

Prediction and Sampling with Local Graph Transforms for Quasi-Lossless Light Field Compression

Mira Rizkallah, Thomas Maugey, Christine Guillemot

▶ **To cite this version:**

Mira Rizkallah, Thomas Maugey, Christine Guillemot. Prediction and Sampling with Local Graph Transforms for Quasi-Lossless Light Field Compression. IEEE Transactions on Image Processing, In press, pp.1-13. hal-02373664

HAL Id: hal-02373664

https://inria.hal.science/hal-02373664

Submitted on 21 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prediction and Sampling with Local Graph Transforms for Quasi-Lossless Light Field Compression

Mira Rizkallah, *Student Member, IEEE,* Thomas Maugey, *Member, IEEE,* and Christine Guillemot, *Fellow, IEEE*
INRIA Rennes Bretagne Atlantique, Campus Universitaire de Beaulieu, 35042 Rennes, France

***Abstract*—Graph-based transforms have been shown to be powerful tools in terms of image energy compaction. However, when the support increases to best capture signal dependencies, the computation of the basis functions becomes rapidly untractable. This problem is in particular compelling for high dimensional imaging data such as light fields. The use of local transforms with limited supports is a way to cope with this computational difficulty. Unfortunately, the locality of the support may not allow us to fully exploit long term signal dependencies present in both the spatial and angular dimensions in the case of light fields. This paper describes sampling and prediction schemes with local graph-based transforms enabling to efficiently compact the signal energy and exploit dependencies beyond the local graph support. The proposed approach is investigated and is shown to be very efficient in the context of spatio-angular transforms for quasi-lossless compression of light fields.**

***Index Terms*—Light Fields, Energy Compaction, Transform coding, Super-rays, Graph Fourier Transform, Prediction, Sampling**

I. Introduction

Light field imaging is gaining in popularity due to its potential for a number of computer vision applications. However, light fields represent very large volumes of high dimensional and redundant data. Depending on the light field sampling and capturing device, raw data rates can indeed vary from a few hundreds of Mbytes to Gbytes, hence the need to design efficient tools to best capture signal dependencies and compact its energy for compression.

In this paper, we consider graph based transforms to exploit correlation in the 4D ray space. Graphs have indeed been shown to be useful tools to describe intrinsic image structures, hence to define the supports of de-correlating transforms. Fourier-like transforms, called graph Fourier transform (GFT) [1] and many variants [2], [3], [4], [5], [6], [7] have been shown to be powerful tools for coding piecewise smooth and natural 2D images. An interesting review can be found in [8]. However, when the dimension of the signal increases, the dimension of the graph increases and the complexity inherent to the computation of the GFT basis functions rapidly becomes untractable. This is obviously the case for light field data making a complete graph connecting all light rays unsuitable for this task.

To cope with this difficulty, we consider instead local transforms with limited supports. In order to take into account the scene geometry, the supports of the non separable graphs are defined by super-rays. Super-rays have been first introduced in [9] as an extension of super-pixels in the 3D domain to group light rays coming from the same 3D object, i.e. to group pixels having similar color values and being close spatially in the 3D space. While the locality of the support allows us to reduce the computation complexity of the basis functions, it does not allow us to capture long term spatial dependencies of the signal, unlike efficient predictive schemes used in state of the art codecs (e.g. HEVC). The correlation between different super-rays is not exploited.

In this paper, we introduce sampling and prediction schemes to exploit correlation beyond the limits of the local graph transform support. More precisely, based on the graph sampling theory, the proposed methods allow taking advantage of the good energy compaction property of graph transforms on local supports, i.e. with a limited complexity, while benefiting from well established but powerful prediction mechanisms in the pixel domain. The idea is to first sample the light field data and to encode these references samples with any image coder having powerful Intra prediction mechanisms. The local non separable graph transform is then computed, but only its high frequency coefficients are coded and transmitted. We derive the equations allowing us to recover the low frequency coefficients of the local graph transforms from its coded high frequency coefficients and from the encoded reference samples. The encoding of these reference samples is a way to efficiently encode the low frequency coefficients containing most of the light field energy, using intra prediction mechanisms of state-of-the-art coders. In the experiments, we used HEVC Intra (HM-16.10).

In this general framework, one key question to address is the best choice of the reference samples. The most natural way would be to take all the pixels of a reference view. However, due to matrix conditioning problems, that we will discuss in the paper, the recovered low frequencies are, in that case, very sensitive to high frequencies coefficients quantization. In order to overcome this issue, we sample the graph in each super-ray, across views, and project the samples into one reference image. Although this approach gives good performance in terms of energy compaction and quasi-lossless compression of the light field, using a complete graph per super-ray still suffers from complexity limitations and high sensitivity to the quantization noise present in the high frequency coefficients.

To further decrease the basis function computational com-

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

plexity, we also consider separable local graph transforms applying first a spatial followed by an angular transform. Unlike in the non separable case, the prediction equations do not suffer from numerical instabilities and from the presence of quantization noise in the high frequencies. The reference samples can thus be taken from a light field view.

This second approach keeps the advantages of both the reduced basis function computational complexity due to the limited support and of the structured set of reference samples (one entire view) that can be easily coded with intra-prediction mechanisms. It however keeps only in part the advantage of the energy compaction of the graph transform since the recovered frequencies do not necessarily correspond to the low frequencies. This second approach extends the spatio-angular prediction scheme described in [10] for separable spatio-angular graph transforms.

The proposed methods can be seen as graph-based prediction schemes deriving low frequency spatio-angular coefficients from one single compressed reference image (e.g. the projected set of reference samples in the non-separable case, or the top-left view in the separable case) and from the high frequency coefficients. The methods have been assessed in the context of quasi-lossless encoding of light fields. Experimental results show that, when coupled with a powerful intra-prediction tool, the graph-based spatio-angular prediction brings a substantial gain in bitrate reaching almost 30%.

The rest of the paper is organized as follows. After a brief overview of state of the art graph transforms, graph sampling and light field compression methods in Section II, we introduce the local graph transforms in Section III. Section IV then describes the proposed sampling and prediction techniques for both the non separable and separable spatio-angular graph transforms. The coding schemes are detailed in Section V and experimental results are presented and discussed in Section VI.

II. Notation and Preliminaries

In this section, after introducing the notations, we review the basics of graph-based transforms and graph sampling theory. We then give a quick overview of the approaches considered so far for light field compression.

A. Notation

The notations used throughout the paper are introduced in Table I. Please note that lowercase normal (e.g. α), lowercase bold (e.g. $\mathbf{x}$), uppercase bold (e.g. $\mathbf{U}$) letters denote scalars, vectors and matrices, respectively. Unless stated otherwise, calligraphic capital letters (e.g. $\mathcal{E}$ and $\mathcal{V}$) represent sets.

B. Graph transforms

A graph has been shown to be a useful tool to describe the intrinsic image structure, hence to capture correlation, which is necessary for image compression. An interesting review of graph spectral image processing can be found in [8].

For image compression, the signal is defined on an undirected connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ which consists of a finite set $\mathcal{V}$ of vertices corresponding to the pixels. A set $\mathcal{E}$ of edges connect each pixel and its 4-nearest neighbors in the spatial domain. By encoding pixel similarities into the weights associated to edges, the undirected graph encodes the image structure. A Fourier-like transform for graph signals called graph Fourier transform (GFT) [1] and many variants [2], [3], [4], [5], [6], [7] have been used as adaptive transforms for coding piecewise smooth and natural images.

TABLE I: Notations

	Symbol	Meaning
Graph	$\mathcal{G}$, $\mathbf{L}$	Unweighted Graph, graph Laplacian matrix
	$\mathcal{E}$, $\mathcal{V}$	Set of edges, set of vertices
	$\mathbf{A}$, $\mathbf{D}$	Adjacency, Degree matrices
Sampling	$\mathcal{S}$	Set of graph vertices
	$\mathcal{S}_c$	the complement of the set $\mathcal{S}$
	$\mathcal{T}$	Set of low frequency coefficients,
	$\mathcal{T}_c$	the complement of the set $\mathcal{T}$
Local Graph Transform	$\mathcal{G}_k$, $\mathcal{V}_k$, $\mathbf{L}_k$	Graph, vertices and Laplacian inside a super-ray k
	$\mathbf{L}_{k,v}$	spatial graph in super-pixel v of super-ray k.
	N_k	Number of pixels in the k^{th} super-ray
	$N_{k,v}$	Number of pixels in the v^{th} superpixel of super-ray k
	$\mathbf{L}_k^b$	angular graph for the band b of super-ray k.
	$\mathbf{x}_k$	Graph signal inside the k^{th} super-ray
	$\mathbf{x}_{k,v}$	Graph signal inside the v^{th} superpixel of the k^{th} super-ray
	$\hat{\mathbf{x}}_k$	non-separable transformed coefficients in the k^{th} super-ray
	$\hat{\mathbf{x}}_{k,v}$	spatial transformed coefficients in the k^{th} super-ray
	$\mathbf{U}_k$, $\mathbf{U}_{k,v}$,$\mathbf{V}_k^b$	Eigenvector matrices of the laplacians $\mathbf{L}_k$,$\mathbf{L}_{k,v}$ and $\mathbf{L}_k^b$
LF	$\mathbf{I}_l$	The l^{th} subaperture image of a light field
	$\mathbf{SM}_{ref}$	Segmentation map of a reference view
	$\tilde{\mathbf{I}}_l$	The l^{th} reconstructed subaperture image of a light field

A spectrum of graph frequencies can be defined through the eigen-decomposition of the graph Laplacian matrix $\mathbf{L}$ defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, 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 matrix $\mathbf{A}$ is the adjacency matrix with entries $A_{mn} = 1$, if there is an edge $e = (m, n)$ between two vertices m and n, and $A_{mn} = 0$ otherwise.

