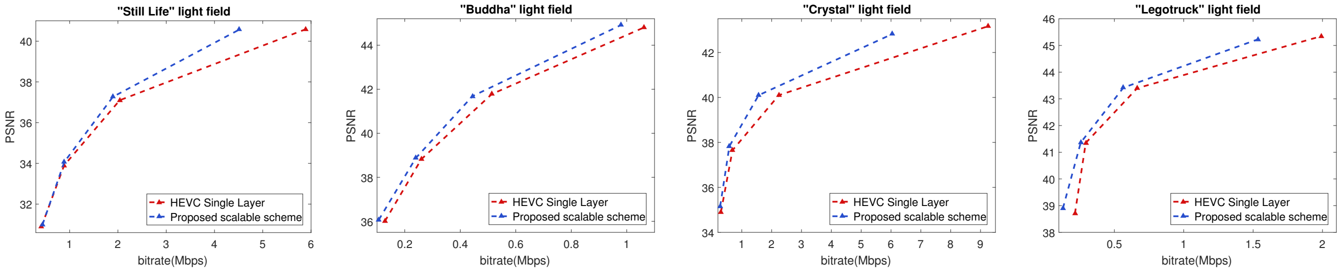

Fig. 5: PSNR-rate performance of the proposed coding method (using QP=22, 27, 32, 37) compared to HEVC Single Layer.

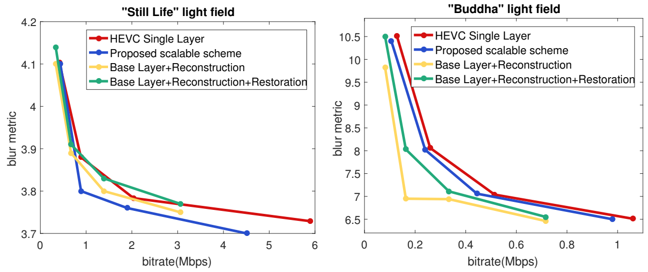

Fig. 6: Amount of blur in the all-in-focus image resulting from different output data: a higher blur measure indicates a lower quality of the all-in-focus image.

The QP values are set to 22, 27, 32 and 37. The matrix of views is scanned row-by-row from left to right and from right to left to form a video sequence, which is converted to YUV (4:2:0). This sequence is then encoded using the SHVC Test Model (SHM 12.1) with the same parameters as above, yielding the EL bitstream. We compare our compression scheme with a HEVC single layer (SL) coding of the original matrix of views converted into a YUV sequence, in the same way as the EL sequence.

Light field images *Buddha* and *Still Life*[1] (768x768), as well as *Crystal* (1024x1024) and *Lego Truck* [2](1280x860) are used for tests. The BL sequence is formed by 45 views out of 81 for the HCI synthetic light fields, and by 93 out of 289 for the Stanford real light fields. The objective quality is assessed on the YUV components with PSNR averaged on all views of the matrix, and the bit-rate is calculated from the coded bitstream for all YUV components. Rate distortion curves are plotted for PSNR vs. bitrate (Fig.5). The results in Fig.5 show that significant gains are obtained for the proposed compression scheme, compared to direct HEVC encoding in a single layer. Furthermore, estimations performed with the Bjontegaard metric [26] demonstrate the bit-rate reductions of 8.2% and 11.83% for synthetic data *Still Life* and *Buddha* respectively, while even more prominent bit-rate reductions of 18% and 24.24% are achieved for real light fields *Lego Truck* and *Crystal*.

The impact of compression on the quality of the all-in-focus images is then analyzed for the HCI synthetic light fields, for which the ground truth depth map is available. We first generate the focus stack $\{\mathfrak{F}_{\alpha_i}\}$ of light field. A digital refocusing of the central view is performed as explained in [2]. The all-in-focus image $E(x, y)$ is obtained by choosing, for a pixel at position (x, y) of depth α_i, the pixel of the image from the focus stack which is refocused at this depth $(\mathfrak{F}_{\alpha_i}(x, y))$. Fig.6 shows the blur measure as a function of the bitrate. The blur is measured with a perceptual edge-based blur metric [27].

The results in Fig.6 demonstrate that the proposed scheme outperforms HEVC single layer coding and introduces less distortion in decoded images. High compression ratios are achieved, without altering the visual quality of the all-in-focus image. Besides, it can be remarked that the output of the reconstruction based on BL-decoded views (illustrated in yellow) provides an equivalent quality of the all-in-focus image as the HEVC single layer coding, but at a significantly lower compression cost.

