IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 30, 2020, accepted February 13, 2020, date of publication March 2, 2020, date of current version March 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977767

Dense Light Field Coding: A Survey

CAROLINE CONTI, (Member, IEEE), LUÍS DUCLA SOARES, (Senior Member, IEEE), AND PAULO NUNES, (Member, IEEE)
Instituto de Telecomunicações, Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisboa, Portugal

Corresponding author: Caroline Conti (caroline.conti@lx.it.pt)

This work was supported by FCT/MCTES through national funds and when applicable co-funded EU funds through the Project under Grant UIDB/EEA/50008/2020.

ABSTRACT Light Field (LF) imaging is a promising solution for providing more immersive and closer to reality multimedia experiences to end-users with unprecedented creative freedom and flexibility for applications in different areas, such as virtual and augmented reality. Due to the recent technological advances in optics, sensor manufacturing and available transmission bandwidth, as well as the investment of many tech giants in this area, it is expected that soon many LF transmission systems will be available to both consumers and professionals. Recognizing this, novel standardization initiatives have recently emerged in both the Joint Photographic Experts Group (JPEG) and the Moving Picture Experts Group (MPEG), triggering the discussion on the deployment of LF coding solutions to efficiently handle the massive amount of data involved in such systems. Since then, the topic of LF content coding has become a booming research area, attracting the attention of many researchers worldwide. In this context, this paper provides a comprehensive survey of the most relevant LF coding solutions proposed in the literature, focusing on angularly dense LFs. Special attention is placed on a thorough description of the different LF coding methods and on the main concepts related to this relevant area. Moreover, comprehensive insights are presented into open research challenges and future research directions for LF coding.

INDEX TERMS Camera array, image compression, light field, plenoptic, video compression.

I. INTRODUCTION

Light Field (LF) imaging is a promising solution for providing more immersive and closer to reality multimedia experiences to end-users with unprecedented creative freedom and flexibility for applications in different areas, such as Virtual Reality (VR) and Augmented Reality (AR) [1], cinematography [2], three dimensional (3D) television [3], [4], biometric recognition [5], and medical imaging [6]. Due to the recent technological advances in optics, sensor manufacturing and available transmission bandwidth, the research in richer imaging technologies has accelerated and practical designs of novel LF acquisition [2], [7] and LF display [3], [4], [8] systems have been rapidly maturing. These devices allow sampling the light rays from every direction through every point in a volume of space and allow using this information to recreate the correct perspective of the scene from any viewpoint position inside this volume. This also enables a variety of post-capture processing capabilities [9], [10] such as refocusing, changing depth-of-field, extracting depth/disparity information, and 3D modeling.

Many tech giants [11], [12] investing in VR have pointed towards the power of LFs for achieving ultra-realistic VR experiences [13], in which a user can freely walk into the scene with head motion parallax — also called Six Degrees of Freedom (6DoF) movement — and realistically perceive every little detail about the materials and lighting, such as shifting reflections and translucence. Following these trends, two prototype systems have been recently proposed by Facebook [11] and Google [12] for capturing, processing and rendering LFs for VR media applications. In fact, given these trends, it is wise to expect that soon many LF systems will be available to both consumers and professionals [14]–[17]. Moreover, there is a recent advent of high-quality consumer Head Mounted Displays (HMD) with positional head tracking — such as HTC Vive, Oculus Rift, and Windows Mixed Reality — that provide new and compelling opportunities for visualizing LFs in new commercial media applications. To cite just a few, in the area of experimental education, 6DoF VR simulations [18] have the potential to effectively

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

train students in an environment that might not have been available to them before, enabling them to reach their full potential. Additionally, in health care, recent researches [19] have shown the benefits of using true-to-life VR exposure therapy for the treatment of a variety of mental health disorders, such as phobias and social anxiety.

Recognizing the potential of this technology, novel standardization initiatives have recently emerged. Among the requirements being discussed, deploying LF coding solutions to efficiently handle the massive amount of data involved in such systems is one of utmost importance [20]. Notably, the Joint Photographic Experts Group (JPEG) committee has launched the JPEG Pleno standardization initiative [10] that addresses representation and coding of emerging imaging modalities and aims at providing the highest possible compression efficiency trade-off given a set of advanced functionalities, such as [21]: i) spatial random access; ii) low latency and real time processing; and iii) scalability (e.g., depth resolution, viewing angle range, etc.). In addition, the Moving Picture Experts Group (MPEG) has started a new work item on coded representations for immersive media (MPEG-I) [22], focusing on the usage of emerging imaging technologies in applications that provide an increased sense of immersion, such as VR applications [22]. In terms of visual representations, MPEG-I comprises a set of standards that should be developed in several phases so as to support a growing number of DoF in virtual walkthroughs in a bounded volume of space [22]. With the emergence of these standardization initiatives [10], [22], the topic of LF content coding has recently became a very active and relevant research area, attracting the attention of many researchers worldwide and being the topic of many special sessions in international journals and conferences. Notably, JPEG Pleno organized two grand challenges on light field coding at the IEEE International Conference on Image on Multimedia and Expo (ICME) in 2016 [23], and at the IEEE International Conference on Image Processing (ICIP) in 2017 [24].

Although the topic of LF content coding only recently became a booming research area, various LF coding solutions have already been proposed in the literature in the last two decades, following the advancement of digital multimedia systems and coding technologies. However, despite the relevance of the topic, there have been only a few surveys about LF processing and coding recently published in the literature. In [25], the authors review the developments and trends on LF imaging, mainly focusing on acquisition, calibration and pre-processing techniques for lenslet-based LF cameras. In [26], a brief overview of the main research directions in relation to some critical problems in LF processing is presented, focusing on a few relevant solutions proposed in the literature for coding, super-resolving and editing LFs. In [27], an overview of LF imaging and processing is presented, covering techniques for LF acquisition, super-resolution, depth estimation, coding, editing, rendering, and user interaction applications. It is important to notice that all previous works adopt a broader review perspective, covering many different aspects of LF imaging and processing, but do not include a comprehensive review of LF coding techniques. Moreover, most of them do not include explanations of technical details of individual LF coding solutions and a comprehensive discussion about the most relevant results, advantages, and limitations.

Aiming to fill this gap, this paper provides a comprehensive survey of the most relevant lossy coding solutions for LF content proposed in the literature in the last 25 years. This survey paper is mainly intended to assist readers who wish to begin research in the area of LF processing and coding. To accomplish this, special emphasis is placed on a thorough description of the different LF coding methods and on the main concepts and challenges related to this relevant area. In summary, the main contributions of this survey paper are:

1) To better assist new researchers in this area, a brief overview of the principles of LF imaging technology and of LF processing is presented. This overview covers the main stages that are essential for delivering LF content to end-users, including the recent developments related to LF acquisition, representation, rendering, display, and visual quality evaluation. Moreover, challenges in each stage are also discussed and references to seminal works are provided.
2) A comprehensive analysis of LF coding solutions is provided. For this, the LF coding solutions in the literature are carefully categorized in terms of the adopted representation format, coding architecture, and technique for achieving compression.
3) Encompassing insights into future directions for LF coding are presented, considering recent work found in the literature and recent standardization activities.

A. BASIC CONCEPTS

The term light field was firstly adopted by Gershun in 1936 for analytically describing the ''light beams that propagate in a straight line through a homogeneous medium and which is the carrier of radiant energy in a space'' [28]. Generally, the light field concept comes from the necessity of describing and replicating the complete/full/whole visible light information in a given surrounding as accurately as possible. Actually, this notion had been exploited earlier by the polymath Da Vinci (1452–1519) when he suggested the existence of the pyramids of sight [29], as well as by the physics Nobel laureate Lippmann, in 1908, when he introduced the concept of integral photography [30]. More recently, the terms holoscopic (from the Greek hólos (whole) + optikos (vision)) [31], plenoptic (from the Latin plenus (full) + opticus (vision)) [32], and lumigraph [33] have been also adopted almost as synonymous.

In the early 1990s, with the popularity growth of computer vision and computer graphics, the problem of geometrically describing the visible space gained a major importance. Notably, Adelson and Bergen proposed, in [32], to define the total distribution of light as a seven dimensional (7D) function, which models the light rays at every possible location

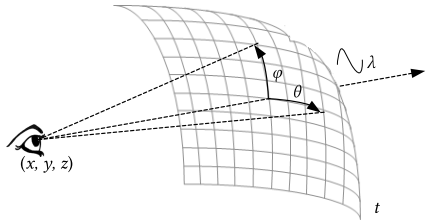


FIGURE 1. The 7D plenoptic function.

in space (x, y, z), toward every possible direction (θ, φ), over any range of wavelengths (λ), and at any time (t) — referred to as the plenoptic function (see Fig. 1):

$$P(x, y, z, \theta, \varphi, \lambda, t) \tag{1}$$

With this definition, it is possible to model different imaging systems, including the Human eye [32], as samplings of the 7D plenoptic function in (1). However, due to the enormous amount of data that would be required for sampling using a 7D representation, it is necessary to make reasonable assumptions to reduce the dimensionality of the plenoptic function and to appropriately sample it.

Therefore, Levoy and Hanrahan proposed, in [34], to use the following three assumptions to reduce from the 7D plenoptic function in (1) to a four dimensional (4D) function:

1) **Static Scene** — Assuming the scene is static, the plenoptic function can be then reduced to $P(x, y, z, \theta, \varphi, \lambda)$.
2) **Constant Radiance along its Path (Free-Space)** — With the assumption that the air is truly transparent and the light ray is transmitted in a free-space (i.e., region free of occluders [34]), the plenoptic function can be then represented by its values along an arbitrary selected surface surrounding the scene (see Fig. 1). Hence, the radiance of any light ray in the space can be always obtained by tracing it back to this selected surface. This assumption allows reducing the plenoptic function to $P(x, y, \theta, \varphi, \lambda)$.
3) **Trichromatic Human Vision System (HVS)** — The Human eye has three types of photosensitive cells (known as cones) in the retina for the perception of colored light. Each of these cone types has its maximum sensitivity in a different wavelength, which corresponds to the primary colors Red (R), Green (G), and Blue (B). Therefore, it is possible to restrict to the HVS and to reduce the wavelength dimension in (1) by assuming three different plenoptic functions (one for each R, G, and B components). Finally, for each color component, a 4D plenoptic function is defined as $P(x, y, z, \theta, \varphi)$.

Moreover, it is common to use a two-plane parameterization to represent the 4D plenoptic function in Cartesian coordinates. In this case, and as illustrated in Fig. 2, a specific light ray intersects the first plane at coordinates (x, y), which defines the spatial location of the ray, and it is propagated in free-space until it intersects the second plane at coordinates (u, v), which specifies the propagation direction. Levoy and Hanrahan [34] baptized this 4D function, $L(x, y, u, v)$, as the 4D light field (a.k.a, lumigraph [33]).

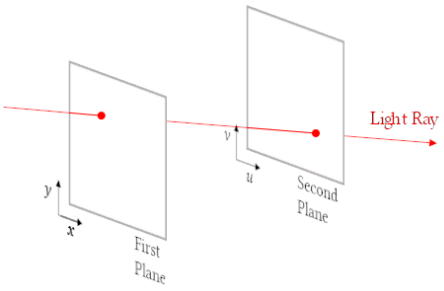


FIGURE 2. Two-plane parameterization for the 4D plenoptic function.

In the context of this paper, it will suffice to describe the complete visible light by this 4D LF function $L(x, y, u, v)$ [34]. This means that light is here understood as a scalar radiance (one value for each color component R, G, and B) traveling along straight lines (rays) with different propagation directions. This 4D LF function will be then used to represent a LF image. Additionally, a LF video can also be represented by regularly sampling LF images per unit of time.[1] Moreover, this paper will particularly focus on reviewing the literature on coding solutions for dense LFs. While a better definition will be given later in the paper, the term dense LF stands for angularly dense LF, and angular density can be understood as the number of viewpoints that are sampled by a LF imaging system in a volume of space.

B. OUTLINE

The remainder of this paper is organized as follows. Section II briefly reviews the main stages that are essential for efficiently delivering dense LF content to the end-users, including the recent developments related to LF acquisition, representation, rendering, display, and visual quality evaluation. Section III focuses on the coding requirements and reviews the most relevant coding solutions for dense LF content in the literature, while simultaneously deriving some conclusions and highlighting some of the remaining challenges. Finally, Section IV concludes the paper.

II. LIGHT FIELD IMAGING AND PROCESSING

Before concentrating on the LF coding solutions available in the literature, this section provides a brief review on the principles of LF imaging technology and on the main stages that are essential for delivering LF content to end-users.

Essentially, an LF imaging processing chain comprises the following functional stages, as illustrated in Fig. 3:

- **LF Acquisition/Creation** — The first step of the LF processing chain is, naturally, the LF content generation, which can be done through an optical setup or computationally created, i.e., through appropriate modeling and rendering of the visual scene and the acquisition setup.

[1] In this case, light can be then described as a 5D function $L(x, y, u, v, t)$.

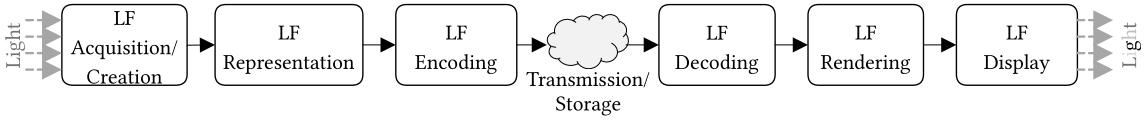


FIGURE 3. LF imaging processing flow [35].

Therefore, Section II-A reviews the principles of LF imaging acquisition and lists some publicly available LF datasets for LF coding.

- **LF Representation** — The LF data acquired in the previous stage may or may not be converted to a representation format that is different from the acquired format. In this context, Section II-B presents a brief review of representation formats that have been proposed in the literature.
- **LF Coding** — Considering the huge amount of data associated to LF transmission systems, efficient LF encoding/decoding solutions become of paramount importance. Since this is the main focus of this paper, Section II-C presents a brief overview of possible LF coding architectures, and a more detailed survey on LF coding approaches proposed in the literature is presented in Section III.
- **LF Rendering** — Rendering the decoded LF content becomes also an important issue, especially to allow adequate visualization of the decoded LF content in conventional two dimensional (2D) and 3D displays. Therefore, Section II-D addresses this issue and reviews some LF rendering algorithms and their capabilities.
- **LF Display** — To take full advantage of the richer visual information of the acquired LF content, new and more immersive display devices are also needed. For this reason, display technologies have been also evolving in recent years, and Section II-E overviews those developments.

In addition to this, the display technology along with the rendering capabilities will also determine what should be expected in terms of the user experience in LF imaging applications. This fact brings up another challenging issue, which is designing appropriate objective and subjective metrics for LF quality evaluation. This issue is then briefly reviewed in Section II-F.

A. LF ACQUISITION

With the target of increasing immersion, more advanced imaging technologies are emerging that allow capturing richer forms of visual data and representing the scene by the 4D light field. Different acquisition techniques can be used to capture LF content with different densities in each of the 4D dimensions, depending on the requirements for spatial or angular resolution. Generally, angularly dense LFs allow a smooth transition between viewpoints without the need for view interpolation. Another parameter that may differ depending on the LF capturing system is the Field of View (FoV) that corresponds to the area of the scene over which objects can be reproduced.

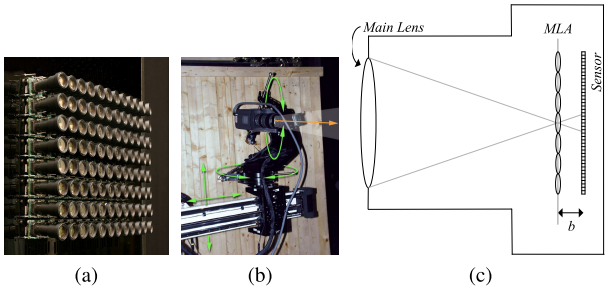


FIGURE 4. Examples of LF acquisition systems: (a) Multi-camera array (From [36]); (b) Camera gantry (From [39]); and (c) Lenslet Camera.

Among the possible techniques for acquiring LFs, three main groups may be listed [14]:

- **LF Imaging with Multi-Camera Array** — In this case, a number of views with parallax on a single or on both directions (full parallax) is captured using an array of multiple cameras (see Fig. 4a) in a linear [36], circular [2], or even arbitrary arrangement [37]. The spatial density depends on the camera's sensor resolution and the angular density depends on the distance between the cameras (baseline) on the array. For acquiring (angularly) dense LFs, the baseline is usually on the order of millimeters to centimeters [38]; this setup is usually referred to as a High Density Camera Array (HDCA). The FoV that is captured depends on the number of cameras and their optical setup, the baseline and the rotation between the cameras. Usually, the camera array configuration is designed targeting the requirements of a specific use case scenario. An example of such LF acquisition system is the Stanford multi-camera array proposed in [36], which is composed of 100 custom cameras, and supports reconfigurable arrangement of the array for three different scenarios (i.e., for panoramic video with high dynamic range; for synthetic aperture photography; and for widely spaced 3D scenes).
- **LF Imaging with Camera Gantry** — In this case, a moving camera gantry (see Fig. 4b) is used to capture different viewpoint images at different instants of time. The spatial density depends on the camera's sensor resolution and the angular density depends on the accuracy of the gantry motion. The FoV depends on the degrees of freedom that is supported by the gantry structure. Examples of LF camera gantry systems are the Stanford LF gantry [39] and the Fraunhofer robot system [40],

both with four degrees of freedom camera movements (translation along x and y and rotation along θ and φ axes). Although camera gantries provide flexible and lower cost capturing systems, they are restricted for capturing static scenes.

- **LF Imaging with Lenslet Camera** — In this case, the integral photography concept proposed by Lippman [30] is adopted, in which LF content with full parallax can be acquired by using a single-tier sensor camera overlaid with a Microlens Array (MLA), as shown in Fig. 4c. The MLA (a.k.a. lenslet array) can be seen as a tiny 2D array of cameras with a very small baseline, sampling the 4D light field and organizing it in a conventional 2D image, known as the lenslet image. As discussed in [41], [42], there are two lenslet camera setups, namely: the unfocused (a.k.a. plenoptic camera 1.0), as used in the commercial Lytro cameras [2], [9]; and the focused (a.k.a. plenoptic camera 2.0), as used in the commercial Raytrix cameras [7]. The difference between these two setups is in the distance b of the MLA to the image sensor (see Fig. 4c). For the unfocused setup, the distance b is equal to the MLA focal length f (i.e., $b = f$), while $b \neq f$ for the focused setup. In practice, varying between these two setups will only change the balance between providing larger angular or spatial density in the captured LF (respectively, using an unfocused or a focused lenslet camera) [41], [43]. In the unfocused lenslet camera, the light of a single ray (or of a thin bundle of rays) from a given angular direction (θ, φ) converges on a specific microlens at position (x, y) in the array and is collected at a single pixel position in the image sensor underneath. Hence, the angular density depends on the number of pixels behind each microlens, and the spatial density depends on the number of microlenses in the MLA. In the focused lenslet camera, the closer distance b is to f, the larger is the angular density (and vice-versa). As in a conventional 2D camera, the FoV is defined by the camera's sensor size and the main lens focal length. Although lenslet cameras can capture highly dense LFs at a single shot, they are specially recommended for capturing objects at small distances due to the their small FoV.

Table 1 presents a list of publicly available datasets for dense LFs in all of the three categories.

It is also possible to combine these techniques to meet some specific requirements for spatial and angular density. For instance, in [55], a gantry structure with a lenslet camera is used for capturing LFs with a baseline varying between micrometers (inside the lenslet camera) to meters (by varying the camera poses). Additionally, in [12], a one dimensional (1D) array of GoPro cameras in a vertical arc are placed in a horizontally rotating gantry structure to capture dense LFs with 360° FoV. Moreover, depth cameras [58] and Light Detection And Ranging (LIDAR) can also be combined with the above LF techniques to acquire geometry information of the scene that may be useful for rendering virtual viewpoints.

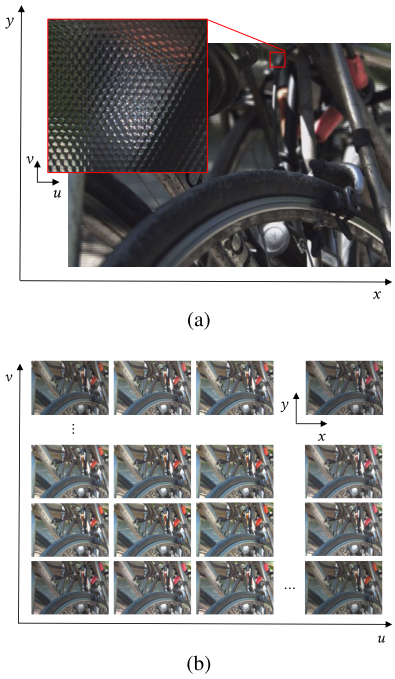


FIGURE 5. Possible acquired LF representation formats: (a) Lenslet format; and (b) Full parallax multiview format.

B. LF REPRESENTATION

Initially, the data acquired by a LF imaging system can have one of the following (raw) formats [59]:

- **Lenslet Representation** — For LF acquired using a lenslet camera, the LF content is represented as a 2D image comprising a 2D grid of microlens images (a.k.a. micro-images, elemental images, and macro-pixels), as depicted in Fig. 5a.
- **Full Parallax Multiview Representation** — For LF acquired using a multi-camera array or a camera gantry, the LF content is represented by a 2D grid of views (see Fig. 5b) and, usually, each view has the same spatial resolution.

Although it is possible to adopt these acquisition formats for representing the LF data, in some cases, it may be necessary to convert from the acquisition format to a more appropriate representation format. It is also possible to convert from one to another and, in some cases, this conversion can be invertible or non-invertible, depending on the type of the camera and on the algorithm used [59]–[61].

In fact, a key issue for successful LF imaging applications is the choice of a convenient representation for the LF data acquired, given a certain set of application requirements. If high compression efficiency is a dominant requirement when choosing the LF representation format, then the decision should be made prioritizing a coding perspective, which means that an efficient coded representation should be at the forefront. In this context, this section briefly describes some relevant LF representation formats that have been

TABLE 1. List of publicly available datasets for dense LF content.

Dataset	Acquisition System	Description	Link
CVIA LF Dataset [44]	Lenslet camera (Lytro Illum) and depth camera	139 LF images and depth map	https://sites.google.com/view/cvia/plenoptic-imaging
DDFF 12-Scene Dataset [45]	Lenslet camera (Lytro Illum) and depth camera	720 LF images and depth maps	https://hazirbas.com/datasets/ddff12scene/
Disney LF Dataset [46]	Camera gantry	5 LF images with high spatio-angular resolution	https://people.csail.mit.edu/changil/light-field-depth/
EPFL LF Dataset [47]	Lenslet camera (Lytro Illum)	5118 LF images in 10 different categories	http://www.epfl.ch/labs/mmspg/EPFL-light-field-image-dataset
Fraunhofer LF Dataset [40]	Camera gantry	9 LF images	https://www.iis.fraunhofer.de/en/ff/amm/dl/lightfielddataset.html
HCI 4D LF Dataset [48]	Synthetic	24 designed scenes	https://lightfield-analysis.uni-konstanz.de/
INRIA LF Dataset [49]	Lenslet camera (Lytro 1st generation, Lytro Illum, and Raytrix R8), and synthetic	More than 100 LF images and 40 LF video sequences	http://clim.inria.fr/DataSoftware.html
InterDigital LF Dataset [50]	Multi-camera array	12 LF video sequences	https://www.interdigital.com/data_sets/light-field-dataset
Matching LF Dataset [51]	Lenslet camera (Lytro Illum and Raytrix R29)	31 LF images captured with the two lenslet cameras from the same viewpoint position	https://doi.org/10.6084/m9.figshare.6115487
MIT Synthetic LF Archive [52]	Synthetic	18 LF images with transparencies, occlusions and reflections	http://web.media.mit.edu/~gordonw/SyntheticLightFields\
MPI LF Archive [53]	Camera gantry, and synthetic	14 LF images (with only horizontal parallax)	http://lightfields.mpi-inf.mpg.de/Dataset.html
SMART LF Dataset [54]	Lenslet camera (Lytro Illum)	16 LF images	http://www.comlab.uniroma3.it/SMART.html
Stanford LF Archive [39]	Camera gantry, multi-camera array, and lenslet-based microscope	22 LF images captured using different acquisition setups	http://lightfield.stanford.edu
Stanford Lytro LF Archive [55]	Lenslet camera (Lytro Illum)	22 LF images captured using different acquisition setups	http://lightfields.stanford.edu/LF2016.html
Stanford Multiview LF Dataset [55]	Gantry structure and 1D array of multiple lenslet cameras (Lytro Illum)	More than 300 images in 9 different categories	http://lightfields.stanford.edu/mvlf/
TUT LF Dataset [56]	1D multi-camera array	6 LF images (with only horizontal parallax)	http://urn.fi/urn:nbn:fi:att:ed60be6d-9d15-4857-aa0d-a30acd16001e
UCSD/MERL LF Repository [57]	Camera gantry, and 1D multi-camera array	9 LF images (with only horizontal parallax)	http://neelj.com/data/lfarchive/

proposed in the literature in the context of LF coding. However, it should be noticed that the analysis and discussion of specific LF coded representations performance will be done in Section III.

1) LENSLET REPRESENTATION

This representation format corresponds to the raw acquisition format of a lenslet camera and, in this case, no conversion and/or further processing is required. Hence, the LF data is represented as a 2D image comprising a grid of micro-images, as depicted in Fig. 5a. As illustrated in Fig. 6, each micro-image captures a low-resolution portion of the scene. Moreover, several packing schemes, shapes and sizes of microlenses are possible in the array (see Fig. 6), and the structure of these micro-images is a consequence of the chosen MLA. In addition to this, the micro-image characteristics may also change depending on the chosen lenslet camera setup (see Fig. 6). For instance, for an unfocused lenslet camera, a micro-image is a picture of the back of the main lens [41] (i.e., it is an image focused on the main lens), while, for a focused lenslet camera, a micro-image is (a low-resolution portion of) the image of the main lens that is relayed through the microlens.

Analyzing this lenslet representation from a coding point of view, it is observed that, independently of the camera setup or MLA used, the LF content presents some inherent spatial correlations, as illustrated by the autocorrelation function in Fig. 7b. Notably, it can be seen that the pixel correlation in a lenslet image is not as smooth as in conventional 2D images (see Fig. 7a). Differently, a regular structure of spikes is evidenced in the autocorrelation function in Fig. 7b, in which the constant distance between these regular spikes corresponds to the micro-image spacing in the array. Moreover, as is commonly observed in 2D images (see Fig. 7a), pixels inside each micro-image are also significantly correlated within a local neighborhood (see Fig. 7b).