The Laplacian matrix $\mathbf{L}$ is symmetric positive semi-definitive and therefore can be diagonalized as:

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

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 eigenvectors $\mathbf{U}$ of the Laplacian of the graph are analogous to the Fourier bases in the Euclidean domain and allow representing the signals residing on the graph as a linear combination of eigenfunctions akin to Fourier Analysis [1]. This is known as the Graph Fourier transform.

C. Graph Sampling

Let us consider a connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ made of N vertices associated with a Laplacian $\mathbf{L}$. It has a complete set of eigenvalues $\lambda_l, l \in [1,\ N]$ and eigenvectors $\mathbf{u}_l, l \in [1,\ N]$. A graph signal is bandlimited and has a bandwidth λ_n if it can be expressed as a linear combination of only the first n eigenvectors of $\mathbf{L}$.

A subset of vertices $S \subset \mathcal{V}$ is a *uniqueness set* [11] for bandlimited signals (with bandwidth λ_n) if for any two bandlimited signals $\mathbf{f}, \mathbf{g}$, $\mathbf{f}(\mathcal{S}) = \mathbf{g}(\mathcal{S}) \implies \mathbf{f} = \mathbf{g}$.

It is also shown that $\mathcal{S}$ is a *uniqueness set* for all λ_n-bandlimited signals $\mathbf{f}$, if and only if

$[\mathbf{u}_1(\mathcal{S}), \mathbf{u}_2(\mathcal{S}), \ldots, \mathbf{u}_n(\mathcal{S})]$ are linearly independent where λ_n is the n^{th} smallest eigenvalue of $\mathbf{L}$ and $\mathbf{u}_i(S) \in \mathbb{R}^{|S|}$ is a reduced eigenvector. The term reduced implies taking the rows of the eigenvectors corresponding to the indices of the sampling set $\mathcal{S}$ [12].

It has also been shown that for any uniqueness set $\mathcal{S}$ of size n for λ_n-bandlimited signals, there is always at least one node $\int_i \notin \mathcal{S}$ such that $\mathcal{S} \bigcup \int_i$ is a uniqueness set of size $n+1$ for λ_{n+1}-bandlimited signals. [12] [13] ($\bigcup$ denotes the union of two sets). This property shows that one can always find a uniqueness set for all bandlimited signals $\forall\ \lambda$. The set of samples can be iteratively selected from the input light field data.

After building a *minimum uniqueness set* $\mathcal{S}$ of size n, let us denote the complement of this set in $\mathcal{V}$ as $\mathcal{S}_c$. A simple way to reconstruct the missing samples on $\mathcal{S}_c$ is to solve a least-squares problem in the spectral domain [11]. The bandlimited signal $\mathbf{f}$ can be written as:

$$\mathbf{f} = \begin{bmatrix} \mathbf{f}(\mathcal{S}) \\ \mathbf{f}(\mathcal{S}_c) \end{bmatrix} \tag{2}$$

$$= \left[\begin{array}{cccc} \mathbf{u}_1(\mathcal{S}) & \mathbf{u}_2(\mathcal{S}) & \ldots & \mathbf{u}_n(\mathcal{S}) \\ \hline \mathbf{u}_1(\mathcal{S}_c) & \mathbf{u}_2(\mathcal{S}_c) & \ldots & \mathbf{u}_n(\mathcal{S}_c) \end{array}\right] \begin{bmatrix} \hat{\mathbf{f}}(1) \\ \hat{\mathbf{f}}(2) \\ \cdots \\ \hat{\mathbf{f}}(n) \end{bmatrix}$$

$$= \left[\begin{array}{c} \tilde{\mathbf{U}}(\mathcal{S}) \\ \hline \tilde{\mathbf{U}}(\mathcal{S}_c) \end{array}\right] \begin{bmatrix} \hat{\mathbf{f}}(1) \\ \hat{\mathbf{f}}(2) \\ \cdots \\ \hat{\mathbf{f}}(n) \end{bmatrix}, \tag{3}$$

where columns of $\tilde{\mathbf{U}}$ are the n first eigenvectors of the $\mathbf{L}$. Since $\mathcal{S}$ is a *minimum uniqueness set*, $\tilde{\mathbf{U}}(\mathcal{S})$ is square invertible. The vector $\left[\hat{\mathbf{f}}(1), \hat{\mathbf{f}}(2), \ldots, \hat{\mathbf{f}}(n)\right]^\top$ can thus be retrieved by

$$\left[\hat{\mathbf{f}}(1), \hat{\mathbf{f}}(2), \ldots, \hat{\mathbf{f}}(n)\right]^\top = \left(\tilde{\mathbf{U}}(\mathcal{S})\right)^{-1} \mathbf{f}(\mathcal{S}), \tag{4}$$

and the missing samples are reconstructed as follows:

$$\mathbf{f}(\mathcal{S}_c) = \tilde{\mathbf{U}}(\mathcal{S}_c) \begin{bmatrix} \hat{\mathbf{f}}(1) \\ \hat{\mathbf{f}}(2) \\ \cdots \\ \hat{\mathbf{f}}(n) \end{bmatrix} \tag{5}$$

where, $[.]^\top$ and $[.]^{-1}$ denote the transpose and inversion of a matrix respectively.

While the previous sampling theorem [11] has been proposed for band-limited signals, we extend those equations to our problem in the following section. More precisely, we deal with signals (i.e. color signals) that might not be necessarily band-limited on the underlying graph supports (i.e. on the super-rays).

D. Light field compression

The availability of commercial light field cameras has given momentum to the development of light field compression algorithms. Many solutions proposed so far adapt standardized image and video compression solutions (in particular HEVC) to light field data. This is the case *e.g.* in [14], [15], [16], [17], where the authors extend HEVC intra coding modes by adding new prediction modes to exploit similarity between lenslet images. This is also the case in [18], [19], [20], where the views are encoded as pseudo video sequences using HEVC or the latest JEM software, or in [21] where HEVC is extended for coding an array of views.

Low rank models as well as local Gaussian mixture models in the 4D rays space are proposed in [22], [23] and [24] respectively. View synthesis based predictive coding has also been investigated in [25] where the authors use a linear approximation computed with Matching Pursuit View synthesis based predictive coding is another research direction followed in [25] where the authors use a linear approximation computed with Matching Pursuit for disparity based view prediction. The authors in [26] and [27] use instead a the convolutional neural network (CNN) architecture proposed in [28] for view synthesis and prediction. The prediction residue is then coded using HEVC [26], or using local residue transforms (SA-DCT) and coding [27]. The authors in [29], use a depth based segmentation of the light field into 4D spatio-angular blocks with prediction followed by JPEG-2000. View synthesis followed by predictive coding is the approach followed in JPEG-Pleno [30]. Moreover, point cloud construction from light fields have been proposed in [31] and point cloud compression techniques [32], [33], [34] can be applied to code the resulting point clouds.

While all prior work mentioned above has been dedicated to lossy compression, much less effort has been dedicated to lossless coding of light fields. One can however mention the approach proposed in [35] using differential prediction.

III. Super-ray based local graph transforms

Let us consider the 4D representation of light fields proposed in [36] and [37] describing the radiance along rays by a function $L(s,t,m,n)$. Based on this representation, the light field of dimensions (S,T,M,N) can be regarded as an array of views at angular positions (m,n), each view being composed of pixels with spatial coordinates (s,t). In the sequel, to denote one view, we will use the pair of indices (m,n) or an index $v=(m,n)$ to simplify the notations.

Such light fields represent very large volumes of high dimensional data. Graphs connecting all light rays spatially and across views can rapidly become intractable, and in particular the computation of the basis functions, if we consider one unique graph for the entire light field. To overcome this difficulty, here we consider local graph transforms with limited supports. Those local supports are defined by super-rays that are carefully constructed in a way that follows the scene geometry.

A. Local supports: Super-rays

The concept of super-ray has been initially introduced in [9] as an extension of super-pixels to address the computational complexity issue in light field image processing tasks. The term *super-pixel*, first coined in [38] can be seen as the

clustering of image pixels into a set of perceptually uniform regions. Similarly, super-rays can be seen as the clustering of rays of the light field within and across views, hence corresponding to the same set of 3D points of the imaged scene.

To construct super-rays, we proceed as follows. We first compute super-pixels in the top-left view using the SLIC algorithm [39] as well as its disparity map using the method in [40]. An example of an original image with its segmentation is shown in Figure 1.



Fig. 1: An example of super-pixel segmentation for the *Cars* dataset, obtained by fixing the number of superpixels to 800.

Then, using the disparity map, we compute the median disparity per super-pixel and use this median disparity to project the segmentation labels to all the other views. The algorithm proceeds row by row. In the first row of views, we perform horizontal projections from the top-left $I_{1,1}$ to the $N-1$ views next to it. For each other row of views, a vertical projection is first carried out from the top view $I_{1,1}$ to recover the segmentation on view $I_{m,1}$, then $N-1$ horizontal projections from $I_{m,1}$ to the $N-1$ other views are performed. At the end of each projection, some labels are projected in all the views without interfering with others. Those typically represent flat regions inside objects. Others mainly consisting of occluded and occluding segments end up superposed in some views. In this case, the occluded pixels are assigned the label of the neighboring super-ray corresponding to the foreground objects (*i.e.* having the higher disparity). As for appearing pixels, they are clustered with the background super-rays (*i.e.* having the lower disparity).