4. CONCLUSION

We introduced a scalable coding scheme for light field data. A sparse viewpoint subset is selected and encoded as a base layer with HEVC. A full light field is reconstructed from the decoded images using a sparse recovery method in the Fourier domain. These reconstructed data are enhanced using the decoded views, and then used as prediction reference for inter-layer coding of the entire original light field.

The proposed scheme is scalable with two layers, so that the data used for rendering can either consist of the plain reconstruction from the sparse view samples, or its refined version with the enhancement layer. Experimental results demonstrate that this scalable scheme outperforms HEVC single layer encoding for the tested light field datasets, both synthetic and real. The analysis of all-in-focus images also shows that our method does not induce visual artifacts, even for data reconstructed from the base layer, which presents an advantageous outcome for further post-capture light field applications.

[1] https://hci.iwr.uni-heidelberg.de/hci/softwares/light_field_analysis

[2] http://lightfield.stanford.edu/lfs.html

5. REFERENCES

[1] B. Wilburn, N. Joshi, V. Vaish, E. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *ACM Trans. on Graphics (TOG)*, vol. 24, no. 3, pp. 765, 2005.

[2] R. Ng, "Digital light field photography," *Stanford University*, pp. 1–203, 2006.

[3] T. Georgiev, G. Chunev, and A. Lumsdaine, "Superresolution with the focused plenoptic camera," *Proc. of SPIE*, vol. 7873, pp. 78730X–78730X–13, 2011.

[4] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. of SIGGRAPH '96*, pp. 31–42.

[5] P. Lalonde and A. Fournier, "Interactive rendering of wavelet projected light fields," in *Proc. of the Conference on Graphics Interface '99*.

[6] I. Peter and W. Straßer, "The wavelet stream - progressive transmission of compressed light field data," in *IEEE Visualization 1999 Late Breaking Hot Topics*. 1999, pp. 69–72, IEEE Computer Society.

[7] M. Magnor, A. Endmann, and B. Girod, "Progressive compression and rendering of light fields," *Vision, Modeling and Visualization*, pp. 199–203, 2000.

[8] D. Lelescu and F. Bossen, "Representation and coding of light field data," *Graph. Models*, vol. 66, no. 4, pp. 203–225, July 2004.

[9] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 338–343, 2000.

[10] B. Girod, C. L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated Lifting & Shape Adaptation," *Proc. of ICME*, vol. 1, no. 4, pp. I373–I376, 2003.

[11] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on focus stack and extended focus images," *EUSIPCO*, pp. 898–902, 2016.

[12] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," *ICASSP*, pp. 539–543, 2014.

[13] C. Conti, P. Nunes, and L. D. Soares, "New HEVC prediction modes for 3D holoscopic video coding," *ICIP*, pp. 1325–1328, 2012.

[14] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. on Image Processing*, vol. 25, no. 1, pp. 80–91, 2016.

[15] G. Wang, W. Xiang, M. Pickering, and C. W. Chen, "Light field multi-view video coding with two-directional parallel inter-view prediction," *IEEE Trans. on Image Processing*, vol. PP, no. 99, pp. 5104–5117, 2016.

[16] X. Jiang, M. Le Pendu, R. A Farrugia, S. Hemami, and C. Guillemot, "Homography-based low rank approximation of light fields for compression," *ICASSP*, 2017.

[17] T. Sakamoto, K. Kodama, and T. Hamamoto, "A study on efficient compression of multi-focus images for dense light-field reconstruction," *VCIP*, 2012.

[18] A. Gelman, P. L. Dragotti, and V. Velisavljevic, "Multiview image compression using a layer-based representation," *ICIP*, pp. 493–496, 2010.

[19] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," *Proc. of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 43–54, 1996.

[20] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[21] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. on Graphics*, vol. 34, no. 1, pp. 1–13, 2014.

[22] G. J. Sullivan, J. M. Boyce, Y. Chen, J. Ohm, C. A. Segall, and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC)," *IEEE Journal on Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1001–1016, 2013.

[23] R. Ng, "Fourier slice photography," *ACM Trans. on Graphics*, vol. 24, no. 3, pp. 735, 2005.

[24] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing," *ACM Trans. on Graphics*, vol. 28, no. 3, pp. 1, 2009.

[25] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *IEEE (CVPR)*, 2008.

[26] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *Doc. VCEG-M33, ITU-T VCEG Meeting*, 2001.

[27] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," *Proc. of the International Conference on Image Processing (ICIP)*, vol. 3, pp. 8–11, 2002.