2) PSEUDO VIDEO SEQUENCE REPRESENTATION

In this case, viewpoints are stacked together along a pseudo temporal axis to be interpreted as a single Pseudo Video

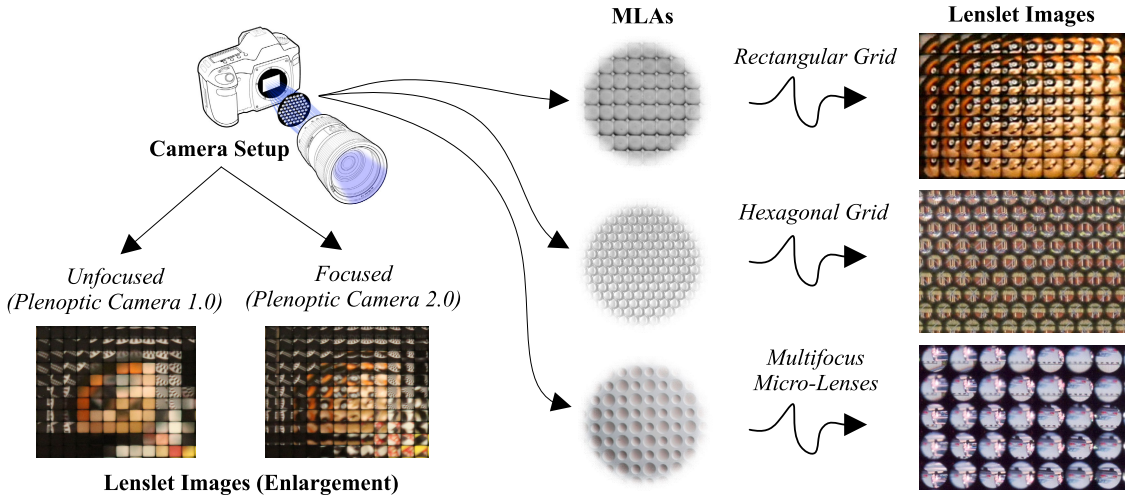


FIGURE 6. Examples of lenslet images captured using different LF camera setups and MLA structures.

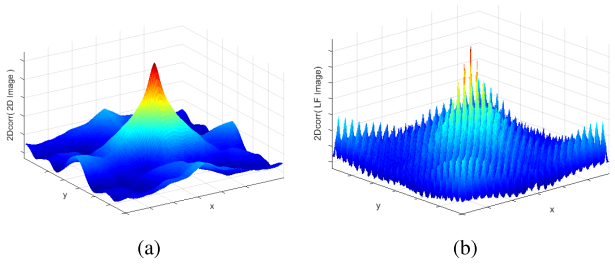
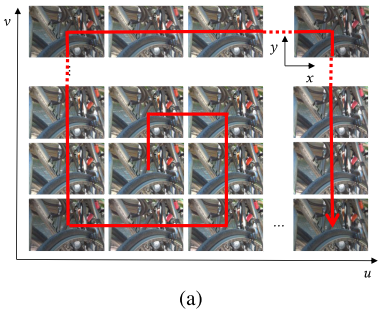


FIGURE 7. Autocorrelation function: (a) 2D image; and (b) Lenslet image.



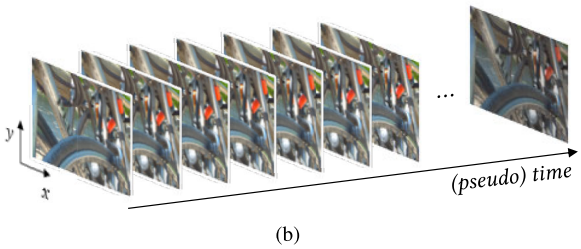


FIGURE 8. PVS representation: (a) Scanning the 2D grid of views in spiral order; and (b) the resulting PVS.

Sequence (PVS), as illustrated in Fig. 8b. If the LF acquisition system supports full parallax, the 2D array of viewpoints needs to be scanned using a specific topology to form a 1D array of views, as depicted in Fig. 8a. Some examples of scanning topologies are illustrated in Fig. 9. The (pseudo) temporal correlation varies depending on the scanning topology used to form the PVS and, consequently, the coding performance will be directly related to this choice.

For representing LF video, the referred to as transposed picture ordering [62] can be used. In this case, all views from the same time instant are concatenated along the time dimension. However, it is worthwhile to note that the temporal correlation between adjacent time instants no longer exists in the final video sequence.

Although for the full parallax multiview acquisition format there is no further pre-processing need prior to coding, for the lenslet acquisition format, it is necessary firstly to convert the lenslet image to a dense array of views. In this context, many approaches have been proposed in the literature for extracting views from a lenslet image:

- **Based on Micro-Images** — In this case, the lenslet image needs to be firstly split into its multiple micro-images, which are then represented as multiple views with low resolution. For this, further calibration/processing is usually required, for instance: i) to compute the micro-image centers; ii) to compensate for any potential optical/geometrical distortions that may result in micro-images with different sizes; iii) to deal with non-integer resolutions; and iv) to discard incomplete micro-images (at the border of the lenslet image). Examples of calibration/processing algorithms for lenslet images can be found in [9], [63]. Apart from the process of discarding incomplete micro-images, this calibration/processing can be invertible.
- **Based on Subaperture Images** — Using the knowledge of the exact LF optical setup (e.g., micro-image coordinates and sizes), a subaperture image can be constructed by extracting one pixel with the same relative position (u, v) from all micro-images. Hence, several low-resolution subaperture images can be extracted at different positions relative to the micro-image center.

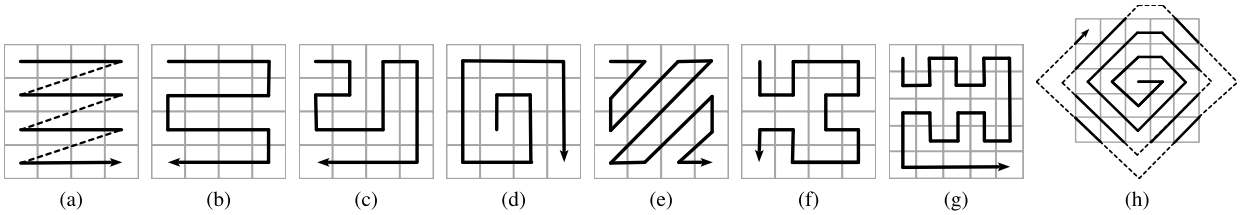


FIGURE 9. Possible scan topologies for arranging views: (a) Raster; (b) Serpentine; (c) Perpendicular; (d) Spiral; (e) Zig-zag; (f) Hilbert; (g) U-shape; and (h) Lozenge.

Usually, extracting subaperture images is not straightforward, needing additional calibration/processing to compute the micro-image centers and to align the micro-image grid to the pixel grid. Moreover, if an MLA with hexagonal grid is used, a transformation is needed to convert from hexagonal to a rectangular MLA grid. Examples of such processing algorithms can be found in [9], [63]. For lenslet images captured using an unfocused lenslet camera, each subaperture image represents an in-camera orthographic projection of the captured scene [41]. On the other hand, for lenslet images captured using a focused lenslet camera, a subaperture image (built by taking one pixel from each micro-image) can be seen as a subsampled perspective of the captured scene [64] or as a low resolution rendered view that is focused at infinite [65], [66] which, consequently, presents aliasing artifacts. Alternatives to deal with these aliased views have been proposed in the literature and involve resorting to depth-based rendering [64], [65] or Laplacian-based rendering [67], [68]. In both cases, an increase in the views' resolution is observed and, consequently, an increase in the LF data size. For this reason, this process to generate views from focused lenslet images is a non-invertible process [59], [60].

- **Based on Epipolar Plane Images** — In this case, the LF data can be decomposed according to its light ray distribution by using the Epipolar Plane Image (EPI) technique [69]. Each EPI can be then interpreted as a 2D cut through the captured 4D light field. As an illustrative example, Fig. 10 shows an EPI built by stacking together views in the same column (which corresponds to fixing the dimension u, as shown in Fig. 10a) and, then, taking a slice from these views in a particular horizontal plane (by fixing the direction x, as shown in Fig. 10a). A prospective characteristic of this EPI-based representation is that the depth/disparity of the objects can be estimated from the slope of the lines that can be observed in Fig. 10b. Examples of this usage can be found in [70]–[72].

3) MULTIVIEW REPRESENTATION

This representation may correspond to the full parallax multiview representation in Fig. 5b, in which the 4D LF content is organized as a 2D grid of multiple views. For representing LF videos, a 5D LF representation may be needed. However, revisiting the literature on LF coding, it can be observed that

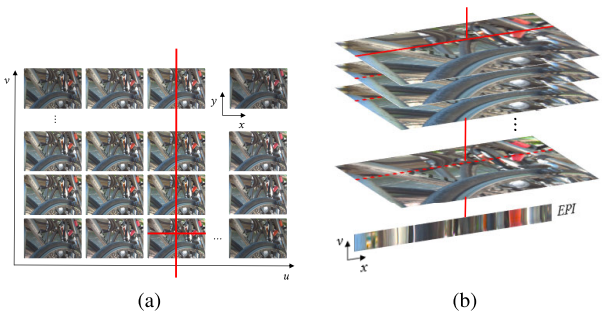


FIGURE 10. Extracting an EPI from the 4D light field: (a) Fixing the dimensions x and u; and (b) the built EPI.

many of the authors have rather proposed to organize the 4D LF content as a conventional 3D multiview format with only horizontal parallax and to use the temporal axis to comprise the fourth LF dimension, as illustrated in Fig. 11. Different to the (single) PVS representation in Section II-B2, this multiview representation comprises a 1D array of multiple PVSs. Thus, with this multiview representation, the correlations in an LF image can be exploited in all dimensions — i.e., spatial, inter-view, and (pseudo) temporal. For representing LF videos, the 2D array of viewpoints can be then firstly scanned using a specific topology (see Fig. 9) to form the 1D array of multiview videos.

From the full parallax multiview acquisition format there is no further pre-processing need prior to coding. Differently, to convert from the lenslet acquisition format, additional calibration and processing is necessary to compute the micro-image centers, to align the micro-image grid to the pixel grid and possibly a transformation to convert from hexagonal to a rectangular MLA grid, as discussed in Section II-B2. Then, views can be constructed based on: i) micro-images; ii) subaperture images; or iii) EPIs.

4) VOLUMETRIC REPRESENTATION

In this case, the dense 2D array of viewpoints needs to be scanned using a specific topology (see Fig. 9) and stacked together in the third dimension to form a 3D block volume as illustrated in Fig. 12. The difference to the PVS representation is on how the LF content is partitioned and processed on the subsequent encoding process. Instead of splitting it into 2D blocks to feed the encoding process, as in the PVS representation, the LF content is split into 3D blocks in the volumetric representation. From this representation, it can be observed that the correlation in the third dimension may

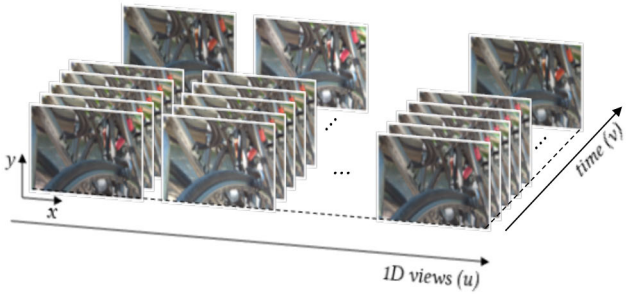


FIGURE 11. Multiview representation constructed using the temporal axis as the fourth dimension.

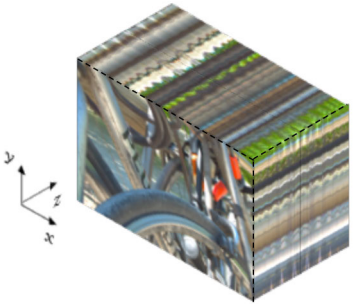


FIGURE 12. 3D volumetric representation constructed scanning the 2D grid of views in spiral order.

vary depending on the scanning topology used to form the 3D volume.

It should be noticed that the full parallax multiview format in Fig. 5b can also be interpreted as a 4D volumetric representation. This 4D volumetric representation has been also considered in the literature for coding. As discussed in Section II-B2, to convert from the lenslet acquisition format to this volumetric representation, additional calibration and processing steps are usually necessary to compute the micro-image centers, to align the micro-image grid to the pixel grid, and possibly a transformation to convert from hexagonal to a rectangular MLA grid.

5) GEOMETRY-ASSISTED REPRESENTATION

In this case, the dense LF data is represented by a sparse number of selected key views together with associated geometry information as depicted in Fig. 13. The geometry information may comprise, for instance, depth, disparity, or a graph model estimated from the LF data. For lenslet-based LF content, the sparse set of key views may comprise micro-images, subaperture images, or views with higher resolution (for LF content captured by a focused lenslet camera) that are extracted from the LF data.

As geometry information can be used for synthesizing views at the decoder side, the amount of views that needs to be coded and transmitted in the processing chain may be reduced. Consequently, the coding performance of this representation is highly dependent on the selection of the key views and the accuracy of the geometry information estimated from the acquired LF data. Trying to deal with the geometry estimation problem, various depth/disparity estimation methods [73]–[78] and graph learning models [79]–[81] have been recently proposed in the literature.

C. LF CODING

For the earliest LF coding schemes proposed in the literature, the most natural choice has been to use the classical image coding architecture shown in Fig. 14 — comprising transform, quantization and entropy coding blocks — and to apply an approach that resembles the JPEG [82] or JPEG 2000 [83] standards.

Later, following the advancement of digital multimedia systems and coding technologies, most of the LF image and video coding frameworks in the literature have been based on a hybrid coding architecture due to its effectiveness for providing high efficiency compression. Currently, the most successful class of visual coding architectures are based on this framework [84], which has been adopted in most 2D video coding standards, including the state-of-the-art HEVC [85] and the future video coding standard Versatile Video Coding (VVC) [86], which has been developed in a joint collaboration effort known as the Joint Video Exploration Team (JVET) between ITU-T Video Coding Experts Group (VCEG) and ISO/IEC MPEG. This model is called hybrid as it combines the advantages of using a transform coding stage from classical still image coding solutions with a prediction modeling loop (see Fig. 14), in which a prediction signal is generated from information available at both encoder and decoder sides. The block diagram of a conventional hybrid encoder is illustrated in Fig. 14 and comprises the following functional blocks:

- **Prediction Modeling** — The prediction modeling aims at reducing the redundancy by exploiting the inherent correlations of the input content. Usually, in a hybrid video coder, a prediction may be formed by using spatially neighboring samples — known as intra prediction — or by using neighboring frames — known as inter prediction. Instead of coding the original pixels values of the current block, only the difference between current and prediction block, called residual block, is encoded and transmitted. In inter-prediction, a motion compensated prediction is used for modeling the translational moving blocks in different frames. In this case, a displacement vector — known as motion vector — is used to indicate the horizontal and vertical positions (relative to the current block position) of the prediction block inside a previously encoded reference picture. Traditionally, inter prediction was designed for exploiting the redundancy between neighboring temporal frames, however, it can actually be generalized to other types of redundancy (e.g., inter-view prediction and non-local spatial prediction as will be seen later on in this paper).
- **Transform** — The goal of transform coding (see Fig. 14) is to convert the residual block into the frequency domain such that it has a representation that is both decorrelated — i.e., separated into components

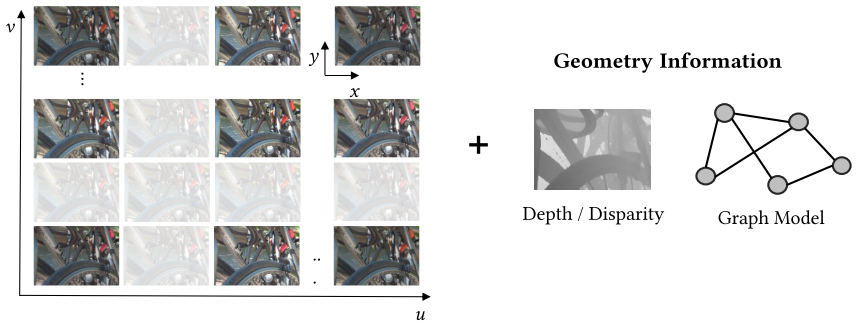


FIGURE 13. Geometry-assisted representation, comprising a sparse set of key views plus geometry information.

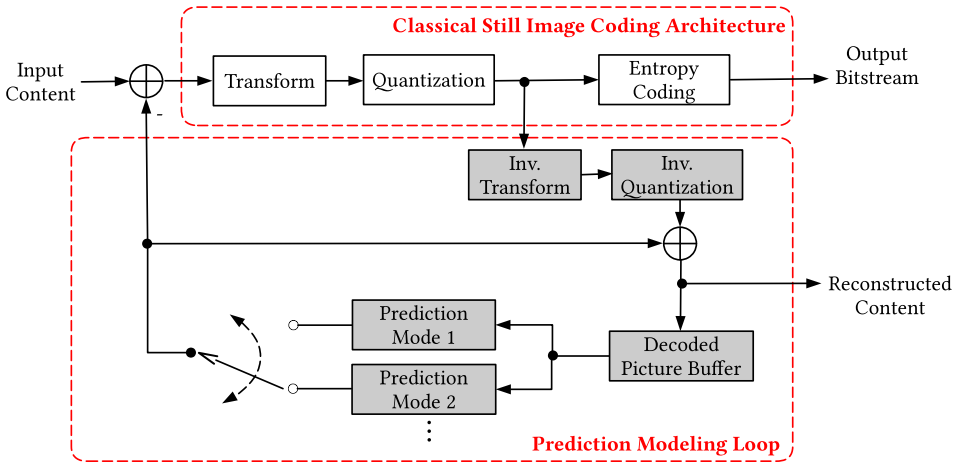


FIGURE 14. Block-diagram of a conventional hybrid video encoder. Built-in decoder is shown in gray shaded blocks.

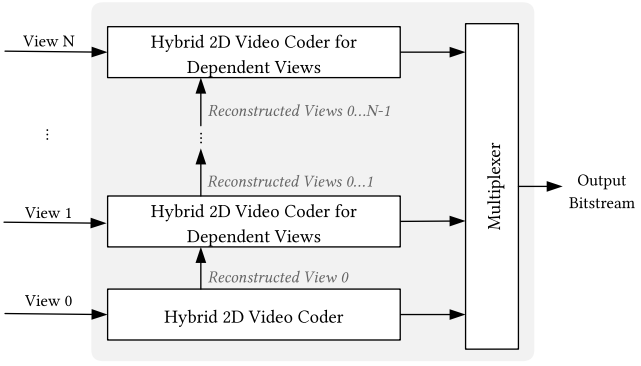


FIGURE 15. Multiview video coding architecture.

with minimal inter dependence — and compact — i.e., where most of the energy is concentrated into a small number of values. The most effective and widely used transform in image and video coders is the 2D Discrete Cosine Transform (DCT). However, JPEG 2000 [83] and MPEG 4 Visual [87] standards have also adopted the 2D Discrete Wavelet Transform (DWT). The output of this process is a transform block with the same size as the residual block, representing the image in the frequency domain. At the decoder side, an inverse transform is used to reverse the operation and reconstruct the residual block in spatial domain.

- **Quantization** — Quantization (see Fig. 14) is applied to the transformed coefficients. The quantizer is designed to discard insignificant values, such as near-zero coefficients, while preserving a small number of significant non-zero coefficients. In this process, the quantization step is used to regulate the range of the quantized values and consequently the output bit rate (or average bits per pixel in the case of still pictures).
- **Entropy Coding** — The small number of significant coefficients, as well the prediction parameters (e.g., quantized residual block, and motion vectors), are entropy coded to remove statistical redundancy. Among the various possible entropy coders in the literature, Context-based Arithmetic Binary Coding (CABAC) has shown to be a powerful method for providing a high degree of adaptation and redundancy reduction. For this reason, the state-of-the-art HEVC standard has also adopted CABAC-based entropy coding. In a nutshell, CABAC starts with a binarization process in which the entries are transformed to binary symbols (bins). For each bin, a suitable context model is then selected depending on the statistics of recently coded bins.

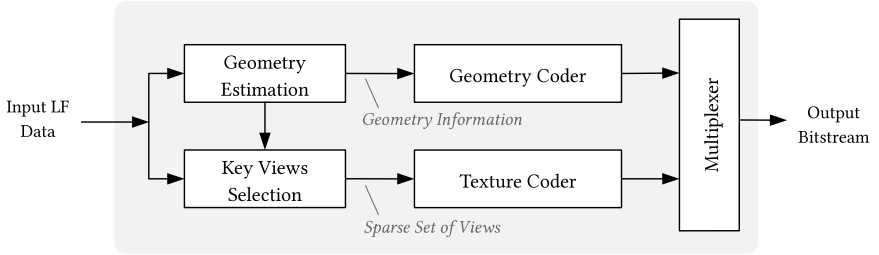


FIGURE 16. Geometry-assisted LF coding architecture.

Thus, each bin is arithmetic coded according to the selected context model. The output of this process is the compressed bitstream, which can then be stored or transmitted.

The encoder duplicates the decoder process to guarantee that they both generate identical predictions for subsequent frames (see shaded gray blocks in Fig. 14).

Instead of using a 2D video codec, some LF coding solutions have adopted a multiview video coding architecture as illustrated in Fig. 15, such as the one used in the HEVC multiview extension MV-HEVC [88]. The basic idea in these multiview video coding solutions is to exploit not only the redundancies that exist temporally between the frames within a given view, but also the redundancies between frames of neighboring views — known as inter-view prediction. In this case, a multilayer approach is used where different 2D video coded representations, called layers, are multiplexed into one bitstream. In MV-HEVC, a layer simply represents the texture data belonging to the same camera perspective (i.e., a view). In the first layer, usually denoted as the base layer, the pictures are coded independently from other layers. The layers that follow the base layer are denoted as enhancement layers. In these enhancement layers, inter prediction methods are used for both inter-view and temporal motion prediction by making the decoded pictures from other views also available as reference pictures.

Mainly motivated by the design of recent 3D coding solutions, such as 3D HEVC [88], geometry-assisted LF coding architectures have been also proposed in the literature for LF coding. In this case, the geometry-assisted representation format (see Section II-B5) is adopted in order to achieve compression. As illustrated in Fig. 16, a sparse set of key-views is selected from the LF data and encoded with a texture coder, which may be a 2D video coder (see Fig. 14) or a multiview video coder (see Fig. 15) solution. The geometry information estimated from the LF data is usually independently coded from the texture data by a dedicated geometry coder (see Fig. 16). Afterwards, data from both coders are multiplexed into one bitstream. At the decoder side, additional intermediate views are synthesized by using a specific view synthesis technique, as it will be seen in Section III-D.

D. LF RENDERING

While traditional 2D and 3D decoded content may be directly forwarded to the display stage without much processing, decoded LF content requires, typically, an appropriate rendering algorithm to be visualized, for instance, in conventional 2D/3D displays or in more advanced HMD and LF displays.

In this context, one important requirement for the design of LF rendering algorithms is to offer the best end-user experience targeting a specific display technology. Moreover, this requirement may eventually consider not only the visual quality of the rendered content, but also the level of user interaction that is enabled.

In this sense, among the advantages of the LF imaging technology is the ability to open new degrees of freedom in terms of content rendering, supporting post-production functionalities not straightforwardly available using conventional 2D and 3D systems, such as:

- **Perspective Shift** — Changing the perspective of the recorded scene can be obtained by simply switching between the captured dense array of views, or by retracing and interpolating light rays as they come from a virtual perspective camera at an arbitrary position, generating a virtual viewpoint [33], [89], [90]. The virtual camera can be positioned along the camera path or can be from within the captured scene - referred to as z-rendering and step-in/out effect [91].
- **Refocusing** — Refocus can be understood as virtually sliding the camera focus plane to a different plane (within the captured depth range). In its basic form, this operation can be obtained by aligning the views at a particular depth plane and integrating them over the directional axis [89]. Therefore, objects that are at the chosen depth plane appear in sharp focus in the rendered view, while the remaining objects are blurred. It is possible to select multiple depth planes to be in focus and to create an all in-focus rendered view. Some relevant algorithms proposed in the literature for refocusing can be found in [41], [65], [89], [92]–[94].
- **Depth-of-Field Control (Synthetic Aperture)** — Extending or narrowing the depth-of-field in the rendered view can be obtained by defining greater or smaller depth ranges where scene objects are rendered in focus. This capability emulates the effect of changing the aperture size in conventional photography where larger apertures produce images with shallower depth-of-field, and with smaller apertures more of the scene appears in focus. As discussed in [89], an arbitrary synthetic aperture size can be obtained when rendering a view by

choosing more or less views to be integrated over the directional axis. The shape of the synthetic aperture and, consequently, the bokeh in the rendered view can also be controlled by weighting differently the samples from each view [89].

- **Super-Resolution** — It is possible to take advantage of the full potential of the LF information to apply super-resolution either in the spatial or in the angular dimensions. LF spatial super-resolution typically uses depth information, estimated from the LF data, to super-resolve a view by propagating light rays intensity values from neighboring views to sub-pixel positions, as proposed in [64], [66], [95]. Angular super-resolution is typically used to synthesize virtual viewpoints from a small set of views, as proposed in [96]–[98]. View synthesis algorithms proposed in the literature for compression purposes will be seen in Section III-D.
- **Distance Measurement** — If the optical properties of the LF system are known precisely, it is possible to translate the relative estimated depth data into absolute distance from the lens, as proposed in [99]–[101].
- **Dooly-Zoom (Vertigo Effect)** — This effect is used to make a subject in the foreground remain in a relatively static position while the background compresses or stretches. This capability can be accomplished by combining a virtual zoom out/in with a step-in/out effect [102].

E. LF DISPLAY

Naturally, since LF imaging systems allow recording the 4D light field, the LF content can be more easily played in a wider variety of display technologies by simply re-creating different displayable versions of the same LF content. In this context, among the possible display technologies that are currently available for LF content visualization, one can cite:

- **2D Displays** — In this case, a single 2D view, or more specifically, a 2D version of the LF content must be rendered from the decoded LF content.
- **Stereo Displays** — In this case, a pair of views need to be rendered from the lenslet image and delivered to the display. This type of display technology allows then improving the user's depth perception (with respect to the 2D display) by presenting a different view to his/her left and right eyes (typically, by means of a pair of eyeglasses).
- **Multiview Autostereoscopic Displays** — Multiview Autostereoscopic is a glassless display technology that allows creating a more natural 3D illusion (with respect to the stereo display) to the end-user by presenting a different perspective as the user moves horizontally around the display (known as horizontal motion parallax). In this case, multiple views need to be rendered from the LF content and delivered to the display.