B. Local graph transforms

1) Local non separable graph transform: We will denote the luminance values of all the light rays (i.e. pixels across all the views) in the k^{th} super-ray, by the vector $\mathbf{x}_k \in \mathcal{R}^{N_k}$, where N_k is the number of rays in the k^{th} super-ray. The k^{th} super-ray $\mathbf{x}_k$ is formed by a set of super-pixels (corresponding super-pixels across the different views). Each super-pixel forming the k^{th} super-ray $\mathbf{x}_k$ will be denoted in a vectorized form, $\mathbf{x}_{k,v} \in \mathcal{R}^{N_{k,v}}$.

We build a non separable graph inside each super-ray where each pixel is connected to its 4 nearest neighbors in the spatial domain (horizontal and vertical neighbors) and to its four corresponding pixels (after projection in the horizontal and vertical neighboring views) in the angular domain. The *local non separable graph transform* of the k^{th} super-ray $\mathbf{x}_k$ is defined as

$$\hat{\mathbf{x}}_k = \mathbf{U}_k^\top \mathbf{x}_k, \tag{6}$$

where the columns of $\mathbf{U}_k$ are the eigenvectors of the local graph laplacian inside the k^{th} super-ray. The inverse graph Fourier transform is then given by

$$\mathbf{x}_k = \mathbf{U}_k \hat{\mathbf{x}}_k \tag{7}$$

2) Local separable graph transform: To further decrease the basis function computational complexity, we also consider the case of a *local separable spatio-angular graph transform*, i.e., applying first a spatial followed by an angular transform. The spatial graph transform coefficients $\hat{\mathbf{x}}_{k,v}$ for each spatial graph $\mathcal{G}_{k,v}$ are obtained by calculating:

$$\hat{\mathbf{x}}_{k,v} = \mathbf{U}_{k,v}^\top \mathbf{x}_{k,v}. \tag{8}$$

where $\mathbf{U}_{k,v}$ are the eigenvectors of the spatial laplacian $\mathbf{L}_{k,v}$ and $\mathbf{x}_{k,v}$ are the luminance values of the super-ray k in view v. Inversely, the luminance values of the pixels belonging to the graph are retrieved from

$$\mathbf{x}_{k,v} = \mathbf{U}_{k,v} \hat{\mathbf{x}}_{k,v}. \tag{9}$$

For each super-ray k, an angular transform is then used to tract similarities between the transformed coefficients $\hat{\mathbf{x}}_{k,v}(b)$ of each band b of the spatial transform coefficients $\hat{\mathbf{x}}_{k,v}$, across the views v. For that, the spatial-band vector is denoted $\hat{\mathbf{x}}_k^b = [\hat{\mathbf{x}}_{k,v}(b)]_{v\in\{1,2,\ldots,M\times N\},\ \text{s.t. } b<|\mathbf{x}_{k,v}|}$, where $M\times N$ is the total number of views. The angular transform coefficients are then obtained by calculating

$$\hat{\hat{\mathbf{x}}}_k^b = {\mathbf{V}_k^b}^\top \hat{\mathbf{x}}_k^b. \tag{10}$$

where $\mathbf{V}_k^b$ is a matrix whose columns are the eigenvectors of the laplacian $\mathbf{L}_k^b$ of the angular graph for the band b.

However, computing the transform on a local support does not allow us to exploit spatial signal dependencies outside the support, resulting in some loss in compression efficiency. To exploit these dependencies, some form of prediction across super-rays is needed.

IV. Super-ray based graph prediction and sampling

The idea we develop here consists in first encoding a selected set of samples, using powerful prediction mechanisms available in state-of-the-art coders (e.g. HEVC). We propose a method that allows us to recover a subset of frequency coefficients of the local graph transforms from these predictively encoded reference samples, hence avoiding to transmit them. This principle is illustrated in Figure 2.

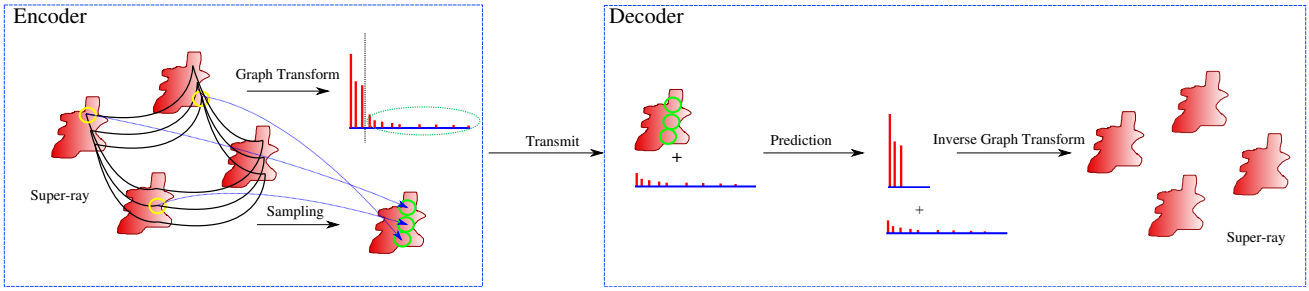


Fig. 2: Overview of the prediction and sampling scheme with a local graph transform applied on a super-ray. On the encoder side, we find the best set of reference samples as explained in section IV-B (algorithm 1) and a local graph transform is applied as in section III-B. The reference samples (surrounded in yellow) are projected in a reference image (as shown with the blue arrows). The high frequencies and the reference image are both sent to the decoder which predicts the low frequency transform coefficients (Equations 13 or 15) to then recover the original super-ray by inverse local graph transform.

A. Graph based prediction

1) Local non separable graph transform: Let us denote $\mathcal{S}$ the set of pixels in a super-ray k belonging to the sampling set (which will be encoded and sent to the decoder side). Let $\mathcal{S}_C$ be the set of all other pixels indices of the super-ray. Let $N_{\mathcal{S}}$ be the cardinal of $\mathcal{S}$. We denote $\mathcal{T}$ the set of $N_{\mathcal{S}}$ lowest frequency coefficients and $\mathcal{T}_C$ the complement of $\mathcal{T}$, i.e. the rest of the frequency coefficients.

Due to the high level of correlation between the different pixels forming a super-ray k, the energy of the transformed coefficients $\hat{\mathbf{x}}_k$ is highly compacted in the low frequencies $\hat{\mathbf{x}}_k(\mathcal{T})$. However, we might still end up with some non-zero high frequencies $\hat{\mathbf{x}}_k(\mathcal{T}_c)$. If we choose an appropriate uniqueness sampling set $\mathcal{S}$ in the k^{th} super-ray, then the inverse graph transform is defined under appropriate permutation as

$$\mathbf{x}_k = \mathbf{U}_k \hat{\mathbf{x}}_k \tag{11}$$

i.e., as

$$\left[\begin{array}{c}\mathbf{x}_k(\mathcal{S})\\ \hline \mathbf{x}_k(\mathcal{S}_c)\end{array}\right] = \left[\begin{array}{c|c}\mathbf{U}_k(\mathcal{S},\mathcal{T}) & \mathbf{U}_k(\mathcal{S},\mathcal{T}_c)\\ \hline \mathbf{U}_k(\mathcal{S}_c,\mathcal{T}) & \mathbf{U}_k(\mathcal{S}_c,\mathcal{T}_c)\end{array}\right] \left[\begin{array}{c}\hat{\mathbf{x}}_k(\mathcal{T})\\ \hline \hat{\mathbf{x}}_k(\mathcal{T}_c)\end{array}\right]. \tag{12}$$

If the signal samples are transmitted separately, $\mathbf{x}_k(\mathcal{S})$ is available at the decoder. If we impose $|\mathcal{S}| = |\mathcal{T}|$, and since we can always find an appropriate uniqueness sampling set $\mathcal{S}$ (as detailed in [12] [13]) i.e. $[\mathbf{u}_1(\mathcal{S}), \mathbf{u}_2(\mathcal{S}), \ldots, \mathbf{u}_{|\mathcal{S}|}(\mathcal{S})]$ are linearly independent as in section II-C, then $\mathbf{U}_k(\mathcal{S},\mathcal{T})$ is a square invertible matrix. Furthermore, if we only transmit $\hat{\mathbf{x}}_k(\mathcal{T}_c)$, then we are able to recover $\hat{\mathbf{x}}_k(\mathcal{T})$ from the following equation:

$$\hat{\mathbf{x}}_k(\mathcal{T}) = \Big(\mathbf{U}_k(\mathcal{S},\mathcal{T})\Big)^{-1}\Big(\mathbf{x}_k(\mathcal{S}) - \mathbf{U}_k(\mathcal{S},\mathcal{T}_c)\hat{\mathbf{x}}_k(\mathcal{T}_c)\Big). \tag{13}$$

Equation (13) is our so-called graph-based spatio-angular prediction. First, $\mathbf{x}_k(\mathcal{S})$ can be seen as a signal composed of a $\lambda_{|\mathcal{S}|}$-band-limited part plus some high frequencies. In this equation, we are actually removing the high frequencies to retrieve the band-limited signal (i.e. $\mathbf{x}_k(\mathcal{S}) - \mathbf{U}_k(\mathcal{S},\mathcal{T}_c)\hat{\mathbf{x}}_k(\mathcal{T}_c)$). Using the least squares reconstruction method in (**??**), we find the low frequency transformed coefficients $\hat{\mathbf{x}}_k(\mathcal{T})$.