Moreover, following the recent developments in sensor and optical manufacturing, the display technologies are also evolving for providing a more natural and immersive visualization. Therefore, some prospective display technologies have also started to show up. Among them, it is possible to cite:

- **AR and VR Displays** — AR and VR HMD allow the user to see different perspectives as he/she moves through the scene. In the case of an AR HMD, the real environment is seen through half-transparent mirrors and then virtual 2D views are seamlessly blended into the real scene [1], [103]. In the case of a VR HMD, a large number of virtual 2D views are delivered to the HMD for providing to the user the impression of immersion in a real environment. Some AR and VR solutions have proposed to take advantage of a microlens-based [1], [104] or a mirror-based [105] LF imaging technology for creating a more natural visualization in AR and VR HMDs.
- **LF Displays** — A display technology using an optical setup similar to the one used in lenslet cameras can be also designed for LF visualization, as proposed by Nippon Hōsō Kyōkai (NHK) Japan Broadcast Corporation [3]. Another LF display technology uses a very dense number of views to create a replica of the 4D light field, as proposed by Holografika [4], Ostendo [8], and Looking Glass Factory [106].

F. LF QUALITY EVALUATION

A new challenging research topic is also to assess the visual quality as perceived by end-users after processing and compressing LF content. Visual quality assessment is generally carried out by determining suitable objective and subjective evaluation techniques.

Among objective quality metrics, Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) [107] have been the most commonly used metrics to assess the visual quality of 2D image and video under compression distortions. In the context of LF content, JPEG Pleno [108] and MPEG-I [59] have also adopted these metrics in core experiments. In this case, PSNR and SSIM are calculated per view for all color components and the average over all viewpoint positions is adopted as an overall quality measure. Apart from this, there have been only a few works specifically addressing objective quality assessment metrics for LF content. In [109], a sparse angle-dependent and a sparse depth-dependent metric are proposed. In the sparse angle-dependent metric, the average SSIM is taken in a set of five views from equidistant viewing angles. In the sparse depth-dependent, depth information is estimated and the average PSNR is taken in a rendered view (with fixed angular position) over sets of pixels in different depth layers. In [110], a quality evaluation method based on contrast and gradient measurements is proposed that measure the impact of LF compression in the visual quality by measuring the amount of compression blur in rendered views. In [111], a no-reference metric is proposed to quantify the amount of distortion due to LF rendering based on the number of light rays per unit area of the scene that is used to estimate an unknown

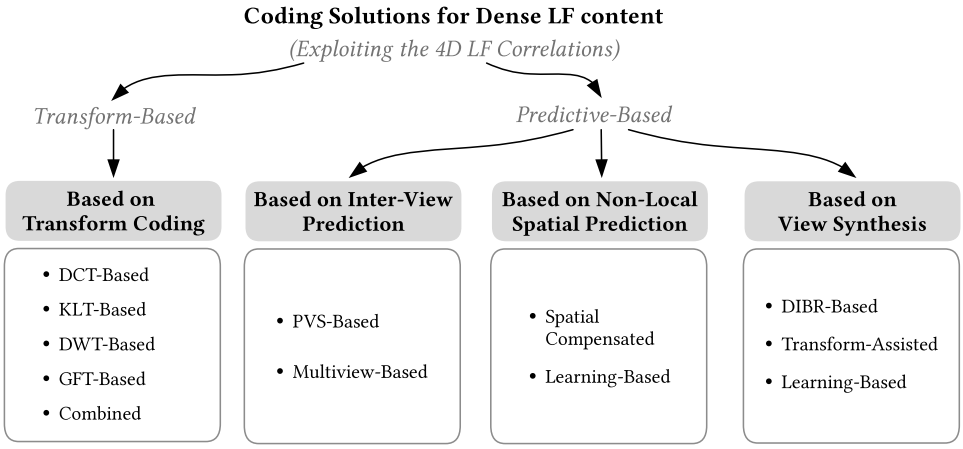


FIGURE 17. The LF coding solutions in the literature are grouped into four categories (in shaded gray blocks).

light ray. In [112], an objective metric that combines a spatial component and an angular component is proposed to evaluate the visual quality of 3D content on light field displays.

Following the standardization initiatives of JPEG Pleno and MPEG-I, methodologies for subjectively assessing the visual quality of LF content have also started being addressed in the literature. In [23], a methodology is proposed for visually evaluating compressed Lytro Illum lenslet images using a conventional 2D display and Double Stimulus Continuous Quality Scale (DSCQS) metric [113]. In [114], a methodology as well as prototype software for performing subjective quality assessment of compressed Lytro Illum lenslet images is proposed that aims at designing a methodology that enables global assessment of quality of experience in a flexible and interactive way [114]. In [115], a LF dataset for quality evaluation is proposed (see Table 1) and an analysis of the subjective quality of compressed LF image is presented. In [54], an analysis of the impact of different visualization techniques — i.e., image-based and animated-based visualization of rendered views in a 2D display — on the obtained scores in subjective quality evaluation is presented. In [116], an analysis of how light field subsampling affects the perceived quality of refocused views visualized in an animated fashion is presented. In [53], a subjective evaluation on a 3D monitor with head tracking is performed to assess the performance of various objective quality metrics on distorted LF contents. In [117], a subject evaluation on a light field display [4] is performed aiming at analyzing the correlation between spatial and angular resolution and discussing ways to improve parallax perception.

III. LITERATURE REVIEW ON DENSE LIGHT FIELD CODING

After reviewing the most relevant aspects of LF imaging and processing in the previous section, it is then possible to better characterize the existing LF coding approaches in the literature. To facilitate this, the various LF coding approaches are clustered into the two major classes depicted in Fig. 17, by identifying which functional part of the codec is responsible for exploiting the inherent LF correlations. Notably:

- **Transform-Based Approaches** — As its name suggests, transform-based approaches exploit the LF correlations in the transform domain, based on a particular transform coding technique. LF coding solutions in this class are reviewed in Section III-A.
- **Predictive-Based Approaches** — Differently, predictive based approaches exploit the LF correlations in a predictive manner. As illustrated in Fig. 17, predictive based approaches can be further categorized depending on the particular data format and prediction schemes adopted. Notably, three categories are identified: i) LF coding based on inter-view prediction, which is reviewed in Section III-B; ii) LF coding based on non-local spatial prediction, reviewed in Section III-C; and, finally, iii) LF coding based on view synthesis, reviewed in Section III-D.

Although transform and predictive based approaches appear separated from each other in Fig. 17, it does not mean that a transform based approach excludes completely any type of predictive coding tool from its architecture (or vice versa). For instance, a predictive based approach may use a hybrid coding architecture, as seen in Section II-C.

It is worth mentioning that, the evaluation procedures and the coding conditions are usually divergent among different publications, which makes it difficult a straightforward comparison between distinct LF coding solutions in the literature. Nevertheless, when discussing about the performance of each LF coding solution presented, an effort was made to include some quantitative results in terms of bit savings as long as it was available in the original publication (alternatively, coding gains in decibel (dB) are presented). Additionally, at the end of each section, a high-level discussion is provided to highlight the most relevant results, advantages, and limitations of the LF coding solutions in each category.

A. TRANSFORM-BASED LF CODING

Starting with transform based approaches, these correspond to LF coding solutions that rely on transform coding for exploiting the inherent LF correlations. Specifically, various transform coding techniques can be used to decorrelate the LF image and then remove the redundant information between neighboring views. Therefore, it is possible to group the transform based approaches into five categories, depending on the type of used transform (see Fig. 17): i) Discrete Cosine Transform (DCT)-based; ii) Discrete Wavelet Transform (DWT)-based; iii) Karhunen Loève Transform (KLT)-based; iv) Graph Fourier Transform (GFT)-based; and v) combined approaches.

These approaches have been mainly proposed for lenslet LF coding. In this case, the lenslet image undergoes a preprocessing operation to convert it to the volumetric representation format presented in Section II-B4. Therefore, not only the existing spatial redundancy within each view is exploited, but also the redundancy between neighboring views.

1) DCT-BASED CODING

Inspired by the approach adopted by JPEG standard, the LF coding schemes in this group make use of the classical image coding architecture shown in Fig. 14, but to apply a 3D or 4D version of the DCT transform.

In [118], a DCT-based coding solution is proposed for lenticular-based imaging [119], where a 1D cylindrical MLA is used for capturing instead of the 2D array of microlenses. The lenslet image is organized into stacks of 8 adjacent micro-images and the 3D DCT is applied to each 8×8×8 block. Then, the resulting DCT coefficients are uniformly quantized and both DC and AC quantized coefficients are equally entropy coded by using a combination of run-length and Huffman coding. It is shown that the proposed solution presents significant improvements compared to JPEG for gray-level lenslet images. In [120], further improvements in compression performance are achieved by using an alternative quantization strategy. In [121], the solution with the 3D DCT from [118] is generalized for lenslet LF with full parallax. In this case, both the horizontal and vertical micro-images are decorrelated simultaneously by the 3D DCT. Hence, it is shown that different scan ordering approaches for gathering the horizontal and vertical micro-images (to form 8×8×8 stacks) result in different Rate-Distortion (RD) performances. This fact has motivated the work in [122], which proposes to use a Hilbert space-filling curve (see Fig. 9f) for scanning the micro-images in the array and forming 8×8×8 stacks to be 3D DCT coded. Various scanning topologies are compared, namely: raster, perpendicular and spiral (see Fig. 9) and it is shown that the 3D DCT in conjunction with the Hilbert scan outperforms all other tested solutions. In addition, an adaptive 3D DCT based framework is proposed in [123], in which the number of micro-images involved in a single 3D DCT is varied (between 8×8×1, 8×8×2, 8×8×4, and 8×8×8 stacks of micro-images) according to the micro-image cross correlation in a neighborhood. In this case, the micro-image mean values are used as a measure of correlation between micro-images. Consequently, it is shown that the adaptive 3D DCT could significantly outperform the non adaptive solution from [118] (for lenticular-based images). An alternative adaptive 3D DCT-based approach has been proposed in [124]. Similarly to [123], the mean value is used as a correlation metric. However, the 3D DCT is applied to stacks of subaperture images and the size of the 3D DCT is varied between 16×16×16 down to 4×4×4, depending on the correlation between neighboring subaperture images. From this, it is shown that further improvements can be achieved compared to the adaptive solution proposed in [123] (when also applied to subaperture images).

More recently, a lenslet LF coding approach using a 4D DCT is proposed in [125] and referred to as Multidimensional Light Field Encoder (MuLE). To explore the 4D redundancy of LFs, the lenslet image is converted to its full-parallax multiview format (see Fig. 5b) and it is divided into 4D blocks (with two spatial plus two directional dimensions) and a separable 1D DCT is applied to each dimension. Then, the resulting DCT coefficients are grouped into bitplanes and processed by an hexadeca-tree coder to cluster zero coefficients into 16 4D subregions — referred to as hexadeca tree partition [125]. Both the hexadeca-tree bits and the bits from the significant coefficients are entropy coded using an adaptive arithmetic coder. This solution is compared against a PVS-based solution in which all subaperture images are scanned in serpentine order (see Fig. 9b) to be coded using HEVC and VP9 codecs, and against a LF solution based on JPEG 2000 and view synthesis proposed in [126] (see Section III-D1. Compared against the PVS-based HEVC solution for coding lenslet images from the *EPFL LF Dataset* (see Table 1), the MuLE achieves, in average, 38.1% of bit savings, while the view synthesis solution in [126] achieves 36.4 % of bit savings, and the PVS-based VP9 solution achieves 20.5% of bit savings. This solution has been recently adopted in the JPEG Pleno Verification Model (VM) [127] (since version 2.0) for lenslet LF coding.

2) KLT-BASED CODING

Instead of using the DCT as in the previous section, other schemes propose to use a KLT-based approach for LF coding. The KLT — a.k.a. Principal Components Analysis (PCA) [128] decomposition and Hotelling [129] transform — is a block based transform that exploits the statistical characteristics of the input data. The KLT consists of decomposing the input data in a set of orthonormal basis functions (known as the principal components) into which the variance of the input data is maximal. This corresponds to ordering the eigenvectors of the covariance matrix (the KLT matrix), which is calculated with the input data, according to the largest eigenvalues.

The idea of applying the KLT transform for compression comes from the fact that a linear combination of any reduced number, k, of eigenvectors corresponds to the best

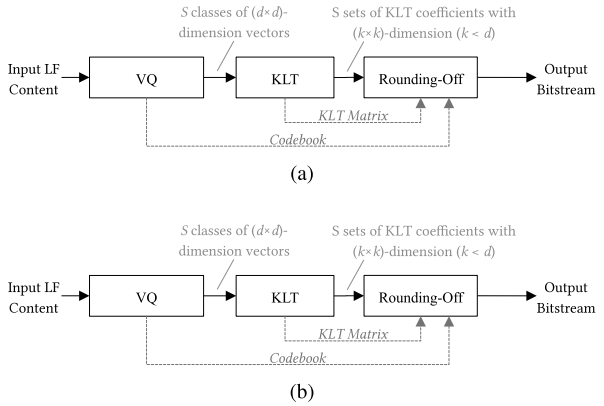


FIGURE 18. KLT-based LF image coding schemes: (a) Proposed by Jang, Yeom, and Javidi [131]; and (b) Proposed by Kang, Shin, and Kim [133], [134].

approximation of the input data in a reduced k-dimension subspace (i.e., the approximation with minimal mean square error) [130]. Therefore, different compression ratios can be achieved by simply discarding less (or more) eigenvectors (i.e., discarding rows from the covariance matrix). Concerning the usage of KLT for image compression, although the KLT is very efficient in compacting the energy in a small number of eigenvectors, there are still some implementation related difficulties, mainly due to the fact that the KLT basis functions are image dependent. However, it may be suitable in applications where the statistics of the data change slowly and the covariance matrix is kept small [130].

Regarding LF coding, a KLT-based coding scheme is proposed in [131] for lenslet LF coding, in which a Vector Quantization (VQ) scheme is used for clustering different micro-images into a representative set of vectors to be then coded with KLT, as illustrated in Fig. 18a. For this, the lenslet image is divided into consecutive blocks of $d \times d$ samples which are treated as a $(d \times d)$-dimensional vector. These vectors are then grouped into S different classes by using the Linde-Buzo-Gray (LBG) optimization algorithm [132]. As a result, a codebook is derived, consisting of S representative vectors (known as code vectors). Then, a KLT with $(k \times k)$-dimension is applied into the vectors from each of the S classes so as to reduce the dimensionality of their vectors from $(d \times d)$ to $(k \times k)$-dimension vectors, where $k \leq d$. Afterwards, the reduced KLT coefficients, together with the codebook and the KLT matrix, are scaled and rounded to the nearest integer to compose the output bitstream. From the presented results, it is shown that varying the $d \times d$ block size does not affect the RD efficiency for lenslet image coding. However, the larger the number of sets S, the better is the observed RD performance. Moreover, the presented KLT scheme always outperforms the JPEG standard for lower bit rates.

An alternative KLT-based coding scheme is also proposed in [133], [134], as illustrated in Fig. 18b. In this case, the lenslet image is decomposed into its subaperture images, which are then KLT coded. It is worth noting that there is no further information on how the resulting KLT coefficients together with the KLT matrix are coded and transmitted in [133], [134]. It is shown that this approach achieves better RD performance compared to JPEG and the same KLT based approach applied to micro-images. In addition, it is observed that the statistical characteristics between subaperture images are more easily decorrelated than between micro-images, having most of the relevant information compacted into a smaller number of eigenvectors. As stated in [134], due to the small FoV of the microlenses in the array, each captured micro-image comprises only a small portion of the 3D scene, which may have different characteristics in different areas of the 3D scene. On the contrary, all subaperture images comprise the complete 3D scene, which are only slightly different on the angles of projection. Consequently, LF redundancy is considerably larger in subaperture images than in micro-images.

3) DWT-BASED CODING

In alternative to block-based transforms, such as the DCT and KLT, some authors proposed to use an approach based on DWT coding, closer to the coding techniques used in JPEG 2000 codecs [83].

In [135], a 3D DWT-based coding scheme is proposed for lenslet LF coding following the classical still image coding architecture illustrated in Fig. 14. For this, the lenslet image is firstly decomposed into a stack of subaperture images and a separate 1D DWT is recursively applied in the third dimension of this stack until the lowest frequency subband contains only two samples (in the third dimension). Then, a two level 2D DWT decomposition is applied to these two sets of lowest frequency bands. Similarly to JPEG 2000, the lowest frequency subbands are quantized using a deadzone quantizer, while the remaining high frequency coefficients are quantized using a uniform scalar quantizer. Following this, a new scanning pattern is proposed to be used to scan samples from all subbands together, which are then arithmetic coded.

In [136], a similar approach with a 3D DWT applied to a stack of subaperture images is proposed. However, in this case, the three (separable) 1D DWTs are recursively applied to each dimension of the stack, producing 8 subbands in each decomposition level. Afterwards, the 3D DWT coefficients are quantized using a deadzone scalar quantizer and coded using the method of Set Partitioning In Hierarchical Trees (SPIHT) [137]. Similarly to the Embedded Block Coding with Optimal Truncation (EBCOT) [138], used in JPEG 2000, the SPIHT algorithm is used as a form of entropy coding applied to bitplanes of quantized coefficients to allow progressive transmission of the LF data. The proposed approach is compared to a 2D version of the coding scheme, in which a 2D DWT is applied to the entire lenslet image followed by SPIHT. The 3D DWT scheme presents significant improvements compared to the 2D DWT. Moreover, several DWT bank filters are analyzed for the 3D DWT coding, and the Biorthogonal 2.2 filters show the best results, but are very similar to the Daubechies filters. In [139], a 4D DWT-based

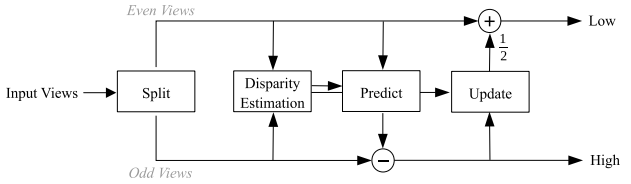


FIGURE 19. 1D DWT lifting structure.

scheme is proposed for coding LF acquired with multi-camera arrays. For this, a separable 1D DWT is applied to each of 4D dimensions of a LF data. The obtained 4D DWT coefficients are then coded by using SPIHT [137].

Regarding standard DWT based coding solutions, a study on lenslet image coding is presented in [140], in which the performance of two DWT based coding solutions (JPEG 2000 and SPIHT) and one DCT based standard solution (JPEG) are compared. The performance is analyzed in terms of the objective quality of views rendered from the coded and reconstructed lenslet image, by using average PSNR and average SSIM index. It is shown that the SPIHT scheme presents better RD performance than JPEG for low bit rates, but JPEG 2000 outperforms them both for either PSNR or SSIM metrics. In [141], a similar study is performed for comparing two standard solutions, JPEG 2000 and JPEG XR [142], for lenslet image coding. This study focuses on comparing the performance, in terms of objective quality of rendered views, of the different transform coding solutions used in each standard, namely: the JPEG 2000 DWT and the block-based DCT transform used in JPEG XR [142]. In the presented results, JPEG 2000 achieves slightly better RD performance than JPEG XR both in terms of PSNR and SSIM metrics. In addition, an empirical performance analysis is presented in [143] for synthetic lenslet images. For this, JPEG 2000 is compared to its extension for volumetric data compression — JPEG 2000 Part 10 [144], known as JP3D — for lenslet image compression. The JP3D solution supports 3D DWT decompositions and extends tiles, code-blocks and Region of Interest (ROI) functionalities accordingly to support volumetric data. In the presented study [143], JPEG 2000 is applied to the entire lenslet image, while two different scenarios are considered for JP3D, in which the 3D DWT is applied to stacks of micro-images, and to stacks of subaperture images. It is shown that the JP3D solution outperforms the JPEG 2000 for both scenarios.

In [145], a 4D DWT scheme is proposed for coding LF captured with a multi-camera array. Firstly, a 2D lifting-based Haar DWT [146] is carried out by applying a 1D DWT lifting structure horizontally and vertically across the 2D array of views. As depicted in Fig. 19, estimated disparity information is used in the prediction and update steps of the 1D DWT lifting structure to exploit the inter-view correlation. For this, the views are divided into two interlaced sets of even and odd views. In the prediction step, the disparity information is used to predict an odd view by warping it from an even view. Then, the resulting prediction residual from the odd view corresponds to the high frequency subband of the Haar DWT. Afterwards, in the update step, this high frequency subband is then warped and added to the even view to generate the low frequency subband, which is approximately the disparity compensated average between even and odd views. After this 2D inter-view transform, a multi-level 2D DWT is applied to each frequency subband images. To encode the transformed coefficients, a modified SPIHT algorithm is adopted to work in a block-wise manner. It is worth mentioning that this coding architecture supports view scalability by progressively decoding the interview frequency subbands. Moreover, the number of scalable layers can vary by applying more or less decomposition levels in the inter-view DWT. Motivated by this fact, a scalable lenslet LF coding scheme based on DWT lifting is proposed in [147]. In this scheme, the disparity information is derived by matching a set of SIFT descriptors extracted from two different subaperture images and estimating the corresponding homography transform. Then, the resulting homography matrix is transmitted to the decoder side. Similarly to [145], a 2D lifting-based Haar DWT is applied to exploit inter-view correlation. Afterwards, the inter-view frequency subband images are coded using JPEG 2000 coder. Experimental results for coding lenslet images from the *EPFL LF Dataset* (see Table 1) show that it is possible to reach, in average, 62.85% and 78.80% of bit savings compared to coding each subaperture image independently using JPEG 2000 and JPEG, respectively. In [148], a lifting-based DWT scheme is also proposed for coding LFs captured with a multi-camera array. Similarly to [147], the DWT subband images are coded using JPEG 2000. However, in this case, the views are divided into Groups of Views (GOV) and the disparity is coded for only one reference view from each GOV. For this, an anchored disparity modeling is proposed for representing the disparity data and a backfilling methodology is proposed for deriving disparity relationships in disoccluded areas. Then, the disparity data in reference views are coded using a break point adaptative DWT with 5 levels of decomposition followed by EBCOT. Experimental results are presented for coding LF images from the *Fraunhofer LF Dataset* (see Table 1), which compare the proposed solution against a PVS-based solution with serpentine ordering (see Fig. 9b) and HEVC. It should be noticed that, in these experiments, only a subsampled set of views are coded with the proposed scheme and the remaining views are interpolated using the 4 nearest coded views. The results show a superior performance of the proposed solution, achieving gains of 2 dB and 3–4 dB, respectively, at high and low bit rates.

4) GFT-BASED CODING

Graph signal processing has proved to be a powerful tool for modeling irregular structures and the complex interactions among them [149]. Notably, the GFT allows extending the notions of classical Fourier transform to signal samples indexed by nodes of an arbitrary directed or undirected graph. A graph is commonly defined as a mathematical struc-

ture $G = (V, E)$ composed of N nodes (a.k.a. vertices), $V = \{v_0, v_1, \ldots, v_{N-1}\}$, and a set of edges E. The set of nodes is used to model elements of a system, while the set of edges is used to encode any relevant relationship between these elements. From this, the graph signal can be defined as a vector $\mathbf{s} \in \mathbb{C}^N$, where its n-th entry s_n denotes the signal value on the node v_n of the graph — called the vertex domain. There are many choices for defining the frequency representation to be used in GFT, and the decision between them usually depends on the problem being considered [149]. The two most widely used representations consider the adjacency matrix $\mathbf{A}$ and the graph Laplacian $\mathbf{L}$ [150] as shift (delay) operators. Generically, the adjacency matrix $\mathbf{A}$ generalizes the shift operator of classical discrete signal processing and applies to directed and undirected graphs, while the graph Laplacian $\mathbf{L}$ applies only to undirected graphs with positive weights, so that $\mathbf{L}$ is symmetric and positive semidefinite, which avoids some numerical difficulties that may arise when choosing $\mathbf{A}$ [149]. Assuming the simplest case when the shift matrix ($\mathbf{A}$ or $\mathbf{L}$) is symmetric, the GFT can be obtained from the eigen decomposition of $\mathbf{A}$ or $\mathbf{L}$. In this particular case, all the eigenvalues are real and nonnegative, and the full set of orthogonal eigenvectors can be obtained. Then, the eigenvectors are the basis vectors to the GFT matrix, and the eigenvalues are the graph frequencies. The basis vectors are ordered from low to high graph frequencies. Similar to the KLT, the corresponding GFT matrix is image dependent.

GFT-based methods have been recently proposed for LF coding. In this case, a graph-based representation is used to model color, depth/disparity or other geometry information from the LF content. In [151], the authors propose a GFT-based scheme for coding LFs captured by a multi-camera array. For this, a graph representation based on the adjacency matrix is constructed for each possible residual block of $d \times d$ samples by defining each pixel position in the block as a node, and the residual value in each pixel as the graph signal. Then, a sparse adjacency matrix $\mathbf{A}$ is built according to a 2D Nearest-Neighbor (NN) graph model [152]. This scheme is based on the HEVC codec, but the DCT is replaced by the proposed GFT to encode the residual data. Hence, the LF data is organized as a PVS (see Section II-B2) to be coded by scanning the views in different orders [151]. To avoid transmitting $\mathbf{A}$ for every single block of $d \times d$ samples in every view, the proposed scheme assumes that blocks in the same position in different views are highly correlated. Results for LF images from the *Stanford LF Archive* and *HCI 4D LF Dataset* (see Table 1) show that it is possible to reduce up to 22% the number of coefficients compared to DCT. In [153], a GFT-based solution for lenslet LF coding is proposed using the graph lifting transform. To achieve further compression, the lenslet image acquired by the sensor is calibrated, converted to subaperture images, and coded prior to demosaicking. Thus, a graph is constructed to represent the sparsely distributed color pixels in each subaperture image, by defining each pixel position as a node, and the color intensity in each pixel as the graph signal. For each non-overlapping block of $d \times d$ samples, a sparse adjacency matrix is constructed for each color component by connecting pixels based on the Euclidian distance between its nodes. Once the graph is constructed, a graph lifting transform [154] is applied. The transformed coefficients are uniformly quantized and entropy coded using an Amplitude and Group Partitioning (AGP) method [153]. Experimental results for lenslet images from the *EPFL LF Dataset* (see Table 1) show significant coding gains only for high bit rates compared to a scheme in which demosaicking is applied prior to coding and each subaperture image is coded independently with HEVC. Since the solution is outperformed by HEVC for low bit rates, the authors suggest using the solution for applications such as archiving and instant storage on lenslet cameras. In [155], a GFT solution with a support defined on super-ray segmentation is proposed for lenslet LF coding. As proposed in [156], a super-ray segmentation is used for grouping light rays of similar color values and being close spatially in the 3D space by taking into account the disparity information. This corresponds to grouping perceptually similar regions across all subaperture images. Then, two separable GFTs are applied locally in each super-ray region for exploiting both spatial and angular dependencies. Firstly, a GFT is constructed for each segmented region in a reference subaperture image by defining each pixel position in the region as a node which is connected (through edges) to its spatially neighboring pixels. To increase the correlation of spatial GFT coefficients across different subaperture images (and, consequently, improve energy compaction), a graph optimization method is proposed for finding coherent spatial GFT basis vectors for segmented regions on the remaining subaperture images. Afterwards, an angular GFT is constructed for each spatial graph frequency inside a super-ray. For this, the nodes are defined as the subaperture images where the spatial graph frequency exists, and the edges are drawn from one node to its direct four neighbors. Finally, the GFT coefficients are quantized and encoded using CABAC. The segmentation map of the reference subaperture image and the disparity value per super-ray are also encoded using an arithmetic coder and transmitted to the decoder side. Experimental results for lenslet images captured by a Lytro Illum camera (including three LF images from the *EPFL LF Dataset* in Table 1) are shown, in which the proposed solution is compared against four different LF coding solutions: i) a PVS-based solution using lozenge scan order and HEVC; ii) a PVS-based solution proposed in [157] (see Section II-B2); iii) a LF coding solution based on view synthesis proposed in [158], [159] (see Section III-D1); and iv) a LF coding solution based on view synthesis proposed in [49] (see Section III-D3). From these results, it is seen that the proposed solution is outperformed by the PVS-based solution in [157] and by the view synthesis-based solution in [49], but is able to present some coding gains at high bit rates when compared to the PVS-based with lozenge scan and the views synthesis-based solution in [158], [159].

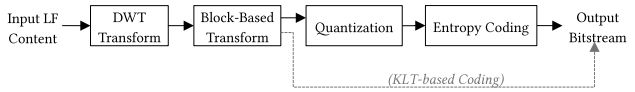


FIGURE 20. Block diagram of combined-transform coding schemes for lenslet image compression. When using the KLT as the block-based transform, quantization and entropy coding processes are bypassed (as illustrated by the dashed line).

5) COMBINED TRANSFORM CODING

This category corresponds to LF coding schemes in which two or more types of transforms are combined to separately exploit the spatial redundancy between samples in a local neighborhood and the inherent LF cross correlation in neighboring views. A common characteristic of most of these LF coding approaches is that an image-based transform (notably, the 2D DWT) is used to exploit the local sample correlation, followed by a block-based transform that is applied to the lowest subbands across different views, as illustrated in Fig. 20. The major motivation for this choice is to reduce the blocking artifacts that are likely to appear in the reconstructed image when using a block based transform coding.

In [160], a combined transform scheme is proposed for lenslet LF coding that combines a 2D DWT with a 2D DCT. In this scheme, the lenslet image is divided into tiles with the micro-image size to be recursively decomposed with a 2D DWT. Following this, a packet partition scheme is used to rearrange the samples from the same DWT subband into blocks of 8×8 samples to be DCT coded. The 2D DCT coefficients are then scalar quantized and entropy coded similarly to JPEG encoding. The presented approach outperforms JPEG with significant gains, mainly at low bit rates.

In [161], an approach combining a 2D DWT and a 3D DCT is proposed for coding lenticular-based images. For this, the 2D DWT is recursively applied to each subaperture image extracted from a lenslet image to decompose them in two levels. Then, the lowest subbands from different subaperture images are stacked into $8\times8\times8$ blocks to be processed by a 3D DCT. Afterwards, all coefficients within all the subbands are quantized using a deadzone scalar quantizer. Then, the 3D DCT coefficients are Huffman coded while all the other coefficients are arithmetic coded. In [162], the previous solution is extended for full parallax lenslet image, which is then compared to the solution in [121]. It is shown that the combined transform approach achieves improved RD performance at low bit rates when compared to the scheme proposed in [121], in which only the 3D DCT is used.

In [163], instead of using the DCT as the block-based transform, the 2D DWT is combined with the KLT. In this case, a one level 2D DWT decomposition is individually applied to all micro-images, resulting in four subbands per micro-image. Then, samples in the same subband from all micro-images are arranged into four arrays, and the KLT is applied to each of them. Then, the four arrays with reduced dimensionality are transmitted together with the KLT matrix. It is shown that the combined solution performs significantly better, in terms of RD performance, compared to a scheme very similar to the one proposed in [131] (see Fig. 18a), where only the KLT is used. Moreover, several filter banks are tested for the 2D DWT, and the Daubechies is the one that results in better RD performance for lenslet image coding.

6) DISCUSSION

Several LF coding schemes proposed to exploit the inherent correlations of the LF content in the transform domain. The greater advantage of such schemes is the simplicity of the proposed coding architecture, inspired by the most popular still image codec, the JPEG standard. Generically, the proposed transform-based LF coding solutions have presented significant coding gains compared to standard transform-based solutions, such as JPEG and JPEG 2000. Additionally, 4D transforms have been shown to be the most suitable for LF coding, independently of the transform family that is adopted (e.g., DCT, DWT, or combined approaches), since they are able to better exploit all the LF correlations.

From the different transform coding techniques proposed, 4D DWT-based approaches, either when applied alone (as in [145], [147], [148]) or combined with a block-based transform (as in [163]), have shown to present the best RD performance while providing other typically required functionalities, such as quality and resolution scalability, random access and ROI coding. Nevertheless, recent experimental results have suggested that the 4D DCT-based MuLE solution in [125], as adopted in the JPEG Pleno VM 2.1 [127], can present competitive RD coding performance. Moreover, in the context of JPEG Pleno VM 2.1 [127], a random access extension of the 4D DCT-based solution is proposed in [164] by independently coding a set of adjacent blocks of 4D DCT coefficients. However, random access is supported at the price of significantly increasing the bit rate (to almost the double [164]).

Compared to other LF coding solutions in the literature (outside the transform-based category), it has been shown [165] that the transform-based solutions [125] can achieve competitive RD performance when compared against LF coding solutions based on inter-view prediction (such as PVS-based approaches) and some solutions based on view synthesis [158], [159], but only for coding LF acquired using lenslet cameras, where the 4D redundancy is considerably larger than in LF acquired using multi-camera or gantry setups.

B. LF CODING BASED ON INTER-VIEW PREDICTION

Instead of exploring the inherent LF correlations in the transform domain, as discussed in the previous section, other authors have proposed to do it relying on a predictive approach to exploit the correlation between views (which may comprise micro-images, subaperture images, or viewpoint images with high resolution) for achieving compression.

LF coding solutions based on inter-view prediction can be divided into two groups, which are distinguished according to the adopted representation format and coding

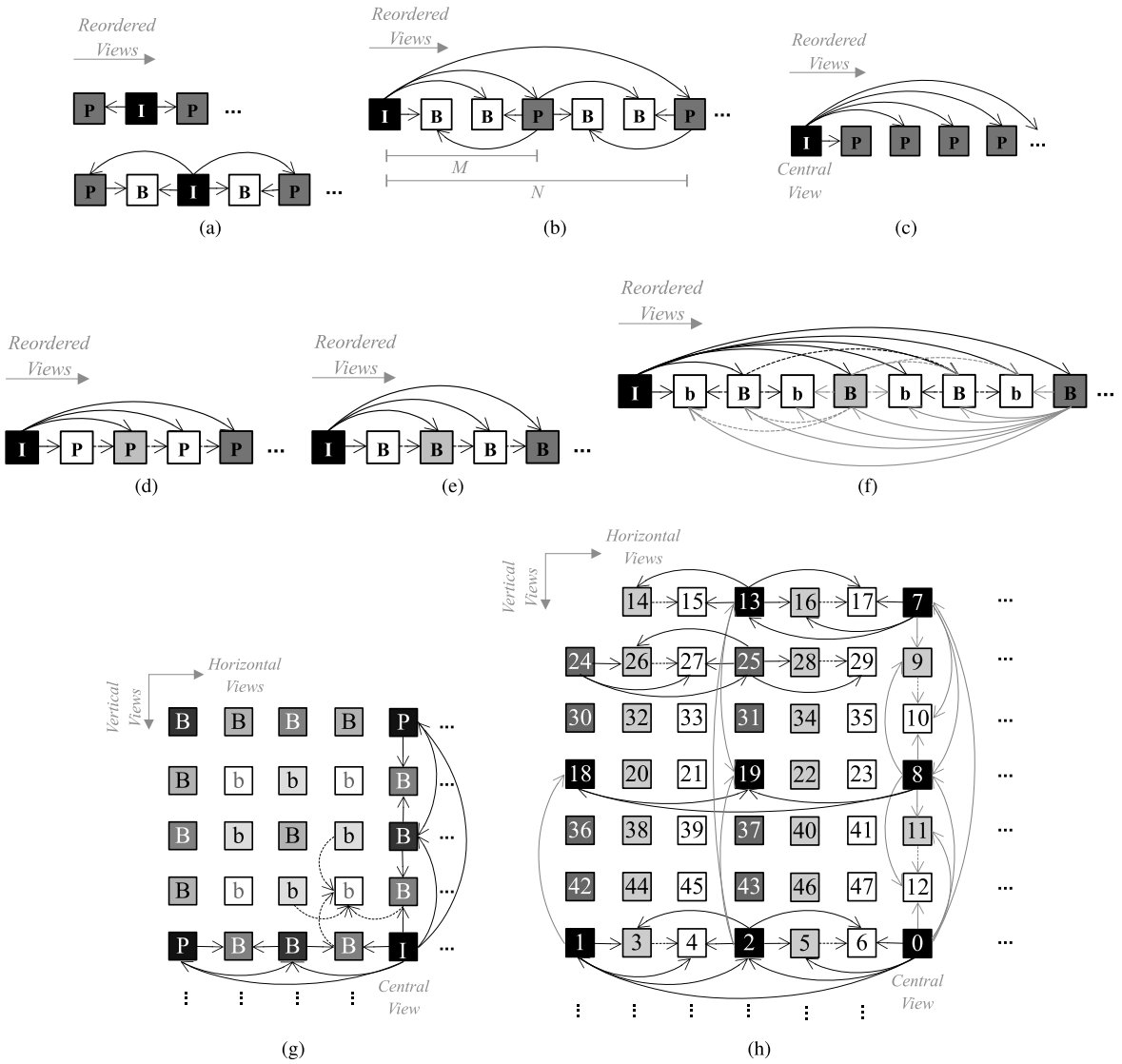


FIGURE 21. Prediction structures proposed in the literature for PVS based LF coding approaches. Different temporal layers are illustrated with different colors and capitalizing formats: (a) PIP (top) and PBI (bottom) [168]; (b) MPEG-2-based structure with $M = 3$ and $N = 6$ [170]; (c) Central 1D structure [173]–[175]; HEVC-based structures [176], [177]: (d) Low Delay P, (e) Low Delay B, and (f) Random Access; (g) 2D hierarchical prediction structure proposed by the winner of ICME 2016 grand challenge on LF compression [157]; and (h) Enhanced 2D hierarchical prediction structure proposed in [178], where the order for coding each view is depicted by the numbers inside the boxes.

architecture: i) PVS-based, in which the LF full parallax multiple views are organized in a PVS format (see Fig. 8b) and encoded with a hybrid 2D video coding solution (see Fig. 14); and ii) multiview-based, in which the LF full parallax multiple views are organized in a multiview format (see Fig. 11) to be encoded with a multiview video coding solution (see Fig. 15).

Although conceptually different, both PVS and multiview based coding approaches have the same basic purpose of proposing an efficient prediction configuration for better exploiting the inter-view correlations. For this, different scanning patterns for ordering the views (as exemplified in Fig. 9), as well as different prediction structures (as summarized in Figs. 21 and 22), are proposed.

1) PVS-BASED LF CODING

Back in 1995, the very first LF coding solution in the literature [166] (to the best of the authors' knowledge) proposed to introduce a Differential Pulse Coding Modulation (DPCM) coding into a DCT based image coding loop in order to encode lenticular-based lenslet images. For this, the lenslet image was organized as a PVS of micro-images and then encoded with the proposed codec, by using the previously encoded micro-image as the predictor. Since then, the PVS-based approaches proposed in the literature have naturally followed the evolution of hybrid 2D video coding standards.

In [139], a PVS-based scheme for coding light field acquired using a multi-camera array is proposed. The coding process starts by selecting a set of evenly distributed

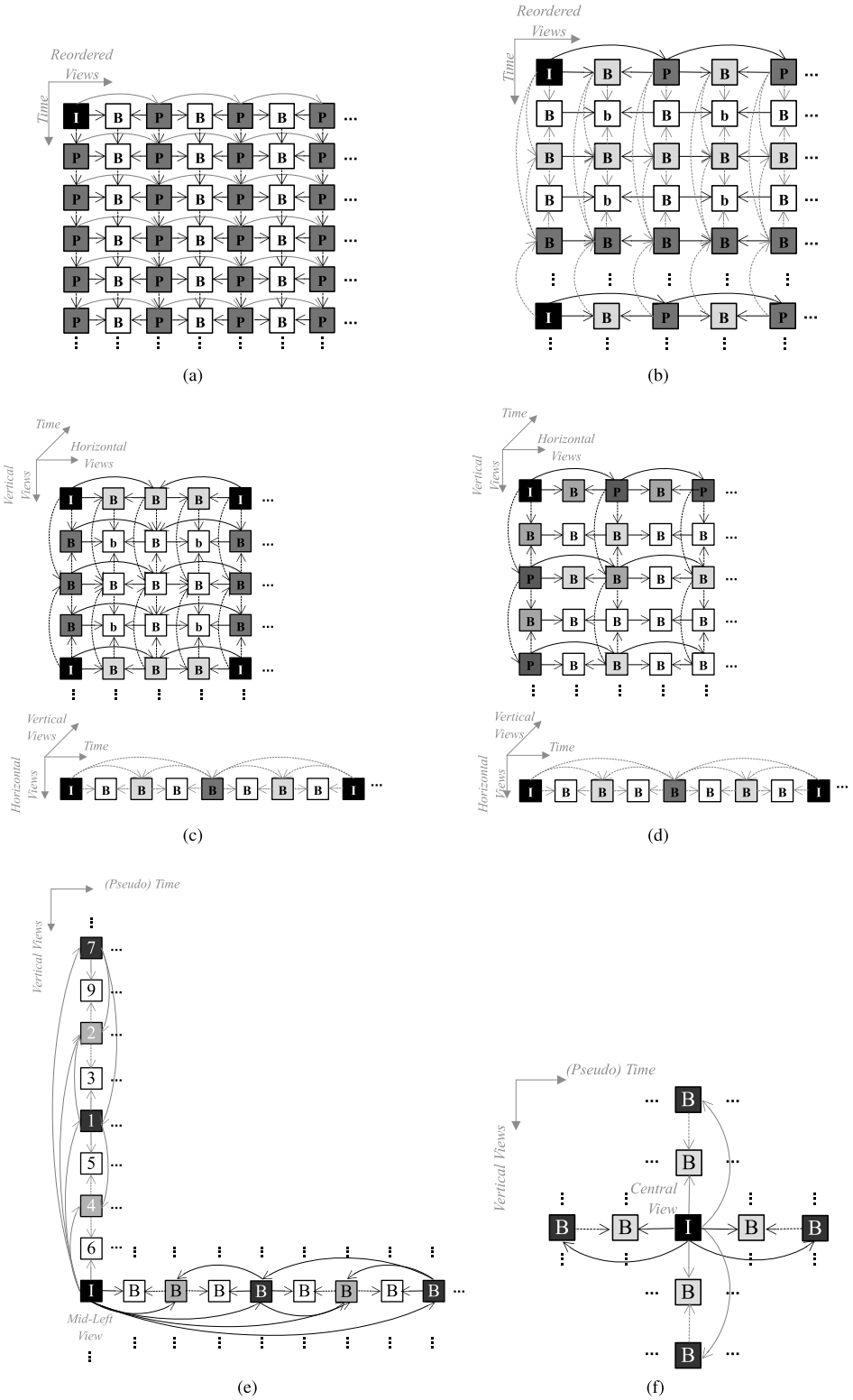


FIGURE 22. Prediction Structures proposed in the literature for multiview based LF coding approaches. Different hierarchical layers are illustrated with different colors and capitalizing formats: (a) IBP structure [196], [197]; (b) Typical prediction structure used in MVC; (c) 2D hierarchical inter-view prediction structure proposed in [198]; (d) 2D parallel inter-view prediction structure proposed in [199]; (e) 2D hierarchical prediction structure propose in [200]; and (f) 2D hierarchical prediction structure propose in [201].

views to be independently coded as intra or I-frames. Then, the remaining views are coded as P-frames by choosing one of the coded I-frames as a reference frame. In [167], a multi-hypothesis prediction scheme is proposed to improve the coding performance of the solution in [139] by using bi-predicted B-frames. Experimental results show 10% of bit savings compared against the solution in [139]. In [168], a combination of JPEG and MPEG-1 [169] standard codecs is used to encode lenslet LF content represented as a PVS of micro-images. For this, two prediction configurations are used to exploit the correlations between neighboring micro-images, as illustrated in Fig. 21a. It is shown that, the proposed hybrid coding solution always outperforms the JPEG standard (used for encoding the entire lenslet image, without micro-image extraction) with significant gains when the PBI configuration in Fig. 21a (bottom) is used. In [170], the PVS of micro-images is proposed to be encoded with MPEG-2 [171], using the prediction structure shown in Fig. 21a and three different scanning patterns: i) raster; ii) perpendicular; and iii) spiral (see Fig. 9). From the presented results, it is shown that the inter-view prediction using raster scan order is much less efficient than spiral and perpendicular due to the reduced number of available vertical inter-view predictions and increased distance between coding frame and reference frame(s). Due to similar reasons, a variation in the inter-view performance is also observed for different prediction configurations, i.e., for different M and N parameters in Fig. 21b. In [172], the authors propose to use a Hilbert scan (see Fig. 9f) for ordering the micro-images in a PVS, which is then encoded using MPEG-2 [171], as in [170]. It is shown that the Hilbert scan results in a significant improvement in RD performance when compared to the raster, perpendicular, and spiral scans (see Fig. 9).

In [173], [175], the authors propose to scan all subaperture images from a lenslet LF content in spiral order (see Fig. 9d) and to encode the resulting PVS with a combination of DPCM and the H.264/AVC. For this, the prediction structure depicted in Fig. 21c is used, in which the central subaperture image is considered the only reference frame available for coding all the remaining subaperture images. Then, the difference between the current subaperture image being coded and its reference frame is encoded as a frame using H.264/AVC. The proposed scheme outperforms a JPEG based solution, in which each subaperture image is independently coded with JPEG. Afterwards, in [179], the authors propose to improve the performance of the solution proposed in [173], [175] by replacing the previous DPCM based scheme by a motion compensated prediction, and to encode the motion compensated residual using H.264/AVC. Moreover, in [180], for further RD performance improvements, the authors propose to equally divide the array of subaperture images into four parts before scanning them in spiral order to form the PVS, so as to reduce the distance between a current subaperture image being coded and its reference frame.

In [181], a PVS is constructed by scanning all subaperture images from a lenslet LF content in raster order (see Fig. 9a), which is then encoded with H.264/AVC. The proposed encoder is then compared against JPEG and JPEG 2000, where both standard solutions are used to encode the entire lenslet image, without subaperture image extraction. From the presented results (for synthetic images computationally generated using a pinhole lens array model approximation [181]), JPEG 2000 is shown to be more efficient than JPEG for lenslet image coding. However, the PVS coding solution based on subaperture images outperforms JPEG 2000 with significant gains. In [182], the authors propose to extend their previous work [181] by considering raster, serpentine, spiral, and zig-zag (see Fig. 9) scan for ordering the subaperture images in the PVS. It is shown that, for synthetic lenslet images as in [181], changing the scanning patterns does not result in significant differences in the RD performance. Moreover, a comparison between the proposed PVS based on subaperture image and a PVS based on micro-images (both coded with H.264/AVC) is also performed for synthetic lenslet images generated with different micro-image and subaperture image sizes. It is suggested that the subaperture images shall be preferred to micro-image in a PVS when the subaperture image size is larger than the micro-image size. However, from the presented results, it is also shown that the difference in RD performance between these two PVS approaches (based on micro-images and subaperture images) may also depend on the scene characteristics. For instance, for LF generated using pinhole lens array model and a scene with highly detailed objects distributed in various depth planes, subaperture images are more difficult to encode than micro-images since near and far objects are equally noticeable in the subaperture image [182] (in other words, the objects size is invariant to depth in the subaperture image due to its orthographic property). Similarly, in [183], a PVS-based approach using H.264/AVC is proposed for coding LF images captured by a Lytro lenslet camera. In this case, the PVS is constructed using raster and spiral scanning topologies (see Fig. 9). Experimental results show significant coding gains compared to a solution in which JPEG is used to encode the entire lenslet image.

In [143], a performance comparison between a PVS based approach and a transform-based approach is proposed. For this, the PVS based approach using H.264/AVC in [182] is compared to a transform-based approach using the JP3D standard, in which a 3D DWT is applied to stacks of micro-images or subaperture images. It is shown in the results that the PVS-based approach (using H.264/AVC) outperforms the transform-based approach (using JP3D) with significant gains at lower bit rates, and produces less visible distortion. Also regarding a comparison against transform-based approaches, a hybrid coding solution combining a motion compensated prediction and a KLT transform coding is proposed in [174], and compared against the KLT-based approach proposed in [133], [134]. For the proposed hybrid solution, a PVS of subaperture images is extracted from the lenslet image in spiral scan order, and the central subaperture image is used as the reference frame for encoding

all the remaining subaperture images (see Fig. 21c). Then, the Normalized Cross-Correlation (NCC) [184] is used as the matching criterion for the motion estimation process and, afterwards, the residual information (from motion compensation) is encoded using the KLT-based coding scheme proposed in [133], [134]. From the experimental results, it is shown that the proposed hybrid solution achieves significant bit savings when compared to the transform-based solution in [133], [134].

In [185], a performance study of HEVC compatible coding solutions for lenslet images captured by the Lytro Illum camera is presented. For this, subaperture images are extracted and organized in a PVS using two different scan orders — raster and spiral (see Fig. 9). The PVS is encoded with HEVC using three different prediction structures — Low Delay P, Low Delay B, and Random Access [177] (see Fig. 21d to 21f). The different PVS based approaches are also compared to the case where the rectified lenslet image (i.e., after calibrating and converting to a rectangular lenslet grid) is encoded in its entirety using HEVC Still Picture Profile [85]. From this study, it is shown that, in general, the PVS-based solution outperforms the HEVC Still Picture Profile coding the entire rectified lenslet image. In addition, it is seen that the relative RD performance of the PVS-based coding approaches is not consistent for all images, and, consequently, it is not possible to reach a general conclusion regarding which would be the best scanning topology and prediction structure to be used in the PVS-based coding scheme. Similar PVS-based schemes for lenslet LF coding have also been proposed in [186], [187] using different scanning topologies to order the subaperture images which are then coded with HEVC. In [188], a study is presented comparing the performance of different image/video coding standards — i.e., JPEG 2000, H.264/AVC, HEVC, Google VP9 [189], and the Joint Exploration Model (JEM) codec [190] (which was the starting point of the future VVC standard) — for coding LF content acquired using a lenslet camera and a camera gantry. Results using LF images from the *EPFL* and *Stanford LF Datasets* (see Table 1) show that a PVS-based solution with HEVC or VP9 significantly outperforms JPEG 2000 and H.264/AVC. However, a PVS-based solution using JEM presents an impressive compression efficiency with about 43% of bit savings compared to HEVC.

In [191], a PVS-based approach using JEM is proposed for coding LF images captured by a Lytro lenslet camera. In this case, the PVS is constructed using a U-shaped scanning topology (see Fig. 9g) and a hybrid topology combining serpentine and U-shaped (see Fig. 9b and 9g) scans. Experimental results show significant coding gains compared to coding the lenslet image with JPEG, and to a solution in which JEM is used to code independently each subaperture image. Similarly, in [192], the same PVS-based approach in [191] is adopted, but including also an enhanced block-based illumination compensation and an adaptive filtering of the reconstructed subaperture images. Experimental results for the *EPFL LF Dataset* (see Table 1) show that further bit savings, of 3.5% in average, can be achieved compared to the solution in [191]. In [193], a PVS scheme is proposed in which the lenslet image is partitioned into tiles of equal size that are scanned in serpentine order (see Fig. 9b). The tile-based PVS is then coded with HEVC using the configuration Low Delay B (see Fig. 21e). The proposed solution achieved significant coding gains at low bit rates when compared against a solution using JPEG standard for coding the entire lenslet image. In [157], a PVS is constructed by organizing the subaperture images in the 2D hierarchical structure illustrated in Fig. 21g, which are then coded using the JEM codec [190]. Similar to what is done in conventional HEVC and JEM inter coding, each subaperture image in the PVS is assigned to a different temporal layer. The lowest layer is assigned to the central subaperture image that is compressed as an I-frame. The remaining subaperture images are then compressed as P- or B-frames by using the nearest subaperture images (at top, bottom, left, and right directions, as illustrated in Fig. 21g) as reference frames. From the presented results, significant coding gains were observed (average of 4.5 dB) compared to a solution using JPEG standard for coding the entire lenslet image. This solution in [157] was selected as the winner of the ICME 2016 grand challenge based on the adopted objective and subjective quality evaluation criteria [194] using the *EPFL LF Dataset* (see Table 1). In [178], the authors propose to improve their previous work by equally dividing the array of subaperture images into four parts and coding them independently using an enhanced 2D hierarchical prediction structure depicted in Fig. 21h. Additionally, an optimal bit allocation for the proposed prediction structure is presented. Results for the *EPFL LF Dataset* (see Table 1) show that their improved solution is able to achieve 18.3% of bit savings compared to their previous solution in [157].

LF video can also be coded with a PVS-based approach by using the called transposed picture ordering as proposed in [62]. However, it is worthwhile to note that the temporal correlation between adjacent time instants no longer exists in the final video sequence since all views from the same time instant are concatenated along the time dimension.

2) MULTIVIEW-BASED LF CODING

In this case, the LF content is organized as a conventional 3D multiview format (see Section II-B3) to be coded using standard 3D video coding solutions — such as MVC [195], and MV-HEVC [88]. This solution can be used in alternative to the PVS-based solutions to exploit the correlations in LF content in all dimensions. For coding LF still content with full parallax, the temporal axis of the multiview representation can be used to organize the fourth LF dimension, as discussed in Section II-B3. For representing LF videos, the 2D array of viewpoints are usually scanned using a specific topology (see Fig. 9) to form a 1D array of multiview videos. Then, different to the transposed picture ordering [62], the temporal correlation can be fully exploited in this scheme.