Moreover, the high-frequency coefficients $\hat{\mathbf{x}}_k(\mathcal{T}_c)$ can be also seen as prediction coefficients, transmitted to recover the exact light field at the decoder. The basis of the linear prediction is the graph-transform basis, which makes these coefficients low-energetical and thus easy to transmit.

The signal values at $\mathcal{S}_c$ are then retrieved as

$$\hat{\mathbf{x}}_k(\mathcal{S}_c) = \mathbf{U}_k(\mathcal{S}_c,\mathcal{T})\hat{\mathbf{x}}_k(\mathcal{T}) + \mathbf{U}_k(\mathcal{S}_c,\mathcal{T}_c)\hat{\mathbf{x}}_k(\mathcal{T}_c),$$

where the first term is equivalent to the $\lambda_{|\mathcal{S}|}$-band-limited signal recovered on $\mathcal{S}_c$ and the second term is added in order to take into account the high frequency components.

2) Local separable graph transform: Let us assume that view 1 is coded as a reference. In order to perform the prediction, we follow the same reasoning as in the previous (non separable graph) case but we apply it to each band that exists in view 1. For a given super-ray k, the spatial transform in view 1 is $\hat{\mathbf{x}}_{k,1} = \mathbf{U}_{k,1}^{\top}\mathbf{x}_{k,1}$ according to Equation (8).

We choose one sample for each band. It corresponds to the vertex $\mathcal{V}_i$ that is in the reference view (labeled by 1 in our case). For a given band b, the inverse angular transform is defined as

$$\hat{\mathbf{x}}_k^b = \mathbf{V}_k^b \hat{\mathbf{x}}_k^b \tag{14}$$

i.e., as

$$\left[\begin{array}{c}\hat{\mathbf{x}}_k^b(1)\\ \hline \hat{\mathbf{x}}_k^b(2)\\ \vdots\\ \hat{\mathbf{x}}_k^b(N_b)\end{array}\right] = \left[\begin{array}{c|ccc}\mathbf{V}_k^b(1,1) & \mathbf{V}_k^b(1,2) & \cdots & \mathbf{V}_k^b(1,N_b)\\ \hline \mathbf{V}_k^b(2,1) & \mathbf{V}_k^b(2,2) & \cdots & \mathbf{V}_k^b(2,N_b)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{V}_k^b(N_b,1) & \mathbf{V}_k^b(N_b,2) & \cdots & \mathbf{V}_k^b(N_b,N_b)\end{array}\right] \times \left[\begin{array}{c}\hat{\hat{\mathbf{x}}}_k^b(1)\\ \hline \hat{\hat{\mathbf{x}}}_k^b(2)\\ \vdots\\ \hat{\hat{\mathbf{x}}}_k^b(N_b)\end{array}\right]$$

where N_b denotes the number of views where the b^{th} band of the k^{th} super-ray is defined. Since the view 1 is transmitted separately, $\hat{\mathbf{x}}_k^b(1)$ is available at the decoder. If we only transmit $\hat{\hat{\mathbf{x}}}_k^b(2), \ldots, \hat{\hat{\mathbf{x}}}_k^b(N_b)$, the we are able to retrieve $\hat{\hat{\mathbf{x}}}_k^b(1)$ from the following equation:

$$\hat{\mathbf{x}}_k^b(1) = \frac{1}{\mathbf{V}_k^b(1,1)} \times \left(\hat{\mathbf{x}}_k^b(1) - \begin{bmatrix} \mathbf{V}_k^b(1,2) & \cdots & \mathbf{V}_k^b(1,N_b) \end{bmatrix} \begin{bmatrix} \hat{\hat{\mathbf{x}}}_k^b(2) \\ \vdots \\ \hat{\hat{\mathbf{x}}}_k^b(N_b) \end{bmatrix} \right) \tag{15}$$

Equation (15) is our graph-based spatio-angular prediction for the separable case. The spatial coefficients of all the views are then retrieved from the following equation

$$\begin{bmatrix} \hat{\mathbf{x}}_k^b(2) \\ \vdots \\ \hat{\mathbf{x}}_k^b(N_b) \end{bmatrix} = \begin{bmatrix} \mathbf{V}_k^b(2,1) \\ \vdots \\ \mathbf{V}_k^b(N_b,1) \end{bmatrix} \hat{\hat{\mathbf{x}}}_k^b(1) + \begin{bmatrix} \mathbf{V}_k^b(2,2) & \cdots & \mathbf{V}_k^b(2,N_b) \\ \vdots & \ddots & \vdots \\ \mathbf{V}_k^b(N_b,2) & \cdots & \mathbf{V}_k^b(N_b,N_b) \end{bmatrix} \begin{bmatrix} \hat{\hat{\mathbf{x}}}_k^b(2) \\ \vdots \\ \hat{\hat{\mathbf{x}}}_k^b(N_b) \end{bmatrix} \tag{16}$$

Once the decoder has recovered the first spatial graph transform coefficients in all the views, it can reconstruct the whole light field by applying a simple spatial inverse GFT since it has access to the graph supports and coefficients.

B. Graph sampling

1) Local non separable graph transform: To be able to carry out our graph-based spatio-angular prediction in section IV-A, we should at first determine the appropriate sampling set. More precisely, we want to find $\mathcal{S}$ that results in the best conditioning of the sub-matrix $\mathbf{U}_k(\mathcal{S},\mathcal{T})$ that guarantees a small reconstruction error. Simultaneously, we seek a sampling set $\mathcal{S}$ that can be wrapped onto one single view to be coded with efficient prediction mechanisms. A first intuitive way to define the sampling set per super-ray would be to choose the set of N_{k,v_0} pixels that reside in the reference view v_0 that can be subsequently coded with intra HEVC. In our experiments however, we have found that the resulting sub-matrix $\mathbf{U}_k(\mathcal{S},\mathcal{T})$ is ill-conditioned for non-consistent super-rays. We choose instead to use an adapted version of the algorithm described in [12] to choose an appropriate sampling set.

More precisely, for each super-ray k, we specify the band-limit frequency as λ_{n_k} with $n_k = N_{k,v_0}$. We seek to find the optimal sampling set $\mathcal{S}$ that guarantees the exact reconstruction of any λ_{n_k}-band-limited signal on $\mathcal{G}_k$. We know that we have a correspondence between the size of the minimum uniqueness set and the signal bandwidth [12]. We therefore want to find a set of N_{k,v_0} samples. In order to find the vertices that belong to this set, we have to find N_{k,v_0} linearly independent rows from the matrix $\tilde{\mathbf{U}}$. We follow the same reasoning as in [12] but with slightly different constraints to adapt it to our coding problem. In summary, the algorithm takes as input the graphs of all super-rays and the number of samples per super-ray. At the output of this stage, we want to wrap all the samples in a reference view to be efficiently coded with HEVC.

While this method allows an optimal sampling per super-ray, yet, it does not guarantee that the output vector is well structured. It is impossible to say that the samples of neighboring super-rays will be efficiently de-correlated using intra-prediction mechanisms of any efficient coder. In extreme cases, we might end up with noisy samples that are very difficult to code. We thus propose to wrap our samples into one reference view taking into account the geometrical information given by our local graph.

We first observe that our graph laplacian is a sum of two laplacians: The first one includes the connections $\mathbf{L}_k^s$ (s for spatial) inside views, and the other $\mathbf{L}_k^a$ (a for angular) made of edges between pixels inside different views. $\mathbf{L}_k^a$ is actually composed of various connected components, each one corresponding to a 3D point in the scene.

Algorithm 1: Light Field Super-ray Graph based Sampling Algorithm

Data: The set of graphs for all super-rays, Segmentation map of a reference view, the sampling set size per super-ray: $\{\mathcal{G}_k = \{\mathcal{V}^k, \mathbf{L}_k\}\}$, $\mathbf{SM}_{ref}$,$\{n_k\}$
Result: A reference image made of samples drawn in all super-rays: $\mathbf{I}_{ref}$
foreach *Super-ray* k **do**
  Initialize: $\mathcal{S} \leftarrow \mathcal{V}_i^k$ where $\mathcal{V}_i^k$ is the vertice corresponding to the centroid of the super-pixel residing in the reference view ;
  Compute $\tilde{\mathbf{U}}$;
  for $m = 2 \rightarrow n_k$ **do**
    Define $\mathcal{T} = [1, m]$;
    Compute $z = null(\tilde{\mathbf{U}}(\mathcal{S},\mathcal{T}))$;
    Normalize rows of $\tilde{\mathbf{U}}(\mathcal{S}_c,\mathcal{T})$;
    Compute $b = \tilde{\mathbf{U}}(\mathcal{S}_c,\mathcal{T})z$;
    $i \leftarrow argmax_i(|b(i)|)$;
    $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_c(i)$
  end
  Fill $\mathbf{I}_{ref}$ at the right positions : $\mathbf{I}_{ref}(\mathbf{SM}_{ref} = k) = \mathbf{x}_k(\mathcal{S})$;
end

Using the angular information provided by $\mathbf{L}_k^a$, we define the matrix $\mathbf{E}$ of size $(N_{k,v_0} \times N_k)$ where each element gives the correspondence between a pixel in a super-ray k in v_0, and any other pixel in the super-ray. Consider a pixel p_1 in the view v_0. If we can access a pixel p_2 from p_1 following the graph connections in $\mathbf{L}_k^a$ then the entry $\mathbf{E}(p_1,p_2) = 1$, otherwise $\mathbf{E}(p_1,p_2) = 0$.