In [197], [202], the authors propose to decompose the lenslet LF video into multiple subaperture video sequences

scanned in raster order (see Fig. 9a) and to jointly exploit motion (temporal prediction) and disparity (inter-view prediction) similarly to what is done in MVC [195]. For this, the prediction structure depicted in Fig. 22a is adopted, and the Evolutionary Strategy (ES) is used to speed up the motion estimation process. In [196], the authors significantly improve their previous work by using motion estimation with half pixel precision. In [203], the authors propose to scan the subaperture video sequences in raster order (see Fig. 9a) and to encode it with MVC by using the conventional prediction structure used in MVC as shown in Fig. 22b. It is seen that the proposed solution outperforms a H.264/AVC based coding solution, in which the LF video sequence is encoded in its entirety using H.264/AVC. It is worthwhile mentioning that these coding schemes [196], [197], [202], [203] consider only lenticular-based content (i.e., only horizontal parallax) with a small number of subaperture images (up to eight). The same approach as in [189] is proposed in [190] for coding synthetic lenslet LF video with 3×3 subaperture sequences. From the presented results, the multiview-based arrangement outperforms a PVS-based approach (in which the subaperture sequences are reordered using transposed picture ordering [62] and encoded with H.264/AVC), as well as a H.264/AVC based coding solution (in which the lenslet LF video sequence is encoded in its entirety using H.264/AVC) with significant RD gains.

In [198], a hierarchical 2D inter-view prediction structure, shown in Fig. 22c, is proposed for lenslet LF video coding using MVC. The idea is to optimize the inter-view prediction structure to the 2D grid of subaperture sequences, and then to further minimize the distance between the current subaperture image and its inter-view reference frame(s). The proposed hierarchical prediction structure is compared against the Hilbert scan (see Fig. 9f) of subaperture images proposed in [172], and presents expressive RD performance improvements. Moreover, a parallel implementation of the proposed prediction structure is also designed, which significantly reduces the overall encoding time. Similarly, in [199], a 2D parallel inter-view prediction structure, shown in Fig. 22d, is proposed for coding lenslet LF video using MVC. The proposed prediction structure is compared against the conventional MVC prediction structure (in Fig. 22b) and against a spiral scan (see Fig. 9d) of subaperture images proposed in [204], in which the central subaperture image is considered the only inter-view reference for coding all the remaining subaperture images (similar to the prediction structure shown in Fig. 21c). From the presented results (for stop motion video sequences captured using the Lytro 1st generation camera [2] [9]), it is seen that the proposed solution outperformed the other two tested solutions, except at low bit rate values, where it is outperformed by the MVC prediction structure in Fig. 22b. To improve the RD performance, a rate allocation scheme is proposed to efficiently assign the Quantization Parameter (QP) to the multiview sequences. As discussed in Section II-B2, for LF images acquired using a focused lenslet camera, the texture resampling from micro-images to the subaperture images usually results in very low-resolution images with significant aliasing artifacts [73]. Motivated by this fact, an alternative multiview based data arrangement using views with higher resolutions is proposed in [20], [205] which are then coded using MV-HEVC. In this scheme, a scalable coding architecture is also proposed in which lower layers comprise multiple views with higher resolution while the last enhancement layer comprises the entire lenslet image.

More recently, a multiview-based solution is proposed in [200] using MV-HEVC for coding lenslet LF images. For this, the subaperture images are organized as multiview video sequences, as depicted in Fig. 11, which are then coded with MV-HEVC using the 2D hierarchical structure illustrated in Fig. 22e. Experimental results using the *EPFL LF Dataset* (see Table 1) show that the proposed MV-HEVC solution significantly outperforms a PVS-based solution in which all views are scanned in spiral order and coded with HEVC, achieving 2.4 dB of coding gains in average. In [201], the authors propose to improve the coding efficiency of their previous work in [200] by using optimized prediction and rate allocation schemes for coding LF images acquired using a lenslet camera and a multi-camera array. In this case, the 2D prediction scheme in Fig. 22f is adopted, in which the central subaperture image is coded independently as an I-frame, and the remaining subaperture images are coded using an estimated QP for each frame by considering its distance, prediction level, and decoding order with respect to the central subaperture image. Experimental results are shown using LF images from the *EPFL LF Dataset* and the *Stanford LF Archive* (see Table 1) and comparing the proposed solution against: i) a PVS-based solution using serpentine scan and HEVC inter coding; ii) the PVS-based solution in [157] (but using HEVC instead of JEM), iii) their previous multiview-based solution in [200]; and iv) a LF coding solution based on non-local spatial prediction in [206]. From these results, it is shown that the proposed solution outperforms all of the other LF coding solutions with significant coding gains against the PVS-based solutions (0.91 dB, for lenslet images, and 1.51 dB, for LF images capture with a multi-camera array, in average compared to the solution with serpentine scan) and the non-local spatial prediction based solution in [206]. Compared against their previous solution in [200], the proposed solution presents slightly better RD performance at low bit rates.

3) DISCUSSION

Coding based on inter-view prediction has been the most popular approach proposed in literature for dense LF coding so far. Generically, it has been shown that, using standard 2D or 3D video coding solutions, it is possible to achieve competitive RD coding performance for LF content acquired with either lenslet, multi-camera or gantry LF setups. Moreover, the coding approaches in this category leave open the possibility of a huge variety of data arrangements and prediction structures for better exploiting the LF correlations.

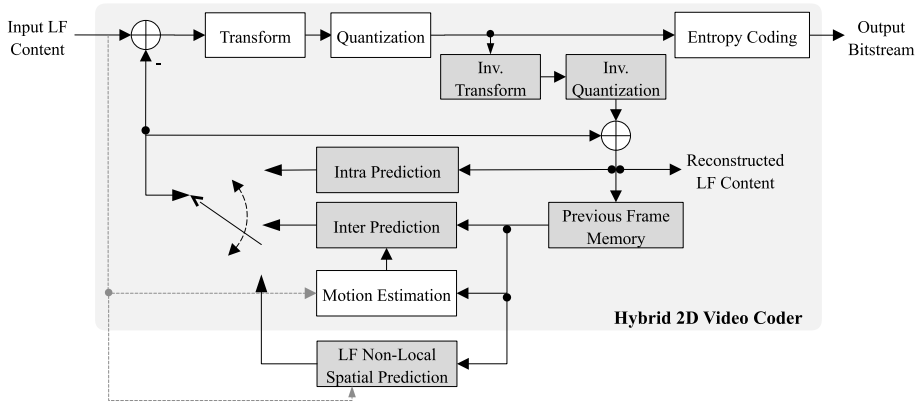


FIGURE 23. LF image and video coding architecture based on non-local spatial prediction.

From the different prediction structures proposed in the literature, 2D hierarchical structures [157], [178], [198], [199] have shown to achieve the best RD performance for LF still images using both PVS and multiview-based coding solutions. However, an advantage of using multiview-based solutions is that scalability and backward compatibility are straightforwardly supported using a 3D coding standard. In addition, multiview-based solutions can be easily extended for coding LF videos.

Compared to lenslet LF coding solutions based on non-local spatial prediction, multiview-based solutions [200] are seen to perform better, especially for coding lenslet LF images captured using the unfocused lenslet camera. Moreover, in [24], the multiview-based solution proposed in [200] shows consistently better objective and subjective RD performance in low bit rates, than two LF coding solutions based on view synthesis in [126], [207] (see Sections III-D1 and III-D3), selected as the winners of the ICIP 2017 grand challenge on LF compression. However, it should be noticed that, the expressive RD gains come with the price of very complex prediction structures, which is undesirable for supporting efficient random access.

C. LF CODING BASED ON NON-LOCAL SPATIAL PREDICTION

Also trying to exploit the LF correlations in a predictive manner, the LF coding solutions in this category propose to exploit the non-local spatial correlation that exists when the lenslet representation is adopted (see Fig. 7b). For this, the lenslet image is encoded in its entirety with a hybrid 2D video codec and by using a special LF non-local spatial prediction as illustrated in Fig. 23.

LF coding approaches in this category can be separated in two groups, depending on the type of non-local spatial prediction that is used (see Fig. 17): i) spatial compensated prediction; and ii) learning-based prediction.

1) SPATIAL COMPENSATED PREDICTION

The idea of exploiting non-local spatial redundancy has been firstly proposed for 2D image and video compression to further enhance the performance of H.264/AVC intra prediction. Notably, the intra macroblock compensation technique [208], [209] proposes to extend the usage of motion compensated prediction to intra I-frames in order to reduce the number of bits needed in conventional intra coding while still supporting random access [209].

Lenslet LF content compression using non-local spatial prediction has been firstly proposed in [210] with the purpose of exploiting the existing micro-image cross-correlation of this type of content to improve the performance of H.264/AVC standard for lenslet image coding. In [211], the Self-Similarity (SS) prediction is proposed to be added to the HEVC coding architecture, so as to take advantage of the flexible partition patterns used in this type of video codecs. Similarly to motion compensation, the SS estimation process uses a block-based matching over a causal search window (i.e., a search window containing only previously coded pixels, as seen in Fig. 24a), to find the 'best' predictor for the current block, in an Rate-Distortion Optimization (RDO) sense. As a result, the relative position of the chosen predictor block is signaled by a displacement vector (see Fig. 24a), referred to as SS vector. In [212], the predictor block is generated from a single candidate block — referred to as uni-SS prediction — and the resulting SS vector can be either encoded explicitly, similar to motion vectors in HEVC or using an SS-skip mode which creates a list of candidate with the SS vectors used to encode neighboring blocks (in the left, above, and above-left positions). To take advantage of the distinctive characteristics of these SS vectors, a novel SS vector prediction scheme, referred to as Micro-Image Vector Prediction (MIVP) is also proposed [212], in which a list of candidate vectors from neighboring micro-images (in the left, above, and above-left positions) are used to predict the SS vectors. Experimental results show that it is possible to achieve significant bit savings compared to HEVC when coding lenslet LF images (38.4% in average) and lenslet LF videos (22.9% in average at low bit rates).

Although not targeting lenslet image coding, a scheme very similar to the SS prediction, known as Intra Block Copy (IntraBC) [213], has been proposed in the context of

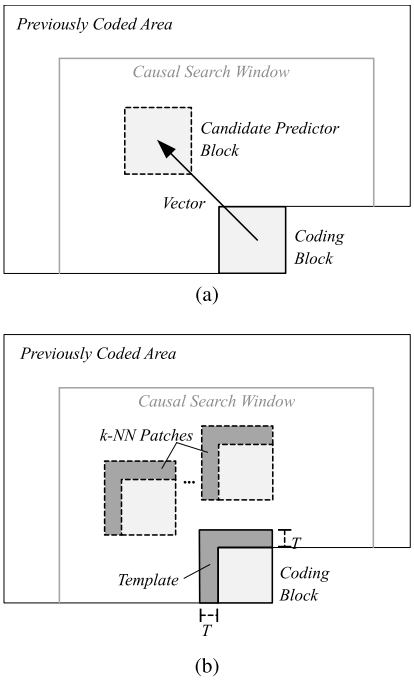


FIGURE 24. Non-local spatial prediction approaches: (a) Spatial compensated prediction; and (b) Learning-based prediction.

HEVC Screen Content Coding (SCC) [213] extension. Firstly proposed in [214], it aims at improving the HEVC coding efficiency for screen video compression, motivated by the fact that this kind of content often contains a substantial amount of still or moving rendered graphics and texts with repetitive patterns. The IntraBC also uses a block-based matching algorithm to estimate a single displacement vector that indicates the relative position of the predictor block to the current block being coded. However, the estimated vector uses only integer pixel accuracy. An improved IntraBC version is proposed in [215], [216], in which the search window is expanded to the entire Coding Block (CB) row or column (for 16×16 CBs), or to the entire previously coded area of the picture by using a hash-based search (for 8×8 CBs).

In [206], [217], the authors propose to extend the SS prediction concept by using HEVC inter B-frame bi-prediction to further improve the RD performance for LF image coding. However, in this case, to guarantee that the two prediction signals come from two different MIs, the search area is separated into two non-overlapping parts. In [218], [219], the authors propose to extend their previous work in [212] with a bi-SS prediction. Different from the solution in [206], [217], the predictor block is generated from a combination of two candidate blocks jointly estimated from the same search area. Experimental results are shown using the *EPFL LF Dataset* (see Table 1) comparing against: i) HEVC SCC; ii) the uni-SS prediction in [212]; iii) the spatial compensated prediction in [206], [217]; and iv) a learning-based spatial prediction proposed in [220] (see Section III-C2). From these results, it is shown that the proposed solution outperforms all of the other solutions with bit savings of (in average): 32.0% against HEVC SCC; 14.4% against the uni-SS prediction in [212]; 9.4% against solution in [206], [217]; and 7.8% against the solution in [220]. In [221], a weighted SS prediction is proposed in which an adaptive set of weighting coefficients are used for combining the two jointly estimated candidate blocks. Experimental results for the *EPFL LF Dataset* (see Table 1) show that it is possible to achieve 3.4% of bit savings comparing against the bi-SS prediction in [218], [219].

In [222], a lenslet image reshaping scheme, referred to as macro-pixel, is proposed to align the micro-image structure in the Lytro Illum lenslet image with the coding grid of a block-wise image/video coding standard. The purpose was to reach further bit savings when coding a lenslet image with any HEVC-based solutions, such as: i) HEVC SCC standard; and ii) a learning-based prediction proposed in [223] (see Section III-C2). From the results, it is shown that it is possible to achieve in average 11.6% and 9.8% of bit savings when applying the reshaping prior to coding the lenslet image with, respectively, HEVC SCC and the solution in [223]. In [224], the authors propose to extend their previous work with three spatial compensated prediction modes, referred to as Boundary Matching Prediction (BMP), Multi-block Weighted Prediction (MWP) and Co-located Single-block Prediction (CSP) [225]. In the MWP mode, an adaptive set of weighting coefficients are used for combining four candidate blocks selected from neighboring macro-pixels (in the left, above, above-left, and above-right positions) to generate the predictor block. In the BMP mode, a linear weighted prediction is also used, but, in this case, only the boundary samples of the candidate blocks are used to derive the weighting coefficients. Differently, in the CSP mode, the predictor block is directly selected from a list of candidate neighboring blocks (in the left, above, above-left, and above-right positions), similar to the SS-skip mode proposed in [212]. Experimental results are shown for coding lenslet LF images from the *EPFL LF Dataset* (see Table 1) and comparing the proposed solution against: i) HEVC SCC standard; ii) a learning-based prediction proposed in [223] (see Section III-C2); and iii) the PVS-based scheme proposed in [193] using tiling of lenslet images. From the results, it is seen that the proposed solution can achieve, in average, bit savings of 37.2% against HEVC SCC, 37.5% against the learning-based prediction in [223], and 10.5% against the PVS-based scheme in [193]. It is worth noting that the BMP mode has been considered in MPEG-I exploration experiments [60] for coding LF content using a lenslet representation and is able to achieve an additional 1.29% of bit savings (in average) when included into the HEVC SCC standard for coding LF video [60].

In [226], a high order compensated prediction is proposed to exploit the fact that neighbor micro-images represent a portion of the scene captured from slightly different perspectives which may not be explored with a translational compensated prediction. For this, geometric transformations with up to 8 degrees of freedom (namely, affine, bilinear and projective) are used to map perspective changes from the block being coded to the causal search window shown

in Fig. 24a). To decrease the computational complexity when choosing the transformation parameters, the SS estimation is applied in a first stage and the best SS candidate block is used as a starting point for applying the geometric transformations. As a result, a set of four vectors is coded and transmitted, each of them defining the position of one corner of the (transformed) prediction quadrilateral.

Although these coding schemes have shown to achieve significant compression gains when compared to state-of-the-art 2D image coding solutions, the proposed 2D-based coding architecture does not support functionalities, such as, scalability, random access, and ROI coding. Motivated by this fact, a lenslet LF coding architecture is proposed in [221], [227] for supporting FoV and quality scalability, random access, and ROI coding. The proposed coding architecture comprises a base layer compliant with HEVC standard, complemented by one or more enhancement layers that progressively support richer forms of the same LF content by hierarchically organizing the angular information of the captured lenslet image. Each enhancement layer is encoded using the bi-SS prediction in [218], [219] and other exemplar-based inter-layer predictions. From the presented results using the *EPFL LF Dataset* (see Table 1), it is shown that the proposed scalable design provides flexibility in the bitstream at no rate cost (in average) compared to HEVC Still Picture Profile.

2) LEARNING-BASED SPATIAL PREDICTION

Learning-based prediction methods have been increasingly considered for still image compression [228], being especially powerful for predicting highly complex textured areas of the image. The idea of this type of prediction is basically to search for an optimized combination of k texture patches that best approximate the sample values of the coding block. These texture patches can be selected from a previously coded neighborhood of the coding block (known as template), as depicted in Fig. 24b, or from a previously learnt dictionary of samples. The former is referred to here as neighbor-embedding prediction and the latter is referred here to as dictionary-based prediction.

In the neighbor-embedding prediction, an optimized linear combination of k-Nearest Neighbor (k-NN) texture patches is estimated and, then, the resulting weighting coefficients are used to approximate the unknown samples in the coding block (see Fig. 24b). Examples of neighbor-embedding prediction methods proposed in the literature for image coding are the Non-negative Matrix Factorization (NMF) [229] and the Locally Linear Embedding (LLE) [230] dimensionality reduction techniques. Moreover, the Template Matching (TM) [231] algorithm can be seen as a particular case of neighbor-embedding prediction, in which a unique 1-NN texture patch is found with the linear weighting coefficient equal to 1 [228].

With regard to lenslet image compression, the work in [232] proposes to replace one of the conventional intra directional prediction modes of HEVC by a prediction scheme based on TM for better adapting to the repetitive micro-image texture patterns. Similar to TM, the proposed prediction method uses an implicit approach to avoid transmitting any information about the used predictor. In this scheme, three neighboring CBs are used as the template and two separate search windows are adopted for finding two best predictors to this template. From the selected predictors, two 1D vectors are derived, and their combination determines the block predictor for the current CB. It is shown that the proposed scheme outperforms the TM algorithm (when including the TM in the HEVC coding framework) for lenslet image coding, being able also to considerably reduce the computational complexity for the same search window.

In [223], an LLE prediction is propose for lenslet LF coding based on HEVC. In this case, the predictor to the current CB is given as a linear combination of its k-NN patches inside a causal search window in the lenslet image (see Fig. 24b). To avoid transmitting information about the selected k-NN patches, the LLE method searches for them, in terms of Euclidian distance, at both encoder and decoder side. Afterwards, the set of weighting coefficients for combining the k-NN patches are determined by solving a least-squares optimization problem with a constraint on the sum of the coefficients that must be 1. After finding the optimally estimated coefficients, the predictor block is determined by using the same linear coefficients estimated for the template to combine the square blocks associated to each k-NN patch (see Fig. 24b). For improved performance, the encoder tests different k values, from 1 up to 8, and the one that produces the best block prediction result (in RDO manner) is explicitly transmitted to the decoder. To avoid further signaling in the HEVC bitstream, up to 8 HEVC intra directional modes are replaced by the LLE based mode using a different number of k-NN patches. From the presented results, it is shown that the proposed LLE-based coding solution always outperformed HEVC Still Picture Profile (with bit savings of up to 38%) and the SS-based solution proposed in [211], [233] (up to 15% of bit savings when the number of k-NN templates is adaptively chosen by varying from 1 to 8). In [234], the LLE-based prediction [223] is combined with the uni-SS prediction [212] (referred here to as LLE+SS solution) to further improve the coding performance for lenslet image coding. Results using the *EPFL LF Dataset* (see Table 1) show that it is possible to achieve 16.9% of bit savings in average compared against the uni-SS solution [212].

In [220], the authors propose an HEVC-based neighbor-embedding predictive solution using Gaussian Process Regression (GPR) for lenslet image coding. Similarly to [223], k-NN patches are firstly chosen in terms of Euclidian distance in a causal search window. However, in order to reduce the computational complexity, the causal search window is divided into two different search windows (horizontal and vertical search windows, as in [232]), and the template thickness, T (see Fig. 24b), is substantially reduced. Then, a filtering method based on the NCC [184] is used to judge

the reliability of the obtained k-NN patches. Afterwards, the prediction from the k-NN patches is modeled as a non-linear (Gaussian) process, and GPR is then used for estimating the predictor block. The GPR-based neighbor-embedding prediction is then included into HEVC by replacing one of the HEVC intra directional modes and no further signaling is needed. The proposed GPR based is compared against HEVC SCC [213], as well as against the TM-based solution proposed in [232] and the LLE-based solution in [223] (but fixing 6-NN patches, instead of adaptively varying "k" from 1 to 8 as in [223]). It is shown that the proposed GPR based solution outperforms the TM-based solution [232] and HEVC SCC (with up to 21% of bit savings) with significant coding gains, and always outperforms the LLE-based solution (with up to 5% of bit savings), showing that improved prediction results could be obtained by using GPR instead of LLE for texture and edge regions.

In [235], an ℓ_1-optimized prediction is proposed for coding lenslet LF content with hexagonal-shaped micro-images (such as, the lenslet images captured with Lytro cameras). In this case, HEVC is adapted to consider the micro-image as the elementary coding unit — referred to as macro-pixel. Hence, the current macro-pixel is predicted as a linear combination of three previously coded neighboring macro-pixels, corresponding to the top, left and top-left macro-pixels. The weighting coefficients are chosen by minimizing the ℓ_1-norm of the residual with a constraint on the sum of the coefficients that must be 1. For coding the weighting coefficients more efficiently, they are chosen from 32 possible sets of weights. To further improve the coding performance, a modified directional intra prediction replaces the original HEVC intra prediction modes. From the presented experimental results using the *EPFL LF Dataset* (see Table 1), it is shown that the proposed ℓ_1-optimized solution always outperforms HEVC Still Picture Profile (with bit savings of 59.6% in average) and the PVS-based solution winner of ICME 2016 grand challenge in [157] (with bit savings of 30.9% in average). More recently, in [236], the authors propose to extend their previous work [236] by including a dictionary-based prediction scheme. In this case, the prediction of a current macro-pixel is found as a linear combination of samples in a learnt dictionary. To decrease the complexity for learning the dictionary, a sparse dictionary is considered, in which a fixed base dictionary with large samples is multiplied by an adaptable sparse matrix which selects which samples of the dictionary are considered. The base dictionary was trained on 4 images different than the images used to assess the coding performance. The "best" prediction mode between dictionary-based, ℓ_1-optimized, and directional prediction is chosen in an RDO manner. The proposed dictionary-based solution leads to significant bit savings compared against the LLE+SS solution in [234] (with 54.8% of bit savings), the PVS-based solutions in [157], [178] (with 49.7% of bit savings), and a solution based on view synthesis (see Section III-D3) in [237] (with 47.9% of bit savings).

3) DISCUSSION

LF coding solutions using non-local spatial prediction have been specifically proposed for coding lenslet LF content. The advantage of this prediction scheme is that it enables exploring the particular correlation of lenslet LF content without requiring any explicit knowledge about the used optical system, being less dependent on a very precise calibration pre-process [212]. Generically, it has been shown that, by integrating a non-local spatial prediction in a 2D coding solution, it is possible to achieve significant coding gains compared to standard 2D coding solutions, such as HEVC Still Picture Profile. However, it has been shown that using the explicit knowledge of the MLA structure — as in the lenslet LF coding solutions based on macro-pixel reshaping in [222], [224], [225], [235], [236] — may result in further coding gains.

Comparing different non-local spatial predictions, learning-based techniques, such as the dictionary-based solutions in [235], [236], have been shown to perform significantly better than solutions based on spatial compensated prediction. With respect to LF coding solutions in other categories, exploration experiments performed in the context of the MPEG-I standardization activity [60] show that a multiview-based LF coding solution with a 2D hierarchical prediction structure performs significantly better than coding the lenslet content using HEVC SCC. Nevertheless, the dictionary-based solution in [236] achieves competitive RD performance compared to PVS-based [157], [178] and view synthesis-based [237] solutions.

However, apart from the work in [221], [227], none of the coding solutions in this category (to the best of the authors' knowledge) have addressed supporting other functionalities, such as scalability, random access and ROI coding. Therefore, a question that remains open is if it is possible to support these functionalities without sacrificing the coding performance.

D. LF CODING BASED ON VIEW SYNTHESIS

Differently from the previous predictive LF coding approaches, the coding solutions in this category aim at greatly reducing the amount of LF texture data that is encoded and transmitted to achieve compression. For this, the LF data is represented by a sparse set of key views plus geometry information (see Section II-B5), which are then encoded and transmitted. Hence, at the decoder side, the sparse set of views plus geometry data are used to synthesize the information discarded at the encoder side. Since the reconstructed views are, in many cases, used as a reference frame for prediction, the LF coding solutions in this category are also connected to the node of predictive solutions in Fig. 17.

The proposed LF coding solutions in this category are divided in three main groups, according to the approach used for view synthesis (see Fig. 17): i) synthesis using Depth Image Based Rendering (DIBR); ii) transform-assisted synthesis; and iii) learning-based synthesis.

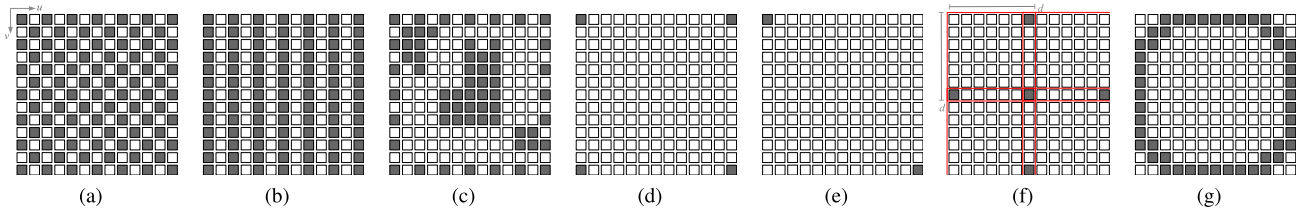


FIGURE 25. Examples of sparse sampling views (in shaded gray blocks) for LF coding solutions based on view synthesis: (a) *k*-Uniform; (b) *k*-Colunm uniform; (c) Adaptative; (d) 4-Corner; (e) 2-Corner; (f) Overlapping cross; and (g) Circular.