For each sample $\mathcal{S}(i)$ corresponding to a point p, we find the corresponding point p_0 in the set of pixels $\mathcal{S}_0$ belonging to the super-ray in the first view i.e. p_0 such as $\mathbf{E}(p,p_0) = 1$. The best case scenario is when each sample has a correspondence to a different pixel in the first view. In this case, the projection is easy following the graph links. In the worst case, more than one sample might have a correspondence with the same point in the first view. In this case, first found, first served. The others are considered as disocclusions, and the pixels



(a) (b)

Fig. 3: Reference images for *Cars* and *Friends*: (a) obtained after projection of the sampling sets in all super-rays in the case of non separable graph transforms; (b) original top-left views used as the reference sampling set in the case of separable graph transforms.

having no correspondence in the first view, will be projected into remaining available positions. The complete algorithm is summarized in Algorithm 1 and has an overall complexity of $\mathcal{O}(|\mathcal{S}|^4 + N|\mathcal{S}|^2)$ with N being the number of vertices in a super-ray and $\mathcal{S}$ the number of samples which is quite small compared to the complexity of a huge non local graph transform.

Examples of images obtained after the sampling and projection of the selected sample set on a reference view (considering here the Luminance pixel values) are shown in Figure 3. Despite the non-optimality of this method, we have ascertained that the selected set of reference samples can be efficiently compressed with HEVC using lossless compression settings. Once we have the samples in hand, they can be sent as prediction information to the decoder side, instead of sending the low frequency coefficients. This strategy is a way to efficiently compress the low frequency graph transform coefficients containing about 99 $\%$ of the light field energy, as shown in Table II (left column).

2) Local separable graph transform: For the separable transform case, in addition to a reduced complexity compared to the non separable case, the prediction equations (Equation (15) and (16)) do not suffer from numerical instabilities and from the presence of quantization noise in the high frequencies. This is mainly due to the fact that we don't need a matrix inversion. Thus intuitively we can choose to send one light field view as reference samples for all super-rays. Examples of reference images are shown in Figure 3.

V. Proposed coding schemes based on local graph based transforms

A. Coding scheme with the non separable graph transform

Figure 4 gives an overview of the coding scheme for the non separable case. The top left view $\mathbf{x}_1$ is separated into uniform regions using the SLIC algorithm ([39]) to segment the image into super-pixels, and its disparity map is estimated with [40]. The disparity values are encoded using simple arithmetic coder. The segmentation is coded with edge arithmetic coder(AEC) [41]. Using both the segmentation map and the geometrical (disparity) information, we can build consistent super-rays and graphs in and across all views, as explained in section III-A, at both the encoder and decoder sides. Once the local graphs are computed, we can find the optimal sampling sets (their actual positions in the light field and the corresponding luminance values) as explained in IV-B. Those samples are reorganized in a reference image coded with HEVC intra and sent as prediction information to the decoder.

We apply the non separable graph transform on the coded version of the reference image (quasi-lossless coding) and the original values of all other samples to compact their energy in fewer coefficients. Since the reference image is coded with very small QP, we are almost sure that we are not adding angular incoherence between the different views. Once we have the graph transform coefficients, instead of sending the whole spectrum with simple arithmetic coding, we propose to use the proposed graph-based prediction to derive the low frequency spatio-angular coefficients from the coded reference set of and high angular frequency coefficients, at the decoder side.

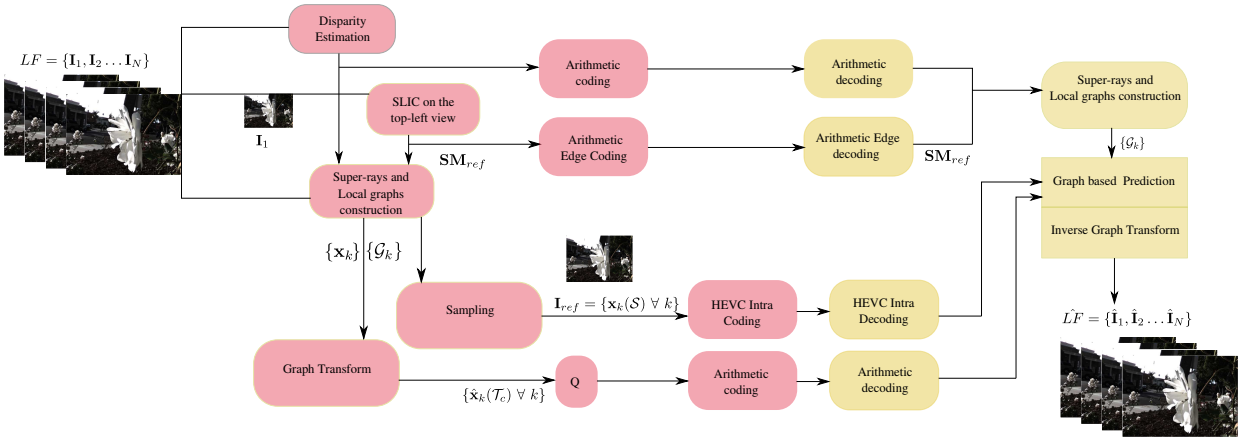


Fig. 4: Overview of proposed coding scheme with the non separable graph transform.

We thus send, for all super-rays, the AC coefficients, i.e., the $(N_k - N_{k,1})$ last bands obtained with the non separable graph transform. (N_k and $N_{k,1}$ are the number of pixels belonging to the super-ray k and those only residing in view 1 respectively). Specifically, after applying the spatio-angular graph transforms on all super-rays, all frequency coefficients are grouped into a two-dimensional array $\mathbf{y}$ where $\mathbf{y}(k, v)$ is the v^{th} transformed coefficient for the super-ray k. Using the natural scanning order (increasing order of eigenvalues), we assign a class number to each super-ray. For a class i, the high frequencies are defined as the last $round(N \times (4-i)/4)$ coefficients where N is the total number of coefficients. Each super-ray belongs to class i if it does not belong to class $i-1$ and the mean energy per high frequency is less than 1. More precisely, we start by finding the super-rays in the first class then remove them from the search space before finding the other classes, and similarly for the following steps. We code a flag with an arithmetic coder to give the information of the class of super-rays to the decoder side. In class i, the last $round(N \times (4-i)/4)$ coefficients of each super-ray are discarded. The remaining high frequency spatio-angular coefficients are quantized uniformly with a small step size $Q = 0.5$. They are grouped into 32 uniform groups to be arithmetically encoded.

B. Coding scheme with the separable graph Transform

The major difference between the coding scheme in the separable case (see Figure 5), compared with the non separable case, resides in the fact that the set of reference samples coded using HEVC-Intra is the top-left view. From this reference view, and the corresponding disparity map, that are transmitted, the decoder can compute the segmentation into super-pixels using the SLIC algorithm and then derive the super-rays used constructing the local graphs and compute the corresponding local graph transforms.

The spatio-angular high frequency graph transform coefficients are coded as in the previous scheme. Using the received reference view and the high frequency coefficients, the decoder can reconstruct all the views as explained in Section IV-A2.

VI. Experimental analysis and results

We applied both coding schemes on real light fields captured by plenoptic cameras from the datasets in [28] and [42]. To avoid the strong vignetting and distortion problems on the views at the periphery of the light field, we only consider the 8×8 central sub-aperture images cropped to 364×524 in [28], and 9×9 cropped to 432×624 from [42], [43]. Some of the light fields considered are shown in Figure 6. The full set of light fields considered for the test is: *Flower2*, *Cars*, *Rock* and *Seahorse* from the dataset in [28] and *StonePillarInside* and *Friends* from the JPEG Pleno Light Field datasets according to common test conditions [42], [43]. The method used to estimate the disparity of the top-left views is described in [40]. Examples of disparity maps provided are shown in Figure 7. A sparse set of disparity values and the segmentation map of the reference view are computed with SLIC [39], and used to construct super-rays, i.e., the local graph supports as described in Section III-A.

The number of super-rays and consequently the super-ray size has an impact on the complexity of the methods, the compaction of the transforms and the side information to be sent and thus on the overall RD performance. More precisely, having a small number of super-rays implies a higher super-ray size hence a tremendous increase in the complexity due to eigen-decompositions. In addition, the disparity errors may have an impact on the disparity compensation and therefore result in a decreased PSNR-Rate performance. On the other hand, having a very large number of super-rays increases the rate needed for segmentation and the disparity values and limits the dimension of each super-ray hence limits the complexity, but results in a smaller benefit in terms of decorrelation of the proposed spatio-angular transforms. Experimentally we have observed for our test datasets that the use of 4000 super-rays gives a good trade-off between computational complexity and compression performance. We thus fix the number of super-rays to 4000 in our experiments.

A. Non Separable vs Separable Graph Prediction

1) Energy compaction: As explained before, we aim at compacting most of the light field energy in few coefficients,

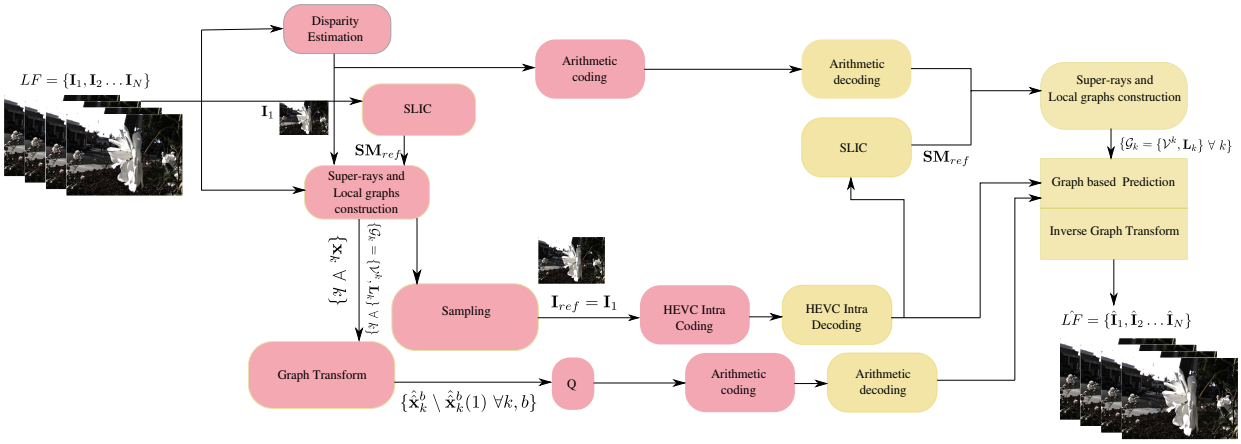


Fig. 5: Overview of proposed coding scheme with the separable graph transform.



Fig. 6: Examples of light fields used in our experiments. Only the top-left view is shown for illustration purpose. From left to right: *Flower2*, *Cars*, *StonePillarsInside* and *Friends*.

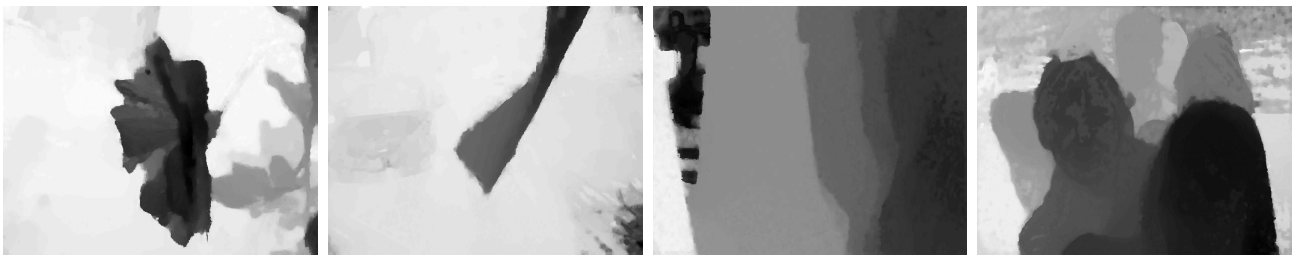


Fig. 7: Disparity maps for examples of light fields used in our experiments.

and at then predicting these coefficients (i.e. they are not transmitted) from a coded reference image and from the high frequency graph transform coefficients that need to be transmitted but at small cost given that they contain little information. Table II gives the percentage of total energy that resides in the predicted DC spatio-angular bands for both non separable ($\hat{x}_k(\mathcal{T}_c) \; \forall \; k$) and separable ($\hat{\mathbf{x}}_k^b(1) \; \forall k, b$) cases. We can observe that most of the energy is compacted in the DC spatio-angular bands, which shows the efficiency in terms of spatio-angular de-correlation of the graph transforms.

The non separable prediction has the benefit of the low energy of the high frequency coefficients of the graph transform that also need to be coded. The separable graph transform, in some cases, looses this benefit as we predict the DC angular (i.e. after the transform across the views) coefficients of all spatial bands. Those low angular frequency coefficients may not contain all the energy otherwise captured by the lower spatio-angular frequency coefficients of the non separable case, although it remains quite efficient in terms of energy compaction as we can see in Table II.

To further illustrate the energy compaction of the transforms, we plot in Figure 8, for two different super-rays, the transform coefficients following the coding order (learned order of frequencies) for both cases: separable and non separable graph transforms. As we can see, in the non separable case, the low frequencies that are predicted on the decoder side (the red dots) correspond to the first frequencies and thus to those who hold most of the energy. However, in the separable case, the coefficients predicted do not necessarily exhibit the highest energy. In the separable case, since we are predicting the first angular frequency coefficient (after the angular transform) for each spatial frequency band(after the spatial transform), some of those coefficients do not hold the highest energy. This is quite clear in the second example, where the red dots in the separable case are assigned to very low values.

2) Compressibility of the reference view : Thanks to the prediction equations introduced in Section IV-A, an efficient encoding of the top-left view in the separable case or the reference view in the non separable case (using any classical encoder with efficient spatial predictors) can be seen as a

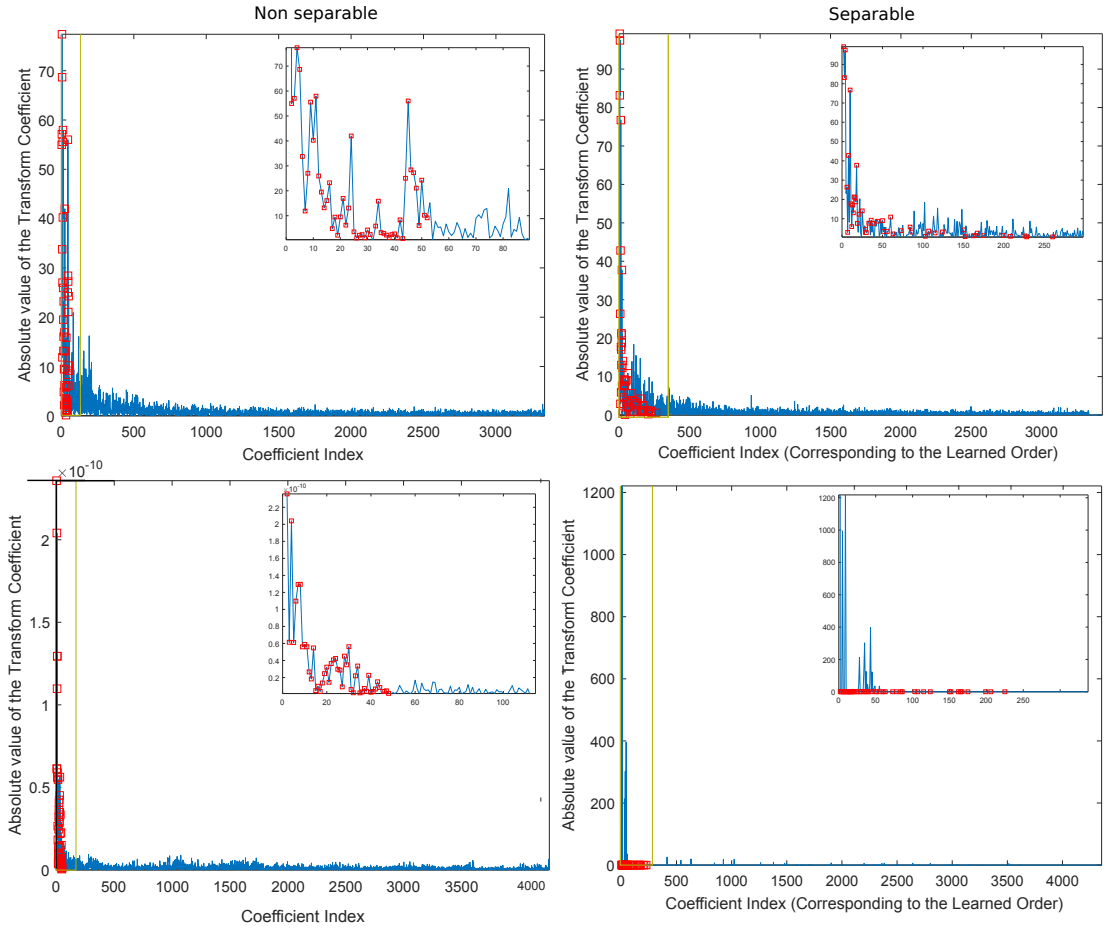


Fig. 8: Energy compaction for two super-rays of *Flower2*. The transform coefficients are ordered with the learned frequency order. The red squares are the predicted DC values on the decoder side. The two rows correspond to two different super-rays and the two columns are for both cases: Non Separable and Separable respectively.

way to encode those DC spatio-angular frequency coefficients which contain most of the light field energy.

The separable graph transform based prediction takes advantage of the natural structure of the reference view as we can see in Figure 3. It is thus efficiently coded using intra-prediction tools. For the non-separable graph prediction, however, this is not the case since the optimal sampling does not totally guarantee that the samples are well structured in each super-ray of the 2D reference view. Yet, the super-ray segmentation preserves in a certain way the natural structure of the reference view.

In our experiments, we choose HEVC intra to encode this information (i.e. top left view or reference view). Tables III and IV give the bit rate obtained when encoding the reference view (from which are derived the DC spatio-angular frequency coefficients) with HEVC-Intra (with QP set to 0). The bit rates are compared with those obtained when using a simple arithmetic coder for directly encoding the spatio-angular DC coefficients. In order to apply the arithmetic coder For each frequency band b, we first group all the coefficients $\hat{\mathbf{x}}_k^b(1)\ \forall k$ of the super-rays in which this band exists, and we code them with an arithmetic coder independently of the other bands. The table shows the rate gain obtained by encoding the set of reference samples with HEVC intra, thanks to the possibility to capture dependencies between super-rays.