1) DIBR-BASED VIEW SYNTHESIS

LF coding schemes in this group synthesize discarded views at the decoder side by simply applying a disparity shift or by using DIBR techniques [238]. Essentially, DIBR techniques use camera calibration information, a sparse set of views, and associated depth map to perform image-based 3D warping, projecting the original views into the 3D space. Then, the resulting 3D world points are forward projected into an intermediate view position to be synthesized. The projected points are then merged and an inpainting algorithm is used to fill missing areas due to disocclusion problems or due to quantization errors when lossy encoding the disparity/depth information. Afterwards, a filtering process can also be used to provide a more natural appearance to the synthesized view [238].

In [239], [240], the authors propose to represent the LF data by a sparse set of micro-images that are uniformly subsampled from the lenslet image, as illustrated in Fig. 25a, to remove the redundancy between neighboring micro-images for achieving compression. Thus, the discarded micro-images are synthesized at the decoder side by simply using the optical geometry used when acquiring the LF content. The proposed scheme is able to improve the RD compression performance when incorporated into the JPEG standard. In [241], [242], the disparity between adjacent micro-images is used to better reconstruct discarded micro-images in the sparse set of micro-images. In [241], JPEG is used as the texture coder and lossless arithmetic coding is used for the disparity data. In [242], the sparse set of micro-images is represented as multiview content and each micro-image is then encoded using a method similar to MVC simulcast coding [195]. Moreover, the disparity is lossless encoded using a run-length coding scheme followed by Huffman coding. At the decoder side, the discarded micro-images are reconstructed by applying a disparity shift, in [241], and by using a DIBR algorithm modified to support the multiple micro-images as input views, in [242], followed by an inpainting algorithm to fill in the missing areas. In [241], the disparity is assumed to be the same for all pixel positions inside a micro-image. On one hand, this can be a valid approximation since each micro-image has a small FoV. On the other hand, this assumption is likely to be inaccurate at object boundaries since a single micro-image can still capture (small) portions of objects in different depth planes. Hence, the quality of the reconstructed micro-images — and, consequently, the quality of rendered views — is severely affected by these disparity inaccuracies at the encoder side. For this reason, instead of uniformly selecting the micro-images in the lenslet image, the selection is carried out adaptively, as exemplified in Fig. 25c, so as to obtain better view reconstruction [241], [242]. In [241], an iterative selection of micro-images is performed based on a cumulative disparity metric. In [242], a visibility test is used to select extra micro-images to be encoded and transmitted by identifying possible hole-causing regions.

However, a common characteristic of these approaches is that the quality of rendered views is negatively affected by inaccuracies in the synthesis of missing micro-images. The reconstruction artifacts are even more challenging for synthesizing micro-images due to their small FoV and resolution (compared to view synthesis in conventional depth assisted 3D coding solutions). For this reason, in [243], an alternative coding architecture is used, as depicted in Fig. 26, in which the entire lenslet image is also encoded and transmitted in an LF enhancement layer so as to provide rendered views with better quality (i.e., rendered from the content in the LF enhancement layer). In this case, an HEVC-based coding scheme is used for encoding the sparse set of micro-images (in the base layer), as well as the disparity information represented as 2D images. Then, the coded texture and disparity are used for reconstructing the lenslet image, which is later used as a reference frame for coding in the LF enhancement layer. From the experimental results, it is shown that the proposed disparity-assisted solution presents significant bit savings (up to 65% when subsampling the grid of micro-images by a factor of 2) compared to encoding the entire lenslet image with HEVC Still Picture Profile. A substantial difference in objective quality between the reconstructed LF content in the lower layers and the LF enhancement layer content is also observed.

An alternative DIBR-based coding approach is proposed in [244] also using the coding architecture in Fig. 26, but, in this case, a sparse set of views is rendered from the lenslet image and then encoded together with the disparity information. For estimating the disparity, the block-based matching algorithm proposed in [65] is adopted, in which a single 4-bit value of disparity is computed for each micro-image. Then, this disparity information is used to render a single view from the lenslet image by using the disparity-assisted weighted blending algorithm proposed in [65]. For encoding

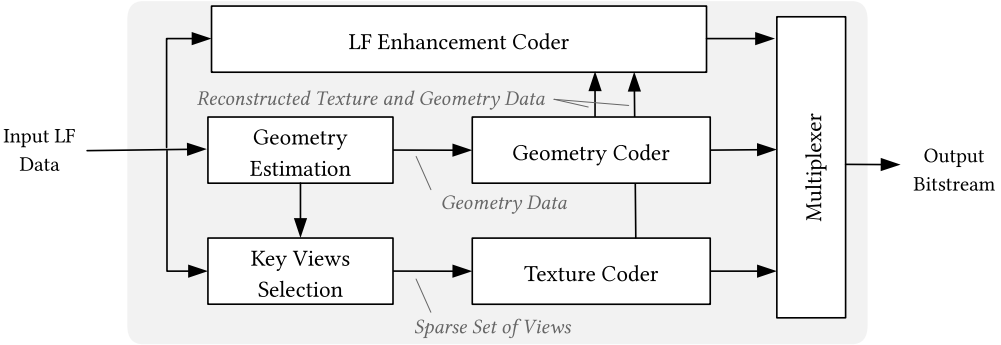


FIGURE 26. Alternative LF coding architecture based on view synthesis, including an LF enhancement coder at the highest enhancement layer.

the rendered view, 3D-HEVC standard is used as the texture coder, and the coding configuration (notably, the QP value) is selected so as to optimize the RD coding performance in the LF enhancement layer encoding process. Differently, disparity information is directly transmitted to the decoder side, bypassing the disparity coder block in Fig. 26. Afterwards, at the LF enhancement layer coder, the reconstructed view is low-pass filtered using an average filter [65]. This filtered view and disparity information are then used to build a reference picture, which is simply subtracted from the original lenslet image and encoded with HEVC intra coding. From the presented results, it is shown that further bit savings could be achieved (up to 31.1% of bit savings compared to HEVC Still Picture Profile [85]) by using the proposed approach when an optimized set of QP values are selected. In [245], the authors propose to extend their previous coding solution in [244] so as to consider more than one extracted view (notably, a set of 3, 5, and 9 views). These views are then encoded with 3D-HEVC using an IPP inter-view prediction structure (similar to Fig. 21c) and a set of optimized QP values. The proposed solution achieved up to 29.1% (when 3 views are extracted), 27.9% (when 5 views are extracted), and 27.2% (when 9 views are extracted) of bit savings compared to HEVC Still Picture Profile [85].

In [246], a solution for coding lenslet LF content is proposed based on DIBR for view synthesis. However, in this case, only a sparse set of four subaperture images on the extreme corner positions, as depicted in Fig. 25d are coded as a PVS using HEVC inter coding and transmitted. At the decoder side, disparity map of the four coded and reconstructed subaperture images are estimated using a deep learning algorithm for computing the optical flow. Afterwards, a DIBR technique is used in which a low rank matrix completion algorithm is used to fill the holes in texture and disparity map since the disoccluded areas of different warped views from only four corner views are unlikely to overlap [246]. Experimental results are shown using the *INRIA LF Dataset* (see Table 1) and comparing the proposed solution against: i) a view synthesis technique using deep learning in [97] (proposed for LF super-resolution); and ii) a PVS-based solution using lozenge scan (see Fig. 9h) of subaperture images coded with HEVC. It is shown that the proposed solution outperforms all the other solutions with significant RD gains compared to the PVS-based solution and slightly coding gains at low bit rates compared to the deep learning-based solution. In [247], a lenslet LF coding approach based on the 3D-HEVC standard is proposed. Using the representation format in Fig. 11, the subaperture images with full parallax are organized as a multiview sequence by interpreting the vertical directional dimension v as a (pseudo) temporal dimension. Hence, half of these multiview videos with the associated depth maps are uniformly selected to be encoded using 3D-HEVC. As illustrated in Fig. 25b, this corresponds to uniformly selecting subaperture images in sparse columns. An algorithm for estimating the depth maps (prior to coding) based on the EPIs is also proposed. At the decoder side, the remainder subaperture images are synthesized using the DIBR technique proposed in [238]. The proposed solution is compared against two PVS-based solutions: i) a solution using spiral scan and HEVC inter coding; and ii) the winner solution at the ICME 2016 Grand Challenge in [157]. From the results using the *EPFL LF Dataset* (see Table 1), it is shown that the proposed solution achieves in average 64.1% of bit savings compared to the HEVC-based solution, and presents significant gains at high bit rates and similar RD performance at low bit rates compared to the solution in [157].

In [126], a lenslet LF coding scheme is proposed using the coding architecture in Fig. 26. In the lower layers, a sparse set of subaperture images are uniformly selected (see Fig. 25a) and stitched together to be encoded with JPEG 2000. Additionally, depth map estimated using the algorithm proposed in [78] are quantized and coded using a context-based coder. These lower layers can be then used to synthesize the remaining subaperture images using the DIBR technique proposed. The DIBR technique uses the coded information from the key set of subaperture views to warp and merge texture and depth maps for the remaining views. In addition to this, in the LF enhancement layer (see Fig. 26), a sparse prediction scheme using a linear regression model is proposed to improve the quality of synthesized subaperture images. For this, the parameters used in the sparse prediction are quantized, coded using Golomb-Rice coding, and transmitted

to the decoder side. This solution was recognized as the most innovative algorithm at the ICIP 2017 grand challenge on LF compression [24]. In [158], [159], the authors propose to extend their previous solution in [126] for coding LF content acquired using a multi-camera array. For this, a coding scheme similar to the one used in [126] is adopted. However, in this case, JPEG 2000 is used as the depth map coder in lower layers. Additionally, in the LF enhancement layer, residual between original and predicted views are encoded with JPEG 2000 and transmitted to the decoder side, along with the sparse prediction parameters. Experimental results using the *Fraunhofer LF Dataset* (see Table 1) show that the proposed solution achieves significant coding gains compared to JPEG 2000 (as shown in [159]), and can outperform a PVS-based solution using HEVC at low bit rates (as shown in [158]). This solution, referred to as Warping and Sparse Prediction (WaSP) [158], has been recently adopted in the JPEG Pleno VM [127] (since version 1.0) for coding LF images acquired by multi-camera arrays.

More recently, in [248], the authors propose to improve their WaSP solution [158] by introducing a more efficient region-based sparse prediction scheme and by using HEVC. In this scheme, the synthesized views are segmented into regions using the disparity map and, for each region, an optimal sparse prediction is estimated. Differently from their previous WaSP solution in [158], [159], the sparse set of key views and the residual of synthesized views are coded using HEVC with serpentine scan-order (see Fig. 9b). Additionally, a hierarchical coding scheme is adopted, in which the views are divided into two or more disjoint subsets representing different hierarchical layers. Thus, the views on a lower layer are used as possible reference views to encode the views in the higher layers. Therefore, as the encoder proceeds towards the higher hierarchical layers, the density of the reference views increases and the intermediate view prediction becomes more efficient [248]. Experimental results are shown for coding LF images from the *EPFL*, *HCI 4D* and *Fraunhofer LF Datasets* (see Table 1) and comparing the proposed solution to eight different LF coding solutions: i) the WaSP solution [158] as in JPEG Pleno VM 2.1 [127]; ii) the DCT-based solution MuLE [125] as in the JPEG Pleno VM 2.1 [127]; iii) the lifting DTW-based solution in [148]; iv) the GFT-based solution in [155]; v) a PVS-based solution using serpentine ordering (see Fig. 9b) and HEVC; vi) the solution in [249] using transform-assisted view synthesis (see Section III-D2); vii) the solution in [250] using learning-based synthesis (see Section III-D3); and viii) the solution in [49] also using learning-based synthesis. For coding lenslet LF images, the proposed solution is seen to significantly outperform the WaSP [158], MuLE [125] and the GFT-based solution [155], while it is outperformed at low bit rates by the solutions using learning-based synthesis in [49], [250]. It is seen that, in average, the proposed solution achieves bit savings of 48.1% compared to WaSP [158] and 27.8% compared to MuLE [125]. For coding LF images acquired using a multi-camera array, the proposed solution achieves 28% of bit savings compared to WaSP [158] and slightly better performance, mainly at low bit rates, than the lifting DWT-based solution [148] and the solution in [249] using transform-assisted synthesis.

2) TRANSFORM-ASSISTED VIEW SYNTHESIS

Transform-assisted view synthesis is built upon the assumption that it is possible to compute a frequency representation of a sparse signal using only a subset of samples. Then, assuming the LF is k-sparse in the angular frequency domain, it can be represented as a linear combination of k non-zero continuous angular frequency coefficients [96]. The view synthesis algorithm then searches for the frequency values and the corresponding coefficients to reconstruct the discarded samples. This approach can be seen as an alternative to DIBR-based solutions for synthesis of non-Lambertian scenes since depth/disparity estimation is shown to fail on such scenes that contain, for instance, refractive and/or mirror-like objects [96], [98].

In [251], a coding scheme is proposed for LF acquired using a multi-camera array in a parallel arrangement. In this scheme, only a sparse set of views are uniformly selected (see Fig. 25a) and converted into a PVS using serpentine scan to be compressed using HEVC. At the decoder side, a view synthesis technique based on a shearlet transform, previously proposed by the authors in [98], is adopted. For this, the coded key views are organized as EPIs (see Fig. 10) and a shearlet transform is used to recover the line slopes in the sparse EPIs. The algorithm is applied interactively for recovering the EPIs taken from horizontal and vertical directions. Experimental results are shown using the *Stanford LF Archive* and *Fraunhofer LF Dataset* (see Table 1) and the proposed solution is compared to a PVS-based solution using HEVC. It is shown that the proposed solution achieves significant coding gains only at low bit rates, but it underperforms the PVS-based HEVC solution at high bit rates. In [249], a LF coding solution based on Fourier Disparity Layer (FDL) decomposition [252] and the coding architecture shown in Fig. 26 is proposed. In this case, the LF content is divided into various sparse sets of views that are hierarchically encoded in different layers. In the base layer, a set of key views are arranged as a PVS in spiral order (see Fig. 9d) and directly encoded using HEVC inter coding. Afterwards, one or more LF enhancement layers are defined, each of which comprising a set of the remaining views to be encoded. The previously coded views in the lower layers are then used for synthesizing the remaining views in an LF enhancement layer by using FDL. As previously proposed in [252], the FDL decomposes, in the Fourier domain, the sparse set of views into a discrete sum of depth layers by solving a linear regression problem. The set of parameters from this FDL decomposition gathers the geometry information and is then transmitted as metadata to the decoder side. Additionally, the residual between synthesized and original views is encoded in a LF enhancement layer as a PVS with spiral ordering using HEVC inter coding. Experimental results are shown using three different

datasets (*HCI 4D*, *INRIA*, and *EPFL LF Datasets*, as seen in Table 1) and comparing the proposed solution against: i) a PVS-based solution with spiral ordering (see Fig. 9d) and HEVC inter coding; and ii) the DIBR-based WaSP solution in [158], [159]. For this comparison, uniform (see Fig. 25a) and circular (see Fig. 25g) sampling patterns are considered for selecting the key views in the proposed solution. From the presented results comparing against the WaSP solution [158], [159], it is seen that the proposed solution can achieve, in average, 56.5% of bit savings while the PVS-based solution achieves 35.42% in average.

In [80], a LF coding scheme is proposed based on a texture plus graph representation (see Fig. 13) and the coding architecture shown in Fig. 26. In the lower layers, only two key views from the extreme left bottom and right top corner positions (see Fig. 25e) are encoded and transmitted to the decoder side. To describe the 3D scene geometry, a graph-based representation is proposed by defining each pixel position in the subaperture image as a node, and the color intensity in each pixel as the graph signal. The adjacency matrix $\mathbf{A}$ is derived from the disparity between the two key views and, based on the concept of epipolar segment [81], it is sparsified in an RDO manner. The key view at left bottom corner and the sparse adjacency matrix $\mathbf{A}$ are coded with HEVC Still Picture Profile. Differently, a diffusion-based prediction of the key view at right top corner is computed and the residual is coded using a GFT. At the decoder side, the coded key views and the sparse adjacency matrix $\mathbf{A}$ are used to synthesize the remaining views. Additionally, in the LF enhancement layer, further views are sparsely selected and the residual between original and predicted views are coded with HEVC and transmitted to improve the quality of the synthesized views in disoccluded areas. Experiments using synthetic LF content from the *HCI 4D LF Dataset* (see Table 1) show that, compared to a PVS-based solution using HEVC, the proposed solution can yield some RD gains only at low bit rates, while it is outperformed by the PVS-based HEVC solution at high bit rates.

In [253], a texture plus graph representation is also proposed for lenslet LF coding. In this case, a sparse set of subaperture images are uniformly selected and encoded as a PVS using HEVC with Low Delay B configuration (see Fig. 21e). Under the assumption that there are smooth transitions between pixels intensities in different subaperture images, a single graph is constructed to model the disparity between the views. The graph learning algorithm proposed in [79] is then used to build a sparse graph, which is then lossless encoded using GFT. At the decoder side, the sparse set of key views and the graph is used to synthesize the remainder subaperture views by solving an optimization problem that enforces smoothness. Experimental results are shown using the *EPFL LF Dataset* (see Table 1) and comparing the proposed solution to: i) a PVS-based solution using serpentine ordering (see Fig. 9b) and HEVC; ii) the DIBR-based WaSP solution in [158], [159]; and iii) a LF coding solution using learning-based view synthesis in [207] (see Section III-D3). It is shown that the proposed solution achieves, in average, 46.7% of bit savings against PVS-based HEVC, 53% against the WaSP solution [158], [159], and 43.0% against the learning-based view synthesis in [207].

3) LEARNING-BASED VIEW SYNTHESIS

The LF coding solutions in this category can be divided into approaches that use conventional machine learning techniques, such as sparse coding and dictionary learning, and approaches that use deep learning techniques, by notably, making use of a Convolutional Neural Network (CNN).

The first set of approaches are inspired by the problem referred to as robust PCA that consists in decomposing a matrix as the sum of a low rank matrix and a sparse matrix representing the noise. Then, the problem is reduced to finding an optimal low rank approximation model of the data that minimizes the noise. These LF coding solutions are then built upon the assumption that if one warps the views of the various viewpoint positions to a common warping center (i.e., to align the views) and considers each warped view as one column in a matrix, then this matrix will have low rank. Usually, a local linearity assumption is adopted to solve this low rank approximation problem — referred to as linear approximation.

In [207], a lenslet LF coding solution is proposed, based on a linear approximation prior and using the coding architecture in Fig. 16. In this case, a sparse set of key views are uniformly selected (see Fig. 25a), organized as a PVS in zig-zag order (see Fig. 9e) and encoded with HEVC. To collect geometry information, a linear approximation model of the discarded views is proposed using the coded key views. The coefficients of this linear approximation are then coded as an image using JPEG standard. At the decoder side, the discarded views are approximated as the weighted sum of the selected views. Experimental results using the *EPFL LF Dataset* (see Table 1) shows that the proposed solution can achieve 37.4% of bit savings, in average, compared to a PVS-based solution with serpentine scan (see Fig. 9b) of subaperture images that are coded with HEVC. This solution was recognized as the winner of the ICIP 2017 grand challenge on LF compression [24] as the best performing algorithm. In [246], a similar scheme is proposed based on linear approximation for coding lenslet LF content. In this case, homography projections are searched for each subaperture image to obtain the best low rank matrix approximation for a given target rank k (where k is less than the total number of subaperture views). In the cases where the scene contains several layers of depth, the method is also extended to search for one homography for each depth plane. The obtained rank k matrix is expressed as a product of a matrix $\mathbf{B}$, containing k basis vectors, with a matrix $\mathbf{C}$ containing weighting coefficients. The matrix $\mathbf{B}$ is then reshaped representing a sparse set of k images which are compressed with HEVC Still Picture Profile. The matrix $\mathbf{C}$ and the homography parameters representing the geometry data are lossless compressed using entropy coding. In case where multiple homographs are applied, one depth map is

also encoded with HEVC Still Picture Profile to be transmitted to the decoder side. At the decoder side, the subaperture images are reconstructed by recovering the low rank matrix and applying inverse warping. Results are shown comparing to a PVS-based solution using lozenge scan (see Fig. 9h) and HEVC, for images from the *EPFL*, *INRIA*, and *HCI 4D LF Datasets* (see Table 1). The proposed solution is shown to achieve 2.2 dB of gains in average. Similarly, in [250], a lenslet LF coding scheme is proposed based on a linear approximation prior. In this case, a sparse set of uniformly selected key subaperture images (see Fig. 25a) arranged as a PVS in serpentine order (see Fig. 9b) is encoded with HEVC. Then, a translation estimation is proposed to search for the best linear approximation model. The estimation is performed in a block-based manner by segmenting the key subaperture images into several depth planes. To limit the encoder complexity, a quadtree segmentation is employed to compute the segmented blocks. The weighting coefficients of the linear approximation are then represented as 16-bit floating point numbers and transmitted. At the decoder side, the discarded subaperture images are reconstructed as the weighted sum of the key subaperture images. In an LF enhancement layer (see Fig. 26), the residual between original and synthesized subaperture images is decomposed using KLT and only the first k coefficients are transmitted to the decoder side. Experimental results are shown using the *EPFL LF Dataset* (see Table 1) and comparing the performance of the proposed solution against: i) a PVS-based solution using HEVC with serpentine scan order (see Fig. 9b); ii) the MuLE solution [125] as adopted in JPEG Pleno VM 2.1 [127]; iii) the WaSP solution [158], [159] as in JPEG Pleno VM 2.1 [127]; and iv) the solution in [253] using graph learning for synthesis. From these results, the proposed solution is seen to achieve bit savings of, in average, 58.5% against PVS-based HEVC, 49.4% against MuLE [125]; 54.8% against WaSP [158], [159], and 24.8% against the graph learning-based solution [253].

In [237], a lenslet LF coding scheme is proposed using the coding architecture in Fig. 16 and a dictionary-based learning approach for view synthesis. In this case, the subaperture images are divided into overlapping coding regions with dimension $d \times d$, and a sparse set of key subaperture views are selected in the overlapping area as seen in Fig. 25f. For each region, a disparity map is estimated using a maximum a posteriori estimator. Both the set of key views and the disparity map are organized as a PVS and coded with JEM using Low Delay P configuration (see Fig. 21d). At the decoder side, the coded texture and depth data in each coding region are used to learn a disparity-guided dictionary. Then, the discarded subaperture views in this coding region are reconstructed as an optimized combination of samples in the constructed dictionary. In addition, in an LF enhancement layer (see Fig. 26), the residual between original and synthesized views are coded with JEM as a PVS using the prediction structure proposed in [157] (see Fig. 21g). From experimental results using the *EPFL LF Dataset* (see Table 1), it is shown that the proposed solution can achieve 37.9% of bit savings compared to a PVS-based solutions using serpentine order (see Fig. 9b) and JEM, and 16.4% compared to the winner of the ICME 2016 grand challenge in [157].

Inspired by the recent success of deep learning in a variety of applications, many LF coding solutions propose to use deep learning for view synthesis. Among the possible different techniques, CNNs are deep learning techniques particularly powerful for image analysis problems, such as recognition and classification. Essentially, the input image is fed into the CNN and then processed through a series of hidden layers before revealing the solution to the problem. Firstly, in a convolution layer, a filter is applied to extract low-level features, such as edges, from the input image. Then, a pooling layer is used to reduce the dimensionality of the convolved features to decrease the computational power required to process the data. Finally, the data is vectorized and fed to a conventional feed-forward neural network.

In [254], a lenslet LF coding scheme using a CNN for view synthesis is proposed. In this case, a sparse set of four subaperture images on the extreme corner positions (see Fig. 25d) are coded as a PVS using HEVC inter coding and transmitted. At the decoder side, a CNN-based view synthesis algorithm proposed in [97] for LF super-resolution is used to synthesize the discarded subaperture images. In this case, two sequential CNNs are used for estimating the disparity and, then, for reconstructing the discarded subaperture images. In an LF enhancement layer (see Fig. 26), the residual between original and synthesized subaperture images are organized as a PVS with raster-scan ordering (see Fig. 9a) and encoded with HEVC inter coding. An RD optimization model is also proposed to select the optimal QPs for coding key and residual of discarded subaperture images. Experimental results using the *EPFL LF Dataset* (see Table 1) show that the proposed solution can achieve 13.2% of bit savings (in average) compared to a PVS-based scheme using HEVC with Random Access configuration (see Fig. 21f). In [255], a similar lenslet LF coding framework is proposed using the same four key subaperture images to be coded (see Fig. 25d) and the same CNN architecture [97] for synthesizing the remaining subaperture images at the decoder side. However, in the LF enhancement layer (see Fig. 26), the residual of discarded subaperture images is coded using GFT. For this, a super-pixel segmentation proposed in [256] is used to subdivide a reference subaperture image into uniform regions where the residual signal is supposed to be smooth. These super-pixel regions are then considered co-located to all subaperture images. Then, a separable GFT is applied to each super-pixel, comprising a local spatial GFT and an inter-view GFT. GFT coefficients are then coded using an arithmetic coder. For the inter-view GFT, the Laplacian matrix is learned from a training set and fixed. Experimental results compare the proposed solution against: i) a PVS-based coding using lozenge scan (see Fig. 9h) and HEVC; and ii) a CNN-based solution similar to [254]. It is shown that the proposed solution slightly outperforms the CNN-based solution and achieves 1.2 dB (in average) compared to the PVS-based solution.

In [257], the authors propose to replace the GFT used in their previous work in [255] by a 4D shape adaptive DCT for coding the residual of discarded subaperture images. For this, the super-ray segmentation proposed in [156] is applied to the residual of discarded subaperture images and used as the support for two separable 2D shape adaptive DCT for exploiting spatial and angular correlations. To further improve the coding performance, a super-ray merging is proposed in an RD optimization manner. Experimental results, comparing to the PVS-based solution using lozenge scan (see Fig. 9h) and HEVC, show that the proposed solution can achieve coding gains of 1.4 dB (in average) at low bit rates, but it is outperformed at high bit rates.

In [258], the authors propose a lenslet LF coding solution where they replace their linear approximation model proposed in [207] by two CNN networks for view synthesis. In this case, only a sparse set of uniformly sampled subaperture images (see Fig. 25a) are coded as a PVS using HEVC. At the decoder side, the first network aims at reducing compression artifacts in the coded key subaperture views, while the second network is used to recover the discarded subaperture images. Experimental results are shown using the *EPFL LF Dataset* (see Table 1) and comparing the proposed solution against: i) a PVS-based solution using serpentine scan of subaperture images (see Fig. 9b) and HEVC; and ii) their previous solution in [207] using linear approximation. The proposed solution is shown to outperform the other solutions with up to 0.78 dB against the PVS-based solution and 0.36 dB against the linear approximation-based solution in [207].