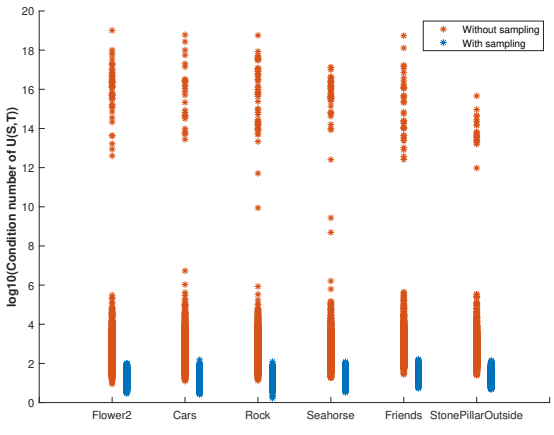


Fig. 9: Efficiency of the sampling and effect on the condition number of the matrix $\mathbf{U}_k(\mathcal{S},\mathcal{T})$. We show for each dataset, for all super-rays the log base 10 of the condition number without (red) and with (blue) sampling.

3) Robustness of the Prediction: In order to assess the efficiency of our prediction and the light field sampling algorithm for the non separable case, we plot in Figure 9 the condition

Light Fields	Energy Percentage in $\hat{x}_k(\mathcal{T}_c)\ \forall\ k$	Energy Percentage in $\hat{\mathbf{x}}_k^b(1)\ \forall\ k, b$
Flower 2	99.15 %	99.02 %
Cars	99.27 %	99.34 %
Rock	98.63 %	98.45 %
Seahorse	99.17 %	98.73 %
Stone Pillars Inside	98.90 %	98.26 %
Friends	99.76 %	99.80 %

TABLE II: Percentage of energy residing in the DC spatio-angular bands $\hat{x}_k(\mathcal{T}_c)\ \forall\ k$ in the non separable case (left column), and in the bands $\hat{\mathbf{x}}_k^b(1)\ \forall\ k, b$ in the separable case (right column).

TABLE III: Bit rate obtained, in the case of the non separable graph transform, when using HEVC intra-coding (left column) of the reference set of samples, and entropy (arithmetic) coding of all the DC spatio-angular bands $\hat{\mathbf{x}}_k(\mathcal{T}_c)\ \forall\ k$ (right column).

Light Fields	HEVC-Intra coding of set of reference samples	Entropy coding of $\hat{\mathbf{x}}_k(\mathcal{T}_c) \forall k$
Flower 2	1.23 Mbits	1.59 Mbits
Cars	1.45 Mbits	1.64 Mbits
Rock	1.31 Mbits	1.49 Mbits
Seahorse	0.74 Mbits	1.38 Mbits
Stone Pillars Inside	1.92 Mbits	1.84 Mbits
Friends	1.86 Mbits	2.07 Mbits

TABLE IV: Bit rate obtained, in the case of the separable graph transform, when using HEVC intra to code the first view (left column), and when using entropy (arithmetic) coding of all the DC spatio-angular bands $\hat{\mathbf{x}}_k^b(1)\ \forall\ k, b$ (right column).

Light Fields	HEVC-Intra of reference view	Entropy coding of $\hat{\mathbf{x}}_k^b(1)\ \forall\ k, b$
Flower 2	1.0 Mbits	1.48 Mbits
Cars	1.06 Mbits	1.54 Mbits
Rock	1.02 Mbits	1.38 Mbits
Seahorse	0.72 Mbits	1.22 Mbits
Stone Pillars Inside	1.81 Mbits	1.66 Mbits
Friends	1.62 Mbits	2.04 Mbits

number in log base 10 of the matrix $\mathbf{U}_k(\mathcal{S}, \mathcal{T})$ for all super-rays k in all the datasets. The condition number is measured to show how much sensible is our prediction in Equation (13) $\hat{\mathbf{x}}_k(\mathcal{T})$ to a small change in $\Big(\mathbf{x}_k(\mathcal{S}) - \mathbf{U}_k(\mathcal{S}, \mathcal{T}_c)\hat{\mathbf{x}}_k(\mathcal{T}_c)\Big)$. On one hand, the condition numbers are computed without sampling i.e. assuming that the reference samples are those in the top-left view. These are shown in red, while the results in blue correspond to the condition numbers after taking the actual samples found with algorithm 1. A major difference is shown in log scale, where the sampling has reduced the condition number from 10^{15} to a maximum of around 10^2. Without sampling, the prediction fails since a tiny change in the high frequency coefficients (even a small rounding procedure) can result in a huge loss in the reconstruction quality.

For the prediction based on the separable graph transform, we do not need a matrix inversion. We only need to invert a number $\mathbf{V}_k^b(1,1)$ whose minimum corresponds to $1/\sqrt{M \times N}$. This inversion does have a smaller impact than the one in the non separable case. This is a major explanation of the PSNR difference between our schemes for a fixed quantization step size $Q = 1$ in Table V.

B. Performance comparison against state of the art coders

We assess the proposed graph-based spatio-angular prediction methods in the context of quasi-lossless light field coding in comparison with a complete HEVC-based scheme with a QP set to 0 and a GOP size of 4. We also compared to an independent coding of sub-aperture images with JPEG2000 [44]. The HEVC version used in the tests is the HM-16.10. The light fields are coded following a raster scan order starting with the top-left view as a reference Intra-coded frame. As additional experiments, we have compared our methods against JPEG-Pleno VM 1.1 ([45] [30]) and the MV-HEVC adapted to light fields proposed in [46] for two datasets *Friends* and *StonePillarsInside* from the JPEG Pleno Light Field datasets [42], [43] with the common test conditions. Note that we compared to those two datasets for which we have obtained the optimized configuration files from the authors.

Results are reported in Tables V and VI where we compare mainly the bit rate needed to code a light field in a quasi-lossless setting (we consider a PSNR higher that 50 dB as a quasi-lossless compression). A substantial gain in bit rate is observed with the proposed graph-based schemes when compared with the MV-HEVC and JPEG-Pleno methods, while preserving a high quality of the reconstructed light fields. This can be justified by the efficiency of the proposed spatio-angular graph transforms in terms of energy compaction along with the ability of HEVC-intra to effectively exploit spatial correlation in the reference view.

For both proposed methods, the disparity information represents around 0.06% of the total bit-streams. For the non separable case, on average, the coded coefficients represent around 95% of the total bit-stream along with 4.7% for the reference view and 0.3% for the segmentation map. As For the separable case, 97% of the total bitrate is allocated to the coefficient and 2.7% for the reference view. The higher bitrate percentage for the reference samples in the non separable case is due to the fact that the reference view is more noisy (see Figure 3). Higher bitrate percentage is assigned to the high frequency coefficients in the separable case. This can be explained by the fact that the energy compaction is not optimal.

TABLE V: Rate comparison between our proposed schemes (with both non separable (NS) and separable (S) graph transforms),HEVC-inter (raster scan) used to code the views in a raster scan order and JPEG 2000, at high quality (PSNR > 50 dB)

Light Fields	HEVC-Inter (QP=0) Raster Scan	JPEG 2000	NS (Q = 0.5)	NS (Q = 1)	S (Q = 1)
Flower 2	3.3129 bpp (54.2 dB)	2.6622 bpp (52.7 dB)	2.4470 bpp (60.4 dB)	2.4457 bpp (52.9 dB)	2.4799 bpp (55.2 dB)
Cars	3.6688 bpp (54.1 dB)	2.6625 bpp (51.9 dB)	2.7759 bpp (60.5 dB)	2.7801 bpp (53.0 dB)	2.6258 bpp (55.2 dB)
Rock	3.2700 bpp (53.8 dB)	2.6602 bpp (53.2 dB)	2.0423 bpp (60.3 dB)	2.0545 bpp (52.6 dB)	2.0162 bpp (54.7 dB)
Seahorse	2.4751 bpp (54.4 dB)	2.6433 bpp (56.8 dB)	1.8224 bpp (60.4 dB)	1.7849 bpp (53.0 dB)	1.9762 bpp (55.3 dB)
Stone Pillars Inside	4.9017 bpp (52.1 dB)	2.6650 bpp (50.5 dB)	2.5559 bpp (59.7dB)	1.5269 bpp (52.4 dB)	3.3094 bpp (55.0 dB)
Friends	3.5400 bpp (52.8 dB)	2.6640 bpp (52.1 dB)	1.9327 bpp (59.7 dB)	1.9311 bpp (52.4 dB)	2.4436 bpp (54.8 dB)

TABLE VI: Rate comparison between our proposed schemes (with both non separable (NS) and separable (S) graph transforms), Multi-view HEVC for light fields (MV-HEVC-LF) in [46] and JPEG Pleno VM 1.1 at high quality (PSNR > 50 dB)

Light Fields	MV-HEVC-LF	JPEG Pleno	NS (Q = 0.5)	NS (Q = 1)	S (Q = 1)
Stone Pillars Inside	5.694 bpp (56.7 dB)	4.9220 bpp (56.7 dB)	2.5559 bpp (59.7dB)	1.5269 bpp (52.4 dB)	3.3094 bpp (55.0 dB)
Friends	3.057 bpp (55.5 dB)	2.4023 bpp (54.2 dB)	1.9327 bpp (59.7 dB)	1.9311 bpp (52.4 dB)	2.4436 bpp (54.8 dB)

C. Computational complexity analysis

We can first note that, for both proposed schemes, the time needed to encode the disparity information and the segmentation map is negligible, given their small size. The time needed for the arithmetic encoding of the transform coefficients is also negligible. If we first consider the non separable graph based compression scheme (depicted in Figure 4), its computational complexity essentially results from the following three steps: the diagonalization of the Laplacians (dimensions in the order 10^3) to find the transform basis functions, the sampling algorithm partly based on the basis functions computed before, and the application of the graph transform on the signals. First, the complexity for the diagonalization of a Laplacian with dimensions $(N \times N)$ is $\mathcal{O}(N^3)$. Second, the sampling algorithm 1 has an overall complexity of $\mathcal{O}(|\mathcal{S}|^4 + N|\mathcal{S}|^2)$ with N being the number of vertices in a super-ray and $\mathcal{S}$ the number of samples, which is far less than the complexity of the diagonalization of the Laplacian since the considered scenario is for $|S| << N$. The complexity of the graph transform application is $\mathcal{O}(N^2)$. When using the separable transform (in Figure 5), the complexity can be significantly reduced as the dimensions of the Laplacians are of the order 10^2 instead of 10^3. Also, in the separable case, no sampling algorithm is needed.