In [259], a LF coding scheme is proposed using a Generative Adversarial Network (GAN) for view synthesis. In this case, a sparse set of uniformly sampled subaperture images (see Fig. 25a) are organized as a PVS and coded with HEVC using the coding configuration proposed in [178] (see Fig. 21h). At the decoder side, the discarded subaperture views are synthesized using three CNNs. The first network uses the coded key subaperture images to generate a high order approximation of the discarded subaperture views. Afterwards, a second network is used to refine the quality of this high order approximation. To ensure sharp edges as well as detailed textures, a discriminative network is applied using adversarial learning. In addition, in an LF enhancement layer (see Fig. 26), the residual of discarded subaperture images is organized as a PVS and coded with HEVC. An RD optimization is also proposed for optimal bit allocation between key and discarded subaperture images. Experimental results are shown for coding LF images from the *EPFL LF*, *Stanford LF Archive*, and *HCI 4D LF Datasets* (see Table 1). The proposed solution is then compared to: i) a PVS-based solution with HEVC using Random Access configuration (see Fig. 21f); ii) the PVS-based solution winner of the ICME 2016 grand challenge in [157]; and ii) the CNN-based solution in [254]. From these results, the proposed solution is shown to achieve 14.8% bit savings compared to PVS-based HEVC, 8.1% compare to the PVS-based solution winner of ICME 2016 grand challenge in [157], and 4.9% compared to the CNN-based solution in [254]. In [260], a lenslet LF coding scheme is proposed using a hybrid scheme for view synthesis. In this case, two set of key subaperture images are uniformly selected from the LF data. The first set is organized as a PVS in zig-zag order (see Fig. 9e) and coded with JEM inter coding. From the second set, a linear approximation model is estimated by spectral projected gradient method and the resulting coefficients are entropy coded. At the decoder side, the key subaperture views of the second set are reconstructed as a linear combination of the key subaperture views in the first set. In addition, two sequential CNNs are used to estimate a disparity map and then to synthesize the remaining subaperture images. The proposed solution is compared against: i) a PVS-based solution using HEVC inter coding; and ii) the solution using only linear approximation in [207]. Results using the *EPFL LF Dataset* (see Table 1) show that the proposed solution achieves, in average, 51.1% bit savings against the PVS-based, and 30.8% against the linear approximation-based [207] solutions.

4) DISCUSSION

LF coding approaches using view synthesis for achieving compression have recently increased in the literature. Generically, the coding performance of these solutions is closely related to the selection of the key views, the performance of the synthesis algorithm, and the accuracy of the geometry information estimated from the acquired LF data. However, there have not yet been in-depth studies (to the best of authors' knowledge) directly analyzing the influence of the key view selection and geometry estimation algorithms on the compression performance.

An advantage of using the LF coding approaches in this category is that scalability, random access and backward compatibility are straightforwardly supported using a 3D coding standard in the lower layers. Quality scalability is also supported by using only information from the lower layers in reconstructing the views from the higher layers.

Among the different view synthesis techniques proposed, learning-based techniques, mainly using a low-rank approximation alone (as in [250]) or combined with CNN-based techniques (as in [260]), have shown to achieve the best coding performance with significant gains when compared to DIBR-based [158], [159] and transform-assisted [253] synthesis. Nevertheless, the DIBR-based solution proposed in [248] has shown competitive RD performance at high bit rates for coding lenslet LFs by making use of a sparse prediction scheme for coding residuals in LF enhancement layers. Additionally, transform-assisted techniques for view synthesis, such as the GFT-based technique in [253] and the FDL-based technique in [249] have shown competitive performance compared to some DIBR-based techniques [158], [159].

With respect to other LF coding solutions in the literature (outside the view synthesis category), view synthesis-based solutions have generally shown better RD performance, mainly at low bit rates, compared to PVS-based LF coding

TABLE 2. List of the most relevant dense LF coding solutions in each category (in chronological order).

Category	Sub-category	Papers
Transform coding	DCT-based	[118] [120] [119] [123] [121] [122] [124] [125]
	KLT-based	[131] [133] [134]
	DWT-based	[139] [145] [143] [135] [136] [140] [141] [147] [148]
	GFT-based	[153] [151] [155]
	Combined approaches	[161] [160] [162] [163]
Inter-view prediction	PVS-based	[166] [139] [167] [168] [170] [181] [182] [172] [174] [173] [175] [179] [180] [183] [185] [191] [193] [157] [186] [188] [192] [178] [187]
	Multiview-based	[197] [202] [196] [203] [198] [205] [204] [199] [20] [200] [201]
Non-local spatial prediction	Spatial compensated	[210] [211] [233] [212] [206] [217] [218] [222] [225] [226] [219] [221] [224] [227] [221]
	Learning-based	[223] [232] [234] [220] [235] [236]
View synthesis	DIBR-based	[239] [240] [241] [242] [244] [243] [245] [246] [126] [247] [158] [159] [248]
	Transform-assisted	[80] [251] [253] [249]
	Learning-based	[207] [246] [237] [254] [255] [257] [258] [259] [260] [250]

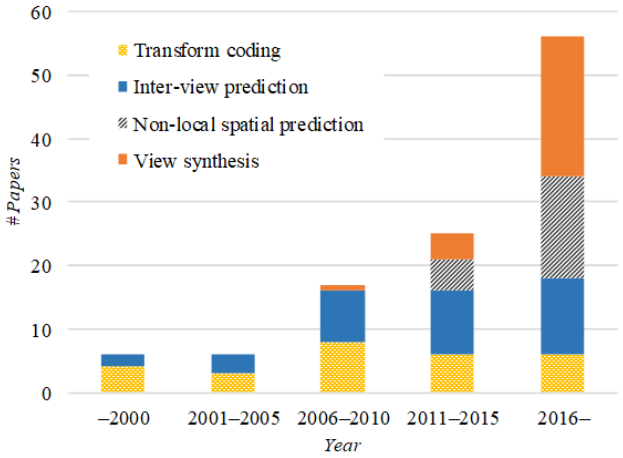


FIGURE 27. Time evolution of the number of papers on dense LF coding in each category.

solutions. Moreover, the most recent DIBR-based [248] and learning-based solutions [250], [260] have shown to significantly outperform the MuLE solution based on 4D DCT [125] for coding all kinds of LF content, including lenslet LFs.

IV. FINAL REMARKS

This paper provided a comprehensive survey of the most relevant lossy coding solutions for dense LF content proposed in the literature in the last 25 years. A detailed analysis of LF coding solutions was provided as well as a careful categorization of the representation format, coding architecture and techniques considered for achieving compression. A high-level discussion was also presented to highlight the most relevant results, advantages, and limitations of the LF coding solutions in each category.

As a summary, Table 2 presents a chronological list of all reviewed papers in each coding category, while Fig. 27 shows the distribution of these papers along time. From this, it is possible to observe that the number of LF coding solutions based on transform coding and inter-view prediction almost exclusively dominated the research efforts until 2010. Moreover, there has been a marked increase in the number of LF coding solutions based on view synthesis (since 2006) and based on non-local spatial prediction (since 2011).

Given the richness of methods proposed in the literature, it is unfeasible to make a single recommendation of what would be the winning solution for LF coding. Moreover, when determining the effectiveness of a solution, there are typically several other requirements, apart from RD compression performance, that need to be considered. These include, for instance, scalability, complexity, random access, and ROI capabilities. Depending on a particular application scenario, some of these requirements might be more important than others. However, analyzing the recent LF coding solutions that present the most promising results, the following characteristics stand out:

- **Use of View Synthesis** — Most of the recent and promising LF coding solutions, in terms of RD performance, make use of view synthesis for achieving efficient compression. This approach has been also under consideration in most JPEG Pleno [108] and MPEG-I [59] core experiments. An advantage of such solutions is that scalability, random access and backward compatibility are easily supported. As geometry information is used for synthesizing views at the decoder side, the amount of views that needs to be encoded and transmitted is greatly reduced. Nevertheless, encoding the residual in an LF enhancement layer by using the synthesized views as a prediction has been shown to be advantageous to improve the perceived quality of synthesized views. It is worth noting that, although many different schemes for selecting the key views have been proposed, there have not yet been done in-depth studies directly analyzing the influence of the key view selection on the coding performance. Moreover, accurate geometry estimation and representation still need to be further studied since they have a strong impact on the coding performance.

- **Use of Image-Based Learning Techniques** — Mainly due to the nature of LFs, represented as a large collection of viewpoint images, image-based learning techniques have shown to be hugely advantageous for prediction and view synthesis. Among the possible techniques, low-rank approximation, as in [250], has demonstrated to be a powerful tool to gather geometry information for view synthesis. In addition, deep learning techniques, mainly based on CNNs as in [260], have presented promising results for depth/disparity estimation and view synthesis. Moreover, learning-based transform decomposition schemes, as in [24], [249], have shown to be advantageous for view synthesis. Regarding prediction, learning-based prediction schemes, such as sparse and dictionary-based prediction, have also shown promising results either when applied in an image basis (to predict the entire view, as in [248]) or in a coding block basis (to predict a coding block, as in [236]).
- **Use of Highly Efficient Hybrid Video Codecs** — Using hybrid video codecs, such as HEVC and VVC, for coding texture and residual information has generally resulted in better RD performance than using classical image coding solutions, such as JPEG and JPEG 2000. In fact, the majority of recent LF coding solutions in the literature has adopted this framework due to its effectiveness for providing high efficiency compression. Nevertheless, it has been seen that 4D transform-based LF coding solutions — such as the 4D DCT in MuLE [125], the 4D shape adaptive DCT in [257], and the 4D lifting-based DWT in [148] — can be promising alternatives, being able to achieve competitive RD performance using a classical image coding architecture.

Based on the recent activities in JPEG and MPEG, the next years will certainly continue to be very productive for the research in LF coding. With respect to JPEG, the Draft International Standards for both JPEG Pleno Part 1 (Framework) and Part 2 (LF coding) have been recently completed [261]. For LF coding, the two solutions MuLE [125] and WaSP [158], [159] are supported as coding modes, respectively, for coding LFs acquired with lenslet camera and multi-camera arrays. Moreover, JPEG has recently created an ad hoc group for studying promising learning-based 2D image codecs, mainly using deep neural network models [261]. In this context, it has been shown that learning-based solutions can achieve competitive objective and subjective qualities when compared to 2D image and video coding standards [261]. Following this trend, it is safe to expect that many new LF coding solutions in this category will also be proposed in the near future. With respect to MPEG-I video, a Multiview plus Depth (MVD) coding solution [262] has been recently under exploration for 3 DoF and 3 DoF with some limited motion parallax (3 DoF+), in which a DIBR-based view synthesis technique is used. It is expected that, after its final standardization in mid–end of 2020, a Call for Proposals will be issued for long-term 6 DoF activities, including LF video coding standardization [262].

REFERENCES

[1] J. Wang, X. Xiao, H. Hua, and B. Javidi, "Augmented reality 3D displays with micro integral imaging," *J. Display Technol.*, vol. 11, no. 11, pp. 889–893, Nov. 2015.
[2] LightField Forum. (2020). *Lytro Archive*. [Online]. Available: http://lightfield-forum.com/lytro/lytro-archive/
[3] J. Arai, M. Kawakita, T. Yamashita, H. Sasaki, M. Miura, H. Hiura, M. Okui, and F. Okano, "Integral three-dimensional television with video system using pixel-offset method," *Opt. Express*, vol. 21, no. 3, pp. 3474–3485, Feb. 2013.
[4] Holografika Kft. (2019). *Holovizio Display System*. [Online]. Available: http://www.holografika.com/
[5] R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1060–1075, Mar. 2015.
[6] D. Shin, M. Cho, and B. Javidi, "Three-dimensional optical microscopy using axially distributed image sensing," *Opt. Lett.*, vol. 35, no. 21, p. 3646, Nov. 2010.
[7] Raytrix. (2012). *Raytrix Website*. [Online]. Available: http://www.raytrix.de/
[8] H. S. El-Ghoroury, C.-L. Chuang, and Z. Y. Alpaslan, "Quantum photonic imager (QPI): A novel display technology that enables more than 3D applications," in *SID Symp. Dig. Tech. Papers*, Jun. 2015, vol. 46, no. 1, pp. 371–374.
[9] T. Georgiev, Z. Yu, A. Lumsdaine, and S. Goma, "Lytro camera technology: Theory, algorithms, performance analysis," *Proc. SPIE*, vol. 8667, Mar. 2013, Art. no. 86671J.
[10] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, "JPEG Pleno: Toward an efficient representation of visual reality," *IEEE Multimedia Mag.*, vol. 23, no. 4, pp. 14–20, Oct. 2016.
[11] A. P. Pozo, M. Toksvig, T. F. Schrager, J. Hsu, U. Mathur, A. Sorkine-Hornung, R. Szeliski, and B. Cabral, "An integrated 6DoF video camera and system design," *ACM Trans. Graph.*, vol. 38, no. 6, p. 216, Nov. 2019.
[12] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–15, Dec. 2018.
[13] J. Yu, "A light-field journey to virtual reality," *IEEE Multimedia Mag.*, vol. 24, no. 2, pp. 104–112, Apr. 2017.
[14] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, Aug. 2006.
[15] M. Harris, "Focusing on everything," *IEEE Spectr.*, vol. 49, no. 5, pp. 44–50, May 2012.
[16] M. Hiramoto, Y. Ishii, and Y. Monobe. (Mar. 2014). *Light Field Image Capture Device and Image Sensor*. [Online]. Available: https://patents.google.com/patent/US20140078259
[17] R. Wilson, B. Olofsson, V. Craciun, D. Copley, N. Parikh, J. Shepherd, X. Li, J. Budney, C. Michaels, S. Saroha, and B. Joshi, "The state of VR/360-degree video," Streaming Video Alliance, Fremont, CA, USA, Working Group no. VR/360-degree Video Study Group, Sep. 2018.
[18] E. Krokos, C. Plaisant, and A. Varshney, "Virtual memory palaces: Immersion aids recall," *Virtual Reality*, vol. 23, no. 1, pp. 1–15, Mar. 2019.
[19] D. Freeman, P. Haselton, J. Freeman, B. Spanlang, S. Kishore, E. Albery, M. Denne, P. Brown, M. Slater, and A. Nickless, "Automated psychological therapy using immersive virtual reality for treatment of fear of heights: A single-blind, parallel-group, randomised controlled trial," *Lancet Psychiatry*, vol. 5, no. 8, pp. 625–632, Aug. 2018.
[20] C. Conti, "Efficient solutions for light field coding," Ph.D. dissertation, Dept. Inf. Sci. Technol., ISCTE-Inst. Univ. de Lisboa, Lisboa, Portugal, Jun. 2017.
[21] P. Schelkens, "JPEG PLENO–scope, use cases and requirements ver1.6," ISO/IEC JTC1/SC29/WG1, Chengdu, China, Tech. Rep. ISO/IECJTC1/SC29/WG1N73030, Oct. 2016.
[22] ISO/IEC JTC1/SC29/WG11, "MPEG-I Visual activities on 6DoF and light fields," ISO/IEC JTC1/SC29/WG11, Macao, China, Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2017/N17285, Oct. 2017.
[23] M. Řeřábek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, "ICME 2016 grand challenge: Light field image compression—Call for proposals and evaluation procedure," Seattle, WA, USA, Tech. Rep., 2016.

[24] I. Viola and T. Ebrahimi, "Quality assessment of compression solutions for ICIP 2017 Grand Challenge on light field image coding," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Athens, Greece, Jul. 2018, pp. 1–6.

[25] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: Briefly revisiting 25 years of research," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 59–69, Sep. 2016.

[26] C. Guillemot and R. A. Farrugia, "Light field image processing: Overview and research issues," *MMTC Commun.-Frontiers*, vol. 12, no. 4, pp. 37–43, 2017.

[27] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[28] A. Gershun, "The light field," *J. Math. Phys.*, vol. 18, nos. 1–4, pp. 51–151, 1939.

[29] L. Da Vinci, *The Notebooks of Leonardo Da Vinci*, J. P. Richter, Ed. New York, NY, USA: Oxford Univ. Press, 1988.

[30] G. Lippmann, "Épreuves réversibles donnant la sensation du relief," *J. de Phys. Théorique et Appliquée*, vol. 7, no. 1, pp. 821–825, Jan. 1908.

[31] A. Bogusz, "Holoscopy and holoscopic principles," *J. Opt.*, vol. 20, no. 6, pp. 281–284, Nov. 1989.

[32] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. Cambrigde, MA, USA: MIT Press, 1991, pp. 3–20.

[33] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, New Orleans, LA, USA, 1996, pp. 43–54.

[34] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, New York, NY, USA, Oct. 1996, pp. 31–42.

[35] F. Pereira, E. A. da Silva, and G. Lafruit, "Plenoptic imaging: Representation and processing," in *Academic Press Library in Signal Processing*, vol. 6, R. Chellappa and S. Theodoridis, Eds. Amsterdam, The Netherlands: Elsevier, 2018, pp. 75–111.

[36] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *Proc. ACM SIGGRAPH*, vol. 24, no. 3. Los Angeles, CA, USA, Jul. 2005, p. 765.

[37] Light. (2019). *Light L16 Camera*. [Online]. Available: https://light.co/camera

[38] ISO/IEC JTC1/SC29/WG1 and ISO/IEC JTC1/SC29/WG11, "Technical report of the joint ad hoc group for digital representations of light/sound fields for immersive media applications," ISO/IEC JTC1/SC29/WG1 and ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. ISO/IECJTC1/SC29/WG1N72033, and ISO/IECJTC1/SC29/WG11N16352, 2016.

[39] S. G. Laboratory. *The (New) Stanford Light Field Archive*. Accessed: Jan. 10, 2020. [Online]. Available: http://lightfield.stanford.edu/index.html

[40] M. Ziegler, R. op het Veld, J. Keinert, and F. Zilly, "Acquisition system for dense lightfield of large scenes," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, Copenhagen, Denmark, Jun. 2017, pp. 1–4.

[41] R. Ng, "Digital light field photography," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2006.

[42] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photograp. (ICCP)*, San Francisco, CA, USA, Apr. 2009, pp. 1–8.

[43] A. Lumsdaine, T. G. Georgiev, and G. Chunev, "Spatial analysis of discrete plenoptic sampling," *Proc. SPIE*, vol. 8299, Jan. 2012, Art. no. 829909.

[44] S. Pertuz, E. Pulido-Herrera, and J.-K. Kamarainen, "Focus model for metric depth estimation in standard plenoptic cameras," *ISPRS J. Photogram. Remote Sens.*, vol. 144, pp. 38–47, Oct. 2018.

[45] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixe, and D. Cremers, "Deep depth from focus," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Dec. 2018, pp. 525–541.

[46] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, p. 1, Jul. 2013.

[47] M. Řeřábek, and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Lisbon, Portugal, 2016, pp. 1–2.

[48] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Annu. Workshop Vis., Modeling Visualizat. (VMV)*, M. B. Hormann, J. Favre, and K. Hormann, Eds. Aire-la-Ville, Switzerland: The Eurographics Association, 2013, pp. 225–226.

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

[50] N. Sabater, G. Boisson, B. Vandame, P. Kerbiriou, F. Babon, M. Hog, R. Gendrot, T. Langlois, O. Bureller, A. Schubert, and V. Allie, "Dataset and pipeline for multi-view light-field video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1743–1753.

[51] W. Ahmad, L. Palmieri, R. Koch, and M. Sjostrom, "Matching light field datasets from plenoptic cameras 1.0 and 2.0," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, Jun. 2018, pp. 1–4.

[52] K. Marwah, G. Wetzstein, A. Veeraraghavan, and R. Raskar, "Compressive light field photography," in *Proc. ACM SIGGRAPH*, New York, NY, USA, 2012, p. 1.

[53] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3720–3729.

[54] P. Paudyal, F. Battisti, and M. Carli, "Effect of visualization techniques on subjective quality of light field images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 196–200.

[55] D. G. Dansereau, B. Girod, and G. Wetzstein, "LiFF: Light field features in scale and depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1–10.

[56] S. Vagharshakyan, R. Bregovic, and O. Suominen. (2018). *Densely sampled light fields, Tampere University of Technology (TUT)*. [Online]. Available: http://urn.fi/urn:nbn:fi:att:ed60be6d-9d15-4857-aa0d-a30acd16001e

[57] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun./2004, pp. 294–301.

[58] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (ToF) cameras: A survey," *IEEE Sensors J.*, vol. 11, no. 9, pp. 1917–1926, Sep. 2011.

[59] M. Teratani, X. Jin, L. Li, G. Lafruit, and L. Yu, "Exploration experiments and common test conditions for dense light fields," ISO/IEC JTC1/SC29/WG11, Marrakesh, Morocco, Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2019/N18173, Jan. 2019.

[60] M. Teratani, X. Jin, G. Lafruit, and L. Yu, "Activity report on dense light fields," ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2019/N18791, Oct. 2019.

[61] R. Monteiro, P. Nunes, N. Rodrigues, and S. Faria, "Light field image coding: Objective performance assessment of Lenslet and 4D LF data representations," *Proc. SPIE*, vol. 10752, Sep. 2018, Art. no. 107520D.

[62] U. Fecker and A. Kaup, "H.264/AVC-compatible coding of dynamic light fields using transposed picture ordering," in *Proc. 13th Signal Process. Conf.*, Antalya, Turkey, Sep. 2005, pp. 1–4.

[63] D. G. Dansereau, O. Pizarro, S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1027–1034.

[64] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[65] A. Lumsdaine, "Focused plenoptic camera and rendering," *J. Electron. Imag.*, vol. 19, no. 2, Apr. 2010, Art. no. 021106.

[66] T. Georgiev, G. Chunev, and A. Lumsdaine, "Superresolution with the focused plenoptic camera," in *Proc. SPIE*, vol. 7873, Feb. 2011, Art. no. 78730X.

[67] S. Wanner, J. Fehr, and B. Jähne, "Generating EPI representations of 4D light fields with a single lens focused plenoptic camera," in *Proc. 7th Int. Symp. Vis. Comput.*, Las Vegas, NV, USA, Sep. 2011, pp. 90–101.

[68] M. Teratani Panahpourtehrani, S. Fujita, W. Ouyang, K. Takahashi, and T. Fujii, "3D imaging system using multi-focus plenoptic camera and tensor display," in *Proc. Int. Conf. 3D Immersion (IC3D)*, Brussels, Belgium, Dec. 2018, pp. 1–7.

[69] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7–55, Mar. 1987.
[70] I. Tosic and K. Berkner, "Light field scale-depth space transform for dense depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, Jun. 2014, pp. 441–448.
[71] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in shearlet domain," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Montreal, QC, Canada, Sep. 2015, pp. 1379–1383.
[72] C. Brites, J. Ascenso, and F. Pereira, "Epipolar plane image based rendering for 3D video coding," in *Proc. IEEE 17th Int. Workshop Multimedia Signal Process. (MMSP)*, Xiamen, China, Oct. 2015, pp. 1–6.
[73] T. E. Bishop and P. Favaro, "Plenoptic depth estimation from multiple aliased views," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops*, Kyoto, Japan, Sep./Oct. 2009, pp. 1622–1629.
[74] J. Chen and L.-P. Chau, "A fast adaptive guided filtering algorithm for light field depth interpolation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 2281–2284.
[75] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1547–1555.
[76] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2792–2799.
[77] O. Fleischmann and R. Koch, "Lens-based depth estimation for multi-focus plenoptic cameras," in *Pattern Recognition* (Lecture Notes in Computer Science), vol. 8753. Münster, Germany: Springer, 2014, pp. 410–420.
[78] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3487–3495.
[79] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, Cadiz, Spain, May 2016, pp. 920–929.
[80] X. Su, M. Rizkallah, T. Maugey, and C. Guillemot, "Graph-based light fields representation and coding using geometry information," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4023–4027.
[81] X. Su, T. Maugey, and C. Guillemot, "Rate-distortion optimized graph-based representation for multiview images with complex camera configurations," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2644–2655, Jun. 2017.
[82] *Information Technology–Digital Compression and Coding of Continuous-Tone Still Images–Requirements and Guidelines*. ITU-T Recommendation document T.81, 1992.
[83] *Information Technology-JPEG 2000 Image Coding System: Core Coding System*. ITU-T Recommendation document T.800, 2015.
[84] I. E. Richardson, *The H.264 Advanced Video Compression Standard*. Hoboken, NJ, USA: Wiley, 2010.
[85] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[86] B. Bross, J. Chen, and S. Liu, "Versatile video coding (draft 5)," ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. ITU-TSG16WP3 and ISO/IEC JTC1/SC29/WG11JVET-N1001, Mar. 2019.
[87] *Information Technology–Coding of Audio-Visual Objects—Part 2: Visual*. ISO/IEC document 14496-2:2004, 2004.
[88] G. Tech, Y. Chen, K. Muller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3D extensions of high efficiency video coding," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 26, no. 1, pp. 35–49, Jan. 2016.
[89] A. Isaksen, L. Mcmillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, New York, NY, USA, 2000, pp. 297–306.
[90] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Process. Mag.*, vol. 24, no. 6, pp. 10–21, Nov. 2007.
[91] G. Lafruit, J. Jung, D. Doyen, and P. Carballeira, "Next-generation light field video coding with view synthesis," ISO/IEC JTC1/SC29/WG11, Turin, Italy, Tech. Rep. ISO/IEC JTC1/SC29/WG11 MPEG2016/M40782, Jul. 2017.
[92] T. Georgiev and A. Lumsdaine, "Rich image capture with plenoptic cameras," in *Proc. IEEE Int. Conf. Comput. Photograp. (ICCP)*, Cambridge, MA, USA, Mar. 2010, pp. 1–8.
[93] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 1–20, Mar. 2015.
[94] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas, "Synthetic aperture confocal imaging," in *Proc. ACM SIGGRAPH Papers (SIGGRAPH)*, vol. 23, no. 3. New York, New York, 2004, p. 825.
[95] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2319–2330, May 2019.
[96] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 1–13, Dec. 2014.
[97] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–10, Nov. 2016.
[98] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 133–147, Jan. 2018.
[99] D. G. Dansereau, I. Mahon, O. Pizarro, and S. B. Williams, "Plenoptic flow: Closed-form visual odometry for light field cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Francisco, CA, USA, Sep. 2011, pp. 4455–4462.
[100] N. Zeller, F. Quint, and U. Stilla, "Depth estimation and camera calibration of a focused plenoptic camera for visual odometry," *ISPRS J. Photogram. Remote Sens.*, vol. 118, pp. 83–100, Aug. 2016.
[101] N. Zeller, F. Quint, and U. Stilla, "From the calibration of a light-field camera to direct plenoptic odometry," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1004–1019, Oct. 2017.
[102] F. Zilly, M. Schoberl, M. Ziegler, J. Keinert, and S. Foessel, "Light-field acquisition system that facilitates camera and Depth-of-Field compositing in post-production," *SMPTE Motion Imag. J.*, vol. 124, no. 1, pp. 16–21, Jan. 2015.
[103] Microsoft HoloLens. (2019). *HoloLens 2: A New Vision for Computing*. [Online]. Available: https://www.microsoft.com/en-us/hololens/
[104] D. Lanman and D. Luebke, "Near-eye light field displays," in *ACM SIGGRAPH Emerg. Technol.*, Anaheim, CA, USA, Jul. 2013, p. 1.
[105] Avegant. (2019). Avegant Light Field Technology and Video Headset. [Online]. Available: https://avegant.com/
[106] Looking Glass Factory. (2020). *Looking Glass 8K Immersive Display*. [Online]. Available: https://lookingglassfactory.com/product/8k
[107] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[108] Z. Alpaslan, F. Pereira, C. Pagliari, E. A. da Silva, I. Tabus, H. Amirpour, M. Bernardo, and A. Pinheiro, "JPEG Pleno light field coding common test conditions v3.2," ISO/IEC JTC1/SC29/WG1, Geneva, Switzerland, Tech. Rep. ISO/IECJTC1/SC29/WG1M83052, Mar. 2019.
[109] R. Olsson, "Synthesis, coding, and evaluation of 3D images based on integral imaging," Ph.D. dissertation, Dept. Inf. Technol. Media, Mid Sweden Univ., Sundsvall, Sweden, 2008.
[110] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on focus stack and extended focus images," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 898–902.
[111] N. Gehrig and P. L. Dragotti, "Geometry-driven distributed compression of the plenoptic function: Performance bounds and constructive algorithms," *IEEE Trans. Image Process.*, vol. 18, no. 3, pp. 457–470, Mar. 2009.
[112] R. R. Tamboli, B. Appina, S. Channappayya, and S. Jana, "Super-multiview content with high angular resolution: 3D quality assessment on horizontal-parallax lightfield display," *Signal Process., Image Commun.*, vol. 47, pp. 42–55, Sep. 2016.
[113] *Methodology for the Subjective Assessment of the Quality of Television Pictures*. ITU-R Recomendation document BT.500-13 (01/2012), 2012.
[114] I. Viola, M. Řeřábek, and T. Ebrahimi, "A new approach to subjectively assess quality of plenoptic content," *Proc. SPIE*, vol. 9971, Sep. 2016, Art. no. 99710X.
[115] P. Paudyal, F. Battisti, M. Sjostrom, R. Olsson, and M. Carli, "Towards the perceptual quality evaluation of compressed light field images," *IEEE Trans. Broadcast.*, vol. 63, no. 3, pp. 507–522, Sep. 2017.