In addition, as each super-ray can be processed independently from others, the proposed methods are well suited for parallelization. While the sequential processing of 3600 super-rays requires 7 hours (with the separable transform), the use of GPU and fast parallel computing libraries such as *MAGMA* [47] (with up to 16 parallel laplacian diagonalizations), allows reducing the time to 30 minutes. The running time for the available matlab implementation of the compared JPEG-Pleno solution [30] is 15 minutes. Although graph-based transforms are more complex than classical wavelet transforms used in JPEG-Pleno, we would like to point out the fact that fast graph transforms that allow reducing the diagonalization and transform running time have been recently proposed in [48] and [49]. The authors show that the transform computation time can be reduced by a factor of up to 27 and that the diagonalization complexity can also be traded against a small alteration of the basis functions.

VII. Conclusion

In this paper, we have proposed sampling and prediction methods with local graph transforms for light field energy compaction and compression. Based on the graph sampling theory, the proposed methods allow taking advantage of the good energy compaction property of the graph transform on local supports with a limited complexity, while benefiting from well established but powerful prediction mechanisms in the pixel domain. We considered both a super-ray based non separable graph transform and a spatio-angular separable simplified version. Two coding schemes have been described based on the non separable and separable graph transforms. The schemes have been assessed for high quality (quasi-lossless) coding. Both proposed approaches are very efficient when the quantization noise on the reference set of samples is low, hence for quasi-lossless compression. If the reference set of samples is too coarsely quantized, drift and noise amplification may appear during the prediction step. This is due to the fact that, in Equation (15), the prediction uses the spatial transform coefficients estimated on the reference set of samples available at the decoder side. Further study will be dedicated to addressing this problem in the case of lossy compression.

References

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[2] G. Shen, W.-S. Kim, S. K. Narang, A. Ortega, J. Lee, and H. Wey, "Edge-adaptive transforms for efficient depth map coding," in *Picture Coding Symposium (PCS), 2010*. IEEE, 2010, pp. 566–569.

[3] W. Hu, G. Cheung, X. Li, and O. Au, "Depth map compression using multi-resolution graph-based transform for depth-image-based rendering," in *2017 IEEE International Conference on Image Processing ICIP*, Sept. 2012.

[4] W.-S. Kim, S. K. Narang, and A. Ortega, "Graph based transforms for depth video coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 813–816.

[5] C. Zhang and D. Florêncio, "Analyzing the optimality of predictive transform coding using graph-based models," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 106–109, 2013.

[6] W. Hu, G. Cheung, A. Ortega, and O. C. Au, "Multiresolution graph fourier transform for compression of piecewise smooth images," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 419–433, 2015.

[7] W.-T. Su, G. Cheung, and C.-W. Lin, "Graph fourier transform with negative edges for depth image coding," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1682–1686.

[8] G. Cheung, E. Magli, Y. Tanaka, and M. Ng, "Graph spectral image processing," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 907–930, 2018.

[9] M. Hog, N. Sabater, and C. Guillemot, "Super-rays for efficient light field processing," *IEEE J. on Selected Topics in Signal Processing, special issue on light field image processing*, Oct. 2017.

[10] M. Rizkallah, T. Maugey, and C. Guillemot, "Graph-based spatio-angular prediction for quasi-lossless compression of light fields," in *Data Compression Conference, DCC*, 2019.

[11] S. K. Narang, A. Gadde, E. Sanou, and A. Ortega, "Localized iterative methods for interpolation in graph structured data," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 491–494.

[12] D. E. Tzamarias, P. Akyazi, and P. Frossard, "A novel method for sampling bandlimited graph signals," in *Proceedings of EUSIPCO*, no. CONF, 2018.

[13] I. Pesenson, "Sampling in paley wiener spaces on combinatorial graphs," *Transactions of the American Mathematical Society*, vol. 360, no. 10, pp. 5603–5627, 2008.

[14] C. Conti, P. Nunes, and L. D. Soares, "New hevc prediction modes for 3d holoscopic video coding," in *2012 19th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2012, pp. 1325–1328.

[15] ——, "Hevc-based light field image coding with bi-predicted self-similarity compensation," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–4.

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

[17] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2016*. IEEE, 2016, pp. 1–4.

[18] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.

[19] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on focus stack and extended focus images," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 898–902.

[20] C. Jia, Y. Yang, X. Zhangy, X. Zhang, S. Wangx, S. Wang, and S. Ma, "Optimized inter-view prediction based light field image compression with adaptive reconstruction," in *2017 IEEE International Conference on Image Processing ICIP*, 2017.

[21] W. Ahmad, R. Olsson, and M. Sjostrom, "Interpreting plenoptic images as multiview sequences for improved compression," in *2017 IEEE International Conference on Image Processing ICIP*, 2017.

[22] 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, 2017.

[23] E. Dib, M. L. Pendu, and C. Guillemot, "Super-ray based low rank approximation for light field compression," in *Data Compression Conference, DCC*, 2019.

[24] R. Verhack, T. Sikora, L. Lange, R. Jongebloed, G. Van Wallendael, and P. Lambert, "Steered mixture-of-experts for light field coding, depth estimation, and processing," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1183–1188.

[25] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4562–4566.

[26] X. Jiang, M. Le Pendu, and C. Guillemot, "Light fields compression using depth image based view synthesis," in *Hot3D workshop held jointly with IEEE Int. Conf. on Multimedia and Expo, ICME*. IEEE, July 2017.

[27] X. Su, M. Rizkallah, T. Maugey, and C. Guillemot, "Graph-based light fields representation and coding using geometry information," in *IEEE International Conference on Image Processing (ICIP)*, 2017.

[28] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 193, 2016.

[29] I. Tabus, P. Helin, and P. Astola, "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and jpeg 2000," in *Image Processing (ICIP), 2017 IEEE International Conference on*, 2017, pp. 4567–4571.

[30] I. J. S. JPEG, "Jpeg pleno light field coding vm 1.1," Doc. N81052, 2018.

[31] C. Galea and C. Guillemot, "Denoising of 3d point clouds constructed from light fields," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1882–1886.

[32] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li *et al.*, "Emerging mpeg standards for point cloud compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2018.

[33] L. H. Z. J. Y. Xu, K. Zhang and W. Zhu, "Introduction to point cloud compression," *Zte Communications*, vol. 16, no. 3, Sep. 2018.

[34] L. He, W. Zhu, and Y. Xu, "Best-effort projection based attribute compression for 3d point cloud," in *2017 23rd Asia-Pacific Conference on Communications (APCC)*. IEEE, 2017, pp. 1–6.

[35] C. Perra, "Lossless plenoptic image compression using adaptive block differential prediction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 1231–1234.

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

[37] S. Gortler, R. Grzeszczuk, R. Szeliski, and M. Cohen, "The lumigraph," in *Proc. SIGGRAPH*, 1996, pp. 43–54.

[38] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. on Computer Vision, ICCV*, 2003, pp. 10–17.

[39] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[40] X. Jiang, M. Le Pendu, and C. Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views," in *IEEE International Conference on Image Processing (ICIP)*, 2018.

[41] I. Daribo, G. Cheung, and D. Florencio, "Arithmetic edge coding for arbitrarily shaped sub-block motion prediction in depth video compression," in *2012 19th IEEE International Conference on Image Processing*, Sept 2012, pp. 1541–1544.

[42] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, no. CONF, 2016.

[43] "Jpeg pleno light field dataset." [Online]. Available: https://jpeg.org/plenodb/lf/epfl/

[44] [Online]. Available: https://jpeg.org/jpeg2000/

[45] [Online]. Available: https://jpeg.org/jpegpleno/

[46] W. Ahmad, R. Olsson, and M. Sjostrom, "Towards a generic compression solution for densely and sparsely sampled light field data," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 654–658.

[47] [Online]. Available: http://icl.cs.utk.edu/magma/

[48] L. Le Magoarou, R. Gribonval, and N. Tremblay, "Approximate fast graph Fourier transforms via multi-layer sparse approximations," *IEEE transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 407–420, Jun. 2018. [Online]. Available: https://hal.inria.fr/hal-01416110

[49] K. Lu and A. Ortega, "Fast graph fourier transforms based on graph symmetry and bipartition," *IEEE Transactions on Signal Processing*, vol. 67, no. 18, pp. 4855–4869, Sep. 2019.