[116] C. Perra, W. Song, and A. Liotta, "Effects of light field subsampling on the quality of experience in refocusing applications," in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May/Jun. 2018, pp. 1–3.

[117] P. A. Kara, A. Cserkaszky, A. Barst, T. Papp, M. G. Martini, and L. Bokor, "The interdependence of spatial and angular resolution in the quality of experience of light field visualization," in *Proc. Int. Conf. 3D Immersion (IC3D)*, Dec. 2017, pp. 1–8.

[118] M. C. Forman, A. Aggoun, and M. McCormick, "A novel coding scheme for full parallax 3D-TV pictures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4. Munich, Germany, Apr. 1997, pp. 2945–2947.

[119] M. C. Forman, "Compression of integral three-dimensional television pictures," Ph.D. dissertation, Dept. Comput. Sci. Eng., De Montfort Univ., Leicester, U.K., 2000.

[120] M. C. Forman, "Quantisation strategies for 3D-DCT-based compression of full parallax 3D images," in *Proc. 6th Int. Conf. Image Process. Appl.*, Dublin, Ireland, 1997, pp. 32–35.

[121] A. Aggoun, "A 3D DCT compression algorithm for omnidirectional integral images," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2. Toulouse, France, May 2006, pp. II-517–II-520.

[122] N. P. Sgouros, D. P. Chaikalis, P. G. Papageorgas, and M. S. Sangriotis, "Omnidirectional integral photography images compression using the 3D-DCT," in *Proc. Digit. Holograp. Three-Dimensional Imag.*, Vancouver, BC, Canada, Jun. 2007, pp. 1–3, Paper DTuA2.

[123] R. Zaharia, A. Aggoun, and M. McCormick, "Adaptive 3D-DCT compression algorithm for continuous parallax 3D integral imaging," *Signal Process., Image Commun.*, vol. 17, no. 3, pp. 231–242, Mar. 2002.

[124] A. Mehanna, A. Aggoun, O. Abdulfatah, M. R. Swash, and E. Tsekleves, "Adaptive 3D-DCT based compression algorithms for integral images," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, London, U.K., Jun. 2013, pp. 1–5.

[125] M. B. de Carvalho, M. P. Pereira, G. Alves, E. A. B. da Silva, C. L. Pagliari, F. Pereira, and V. Testoni, "A 4D DCT-based lenslet light field codec," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 435–439.

[126] I. Tabus, P. Helin, and P. Astola, "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4567–4571.

[127] ISO/IEC JTC1/SC29/WG1, "Verification model software version 2.1 on JPEG Pleno light field coding," ISO/IEC JTC1/SC29/WG1, Geneva, Switzerland, Tech. Rep. ISO/IEC JTC1/SC29/WG1N83034, 2019.

[128] K. Pearson, "On lines and planes of closest fit to systems of points in space," *J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[129] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.

[130] K. Sayood, *Introduction to Data Compression*. San Mateo, CA, USA: Morgan and Kaufmann, 2012.

[131] J.-S. Jang, "Compression of ray information in three-dimensional integral imaging," *Opt. Eng.*, vol. 44, no. 12, Dec. 2005, Art. no. 127001.

[132] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[133] H.-H. Kang, D.-H. Shin, and E.-S. Kim, "Compression of sub-image-transformed elemental images in integral imaging," in *Proc. Adapt. Opt., Anal. Methods/Comput. Opt. Sens. Imag./Inf. Photon./Signal Recovery Synth. Top. Meetings CD-ROM*, Vancouver, BC, Canada, 2007, pp. 1–10, Paper DTuA6.

[134] H.-H. Kang, D.-H. Shin, and E.-S. Kim, "Compression scheme of sub-images using karhunen-loeve transform in three-dimensional integral imaging," *Opt. Commun.*, vol. 281, no. 14, pp. 3640–3647, Jul. 2008.

[135] A. Aggoun, "Compression of 3D integral images using 3D wavelet transform," *J. Display Technol.*, vol. 7, no. 11, pp. 586–592, Nov. 2011.

[136] H. H. Zayed, S. E. Kishk, and H. M. Ahmed, "3D wavelets with SPIHT coding for integral imaging compression," *Int. J. Comput. Sci. Netw. Secur.*, vol. 12, no. 1, pp. 126–133, Jan. 2012.

[137] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.

[138] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Process.*, vol. 9, no. 7, pp. 1158–1170, Jul. 2000.

[139] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[140] R. S. Higa, R. F. L. Chavez, R. B. Leite, R. Arthur, and Y. Iano, "Plenoptic image compression comparison between JPEG, JPEG2000 and SPITH," *Cyber J., Multidisciplinary J. Sci. Technol., J. Sel. Areas Telecommun.*, vol. 3, no. 6, pp. 1–6, Jun. 2013.

[141] C. Perra, "On the coding of plenoptic raw images," in *Proc. 22nd Telecommun. Forum Telfor (TELFOR)*, Nov. 2014, pp. 850–853.

[142] F. Dufaux, G. J. Sullivan, and T. Ebrahimi, "The JPEG XR image coding standard [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 195–204, Nov. 2009.

[143] R. Olsson, "Empirical rate-distortion analysis of JPEG 2000 3D and H. 264/AVC coded integral imaging based 3D-images," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video*, Istanbul, Turkey, May 2008, pp. 113–116.

[144] *Information Technology-JPEG 2000 Image Coding System: Extensions for Three-Dimensional Data*. ITU-T Recommendation document T.809, May 2011.

[145] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.

[146] W. Sweldens, "The lifting scheme: A construction of second generation wavelets," *SIAM J. Math. Anal.*, vol. 29, no. 2, pp. 511–546, Mar. 1998.

[147] J. Garrote, C. Brites, J. Ascenso, and F. Pereira, "Lenslet light field imaging scalable coding," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2150–2154.

[148] D. Ruefenacht, A. T. Naman, R. Mathew, and D. Taubman, "Base-anchored model for highly scalable and accessible compression of multiview imagery," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3205–3218, Jul. 2019.

[149] A. Ortega, P. Frossard, J. Kovacevic, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[150] F. Chung, *Spectral Graph Theory* (CBMS Regional Conference Series in Mathematics), vol. 92. Providence, RI, USA: American Mathematical Society, Dec. 1996.

[151] V. Elias and W. Martins, "On the use of graph Fourier transform for light-field compression," *J. Commun. Inf. Syst.*, vol. 33, no. 1, pp. 92–103, 2018.

[152] A. Sandryhaila and J. M. F. Moura, "Nearest-neighbor image model," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2521–2524.

[153] Y.-H. Chao, G. Cheung, and A. Ortega, "Pre-demosaic light field image compression using graph lifting transform," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3240–3244.

[154] E. Martinez-Enriquez, J. Cid-Sueiro, F. Diaz-de-Maria, and A. Ortega, "Directional transforms for video coding based on lifting on graphs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 4, pp. 933–946, Apr. 2018.

[155] M. Rizkallah, X. Su, T. Maugey, and C. Guillemot, "Geometry-aware graph transforms for light field compact representation," *IEEE Trans. Image Process.*, vol. 29, pp. 602–616, 2020.

[156] M. Hog, N. Sabater, and C. Guillemot, "Superrays for efficient light field processing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1187–1199, Oct. 2017.

[157] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Seattle, WA, USA, Jul. 2016, pp. 1–4.

[158] P. Astola and I. Tabus, "WaSP: Hierarchical warping, merging, and sparse prediction for light field image compression," in *Proc. 7th Eur. Workshop Vis. Inf. Process. (EUVIP)*, Nov. 2018, pp. 1–6.

[159] P. Astola and I. Tabus, "Light field compression of HDCA images combining linear prediction and JPEG 2000," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1860–1864.

[160] E. Elharar, A. Stern, O. Hadar, and B. Javidi, "A hybrid compression method for integral images using discrete wavelet transform and discrete cosine transform," *J. Display Technol.*, vol. 3, no. 3, pp. 321–325, Sep. 2007.

[161] M. Mazri and A. Aggoun, "Compression of 3D integral images using wavelet decomposition," in *Proc. Vis. Commun. Image Process.*, Lugano, Switzerland, Jun. 2003, pp. 1181–1192.

[162] A. Aggoun and M. Mazri, "Wavelet-based compression algorithm for still omnidirectional 3D integral images," *Signal, Image Video Process.*, vol. 2, no. 2, pp. 141–153, Jun. 2008.

[163] S. Kishk, H. E. M. Ahmed, and H. Helmy, "Integral images compression using discrete wavelets and PCA," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 4, no. 2, pp. 65–78, Jun. 2011.

[164] E. A. da Silva, M. P. Pereira, G. Alves, V. Testoni, C. Pagliari, M. B. de Carvalho, F. Pereira, and G. Arcia, "Core experiments set 3 for JPEG PLENO, CE3.2B evaluation of random access extensions to architecture," ISO/IEC JTC1/SC29/WG1, La Jolla, CA, USA, Tech. Rep. ISO/IECJTC1/SC29/WG1M79029, Apr. 2018.

[165] E. A. da Silva, M. P. Pereira, G. Alves, C. Schueler, C. Pagliari, M. B. de Carvalho, F. Pereira, and V. Testoni, "Exploration studies set 3 for JPEG Pleno, CE3.1 performance assessment of verification model 2.0," ISO/IEC JTC1/SC29/WG1, Lisbon, Portugal, Tech. Rep. ISO/IEC JTC1/SC29/WG1 M82030, Jan. 2019.

[166] M. Forman, "Compression of integral 3D TV pictures," in *Proc. 5th Int. Conf. Image Process. Appl.*, Heriot-Watt University, U.K., Aug. 1995, pp. 584–588.

[167] P. Ramanathan, M. Flierl, and B. Girod, "Multi-hypothesis prediction for disparity compensated light field compression," in *Proc. Int. Conf. Image Process.*, vol. 2, Oct. 2001, pp. 101–104.

[168] N. Sgouros, A. Andreou, M. Sangriotis, P. Papageorgad, D. Maroulis, and N. Theofanous, "Compression of IP images for autostereoscopic 3D imaging applications," in *Proc. 3rd Int. Symp. Image Signal Process. Anal.*, vol. 1. Rome, Italy, Sep. 2003, pp. 223–227.

[169] *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s—Part 2: Video*. ISO/IEC document 11172-2, 1993.

[170] S. Yeom, A. Stern, and B. Javidi, "Compression of 3D color integral images," *Opt. Express*, vol. 12, no. 8, p. 1632, Apr. 2004.

[171] *Information Technology–Generic Coding of Moving Pictures and Associated Audio Information: Video*. ITU-T Recommendation document H.262, Feb. 2012.

[172] N. Sgouros, I. Kontaxakis, and M. Sangriotis, "Effect of different traversal schemes in integral image coding," *Appl. Opt.*, vol. 47, no. 19, p. D28, Jul. 2008.

[173] C.-H. Yoo, H.-H. Kang, and E.-S. Kim, "Enhanced compression of integral images by combined use of residual images and MPEG-4 algorithm in three-dimensional integral imaging," *Opt. Commun.*, vol. 284, no. 20, pp. 4884–4893, Sep. 2011.

[174] H.-H. Kang, D.-H. Shin, and E.-S. Kim, "Efficient compression of motion-compensated sub-images with Karhunen–Loeve transform in three-dimensional integral imaging," *Opt. Commun.*, vol. 283, no. 6, pp. 920–928, Mar. 2010.

[175] J.-H. Lee, C.-H. Yoo, H.-H. Kang, and E.-S. Kim, "Compression scheme by use of motion-compensated residual image transformed from elemental image array in three-dimensional integral imaging," *Proc. SPIE*, vol. 7864, Jan. 2011, Art. no. 78640Y.

[176] V. Sze, M. Budagavi, and G. J. Sullivan, *High Efficiency Video Coding (HEVC): Algorithms and Architectures* (Integrated Circuits and Systems). Cham, Switzerland: Springer, 2014.

[177] F. Bossen, "Common HM test conditions and software reference configurations," ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep. JCTVC-L1100, 2013.

[178] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo-Sequence-Based 2-D hierarchical coding structure for light-field image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1107–1119, Oct. 2017.

[179] H.-H. Kang, J.-H. Lee, and E.-S. Kim, "Enhanced compression rate of integral images by using motion-compensated residual images in three-dimensional integral-imaging," *Opt. Express*, vol. 20, no. 5, pp. 5440–5459, Feb. 2012.

[180] H.-W. Lee, J.-H. Lee, H.-H. Kang, and E.-S. Kim, "Compression enhancement using the hybrid motion estimation in sub-image array transformed from elemental image array in three-dimensional integral image," *Proc. SPIE*, vol. 8498, Oct. 2012, Art. no. 849804.

[181] R. Olsson, M. Sjostrom, and Y. Xu, "A combined pre-processing and H.264-compression scheme for 3D integral images," in *Proc. Int. Conf. Image Process.*, Atlanta, GA, USA, Oct. 2006, pp. 513–516.

[182] R. Olsson, M. Sjöström, Y. Xu, "Evaluation of a combined pre-processing and H.264-compression scheme for 3D integral images," *Proc. SPIE*, vol. 6508, Jan. 2007, Art. no. 65082C.

[183] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4733–4737.

[184] S. Mattoccia, F. Tombari, and L. Di Stefano, "Fast full-search equivalent template matching by enhanced bounded correlation," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 528–538, Apr. 2008.

[185] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of lytro-illum light fields," in *Proc. Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Orleans, France, Nov. 2015, pp. 494–497.

[186] H. P. Hariharan, T. Lange, and T. Herfet, "Low complexity light field compression based on pseudo-temporal circular sequencing," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2017, pp. 1–5.

[187] H. Amirpour, M. Pereira, and A. Pinheiro, "High efficient snake order pseudo-sequence based light field image compression," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2018, p. 397.

[188] H. P. Cong, S. Perry, T. A. Vu, and X. HoangVan, "Joint exploration model based light field image coding: A comparative study," in *Proc. 4th NAFOSTED Conf. Inf. Comput. Sci.*, Nov. 2017, pp. 308–313.

[189] The WebM Project. (2019). *VP9 Video Codec*. [Online]. Available: https://www.webmproject.org/vp9/

[190] JVET. (2019). *JVET JEM Software Repository*. [Online]. Available: https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/

[191] S. Zhao, Z. Chen, K. Yang, and H. Huang, "Light field image coding with hybrid scan order," in *Proc. Vis. Commun. Image Process. (VCIP)*, Nov. 2016, pp. 1–4.

[192] C. Jia, Y. Yang, X. Zhang, X. Zhang, S. Wang, S. Wang, and S. Ma, "Optimized inter-view prediction based light field image compression with adaptive reconstruction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4572–4576.

[193] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Seattle, WA, USA, Jul. 2016, pp. 1–4.

[194] I. Viola, M. Rerabek, and T. Ebrahimi, "Comparison and evaluation of light field image coding approaches," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1092–1106, Oct. 2017.

[195] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[196] S. Adedoyin, W. A. C. Fernando, A. Aggoun, and K. M. Kondoz, "Motion and disparity estimation with self adapted evolutionary strategy in 3D video coding," *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1768–1775, Nov. 2007.

[197] S. Adedoyin, W. A. C. Fernando, and A. Aggoun, "A joint motion & disparity motion estimation technique for 3D integral video compression using evolutionary strategy," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 732–739, May 2007.

[198] J. Wei, S. Wang, Y. Zhao, and F. Jin, "Hierarchical prediction structure for subimage coding and multithreaded parallel implementation in integral imaging," *Appl. Opt.*, vol. 50, no. 12, p. 1707, Apr. 2011.

[199] G. Wang, W. Xiang, M. Pickering, and C. W. Chen, "Light field multiview video coding with two-directional parallel inter-view prediction," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5104–5117, Nov. 2016.

[200] W. Ahmad, R. Olsson, and M. Sjostrom, "Interpreting plenoptic images as multi-view sequences for improved compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4557–4561.

[201] W. Ahmad, R. Olsson, and M. Sjostrom, "Towards a generic compression solution for densely and sparsely sampled light field data," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 654–658.

[202] S. Adedoyin, W. A. C. Fernando, A. Aggoun, and W. A. R. J. Weerakkody, "An ES based effecient motion estimation technique for 3D integral video compression," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, San Antonio, TX, USA, Sep./Oct. 2007, pp. III-393–III-396.

[203] J. Dick, H. Almeida, L. D. Soares, and P. Nunes, "3D holoscopic video coding using MVC," in *Proc. IEEE Int. Conf. Comput. Tool (EUROCON)*, Lisbon, Portugal, Apr. 2011, pp. 1–4.

[204] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Full parallax super multi-view video coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 135–139.

[205] C. Conti, P. Nunes, and L. D. Soares, "Inter-layer prediction scheme for scalable 3-D holoscopic video coding," *IEEE Signal Process. Lett.*, vol. 20, no. 8, pp. 819–822, Aug. 2013.

[206] Y. Li, R. Olsson, and M. Sjostrom, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–4.

[207] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 4562–4566.

[208] S.-L. Yu and C. Chrysafis, "New intra prediction using intra-macroblock motion compensation," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q6, Fairfax, VA, USA, Tech. Rep. JVT-C151, May 2002.

[209] S.-L. Yu and C. Chrysafis, "Intra-prediction using intra-macroblock motion compensation," U.S. Patent 7 120 196 B2 Oct. 10, 2006. [Online]. Available: http://www.google.com/patents/US7120196

[210] C. Conti, J. Lino, P. Nunes, L. D. Soares, and P. L. Correia, "Spatial prediction based on self-similarity compensation for 3D holoscopic image and video coding," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 961–964.

[211] C. Conti, P. Nunes, and L. D. Soares, "New HEVC prediction modes for 3D holoscopic video coding," in *Proc. 19th IEEE Int. Conf. Image Process.*, Orlando, FL, USA, Sep. 2012, pp. 1325–1328.

[212] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process., Image Commun.*, vol. 42, pp. 59–78, Mar. 2016.

[213] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.

[214] M. Budagavi and D.-K. Kwon, "Intra motion compensation and entropy coding improvements for HEVC screen content coding," in *Proc. Picture Coding Symp. (PCS)*, San Jose, CA, USA, Dec. 2013, pp. 365–368.

[215] D.-K. Kwon and M. Budagavi, "Fast intra block copy (IntraBC) search for HEVC screen content coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Melbourne, VIC, Australia, Jun. 2014, pp. 9–12.

[216] R. Joshi, J. Xu, R. Cohen, S. Liu, Z. Ma, and Y. Ye, "Screen content coding test model 1 (SCM 1)," ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Valencia, Spain, Tech. Rep. JCTVC-Q1014, Apr. 2014.

[217] Y. Li, M. Sjostrom, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, Jul. 2016.

[218] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Seattle, WA, USA, Jul. 2016, pp. 1–4.

[219] C. Conti, P. Nunes, and L. Ducla Soares, "Light field image coding with jointly estimated self-similarity bi-prediction," *Signal Process., Image Commun.*, vol. 60, pp. 144–159, Feb. 2018.

[220] D. Liu, P. An, R. Ma, C. Yang, and L. Shen, "3D holoscopic image coding scheme using HEVC with Gaussian process regression," *Signal Process., Image Commun.*, vol. 47, pp. 438–451, Sep. 2016.

[221] C. Conti, L. D. Soares, and P. Nunes, "Scalable light field coding with support for region of interest enhancement," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 1855–1859.

[222] X. Jin, H. Han, and Q. Dai, "Image reshaping for efficient compression of plenoptic content," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1173–1186, Oct. 2017.

[223] L. F. R. Lucas, C. Conti, P. Nunes, L. D. Soares, N. M. M. Rodrigues, C. L. Pagliari, E. A. Da Silva, and S. M. M. De Faria, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, Sep. 2014, pp. 11–15.

[224] X. Jin, H. Han, and Q. Dai, "Plenoptic image coding using macropixel-based intra prediction," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3954–3968, Aug. 2018.

[225] H. Han, X. Jin, and Q. Dai, "Lenslet image compression using adaptive macropixel prediction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4008–4012.

[226] R. J. S. Monteiro, P. J. L. Nunes, N. M. M. Rodrigues, and S. M. M. Faria, "Light field image coding using high-order intrablock prediction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1120–1131, Oct. 2017.

[227] C. Conti, L. D. Soares, and P. Nunes, "Light field coding with Field-of-View scalability and exemplar-based interlayer prediction," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2905–2920, Nov. 2018.

[228] M. Türkan and C. Guillemot, "Image prediction based on neighbor-embedding methods," *IEEE Trans.Image Process.*, vol. 21, no. 4, pp. 1885–1898, Apr. 2012.

[229] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, Denver, CO, USA, 2000, pp. 556–562.

[230] S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[231] T. K. Tan, C. S. Boon, and Y. Suzuki, "Intra prediction by template matching," in *Proc. Int. Conf. Image Process.*, Atlanta, GA, USA, Oct. 2006, pp. 1693–1696.

[232] D. Liu, P. An, R. Ma, and L. Shen, "Disparity compensation based 3D holoscopic image coding using HEVC," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Chengdu, China, Jul. 2015, pp. 201–205.

[233] C. Conti, L. Ducla Soares, and P. Nunes, "Influence of self-similarity on 3D holoscopic video coding performance," in *Proc. 18th Brazilian Symp. Multimedia Web (WebMedia)*, 2012, pp. 131–134.

[234] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. da Silva, and L. Soares, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Seattle, WA, USA, Jul. 2016, pp. 1–4.

[235] R. Zhong, S. Wang, B. Cornelis, Y. Zheng, J. Yuan, and A. Munteanu, "Efficient directional and L1-optimized intra-prediction for light field image compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sep. 2017, pp. 1172–1176.

[236] R. Zhong, I. Schiopu, B. Cornelis, S.-P. Lu, J. Yuan, and A. Munteanu, "Dictionary learning-based, directional, and optimized prediction for lenslet image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1116–1129, Apr. 2019.

[237] J. Chen, J. Hou, and L.-P. Chau, "Light field compression with disparity-guided sparse coding based on structural key views," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 314–324, Jan. 2018.

[238] A. Smolic, K. Muller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 2448–2451.

[239] Y. Piao and X. Yan, "Sub-sampling elemental images for integral imaging compression," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Shanghai, China, Nov. 2010, pp. 1164–1168.

[240] P. Yan and Y. Xianyuan, "Integral image compression based on optical characteristic," *IET Comput. Vis.*, vol. 5, no. 3, p. 164, 2011.

[241] C. Choudhury and S. Chaudhuri, "Disparity based compression technique for focused plenoptic images," in *Proc. Indian Conf. Comput. Vis. Graph. Image Process. (ICVGIP)*, Bangalore, India, Dec. 2014, pp. 1–6.

[242] D. B. Graziosi, Z. Y. Alpaslan, and H. S. El-Ghoroury, "Depth assisted compression of full parallax light fields," *Proc. SPIE*, vol. 9391, Mar. 2015, Art. no. 93910Y.

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

[244] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Integral images compression scheme based on view extraction," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug. 2015, pp. 101–105.

[245] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Improved integral images compression based on multi-view extraction," *Proc. SPIE*, vol. 9971, Sep. 2016, Art. no. 99710L.

[246] X. Jiang, M. Le Pendu, and C. Guillemot, "Light field compression using depth image based view synthesis," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 19–24.

[247] X. Huang, P. An, L. Shen, and R. Ma, "Efficient light field images compression method based on depth estimation and optimization," *IEEE Access*, vol. 6, pp. 48984–48993, 2018.

[248] P. Astola and I. Tabus, "Coding of light fields using disparity-based sparse prediction," *IEEE Access*, vol. 7, pp. 176820–176837, 2019.

[249] E. Dib, M. L. Pendu, and C. Guillemot, "Light field compression using Fourier disparity layers," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3751–3755.

[250] B. Heriard-Dubreuil, I. Viola, and T. Ebrahimi, "Light field compression using translation-assisted view estimation," in *Proc. Picture Coding Symp. (PCS)*, Nov. 2019.

[251] W. Ahmad, S. Vagharshakyan, M. Sjostrom, A. Gotchev, R. Bregovic, and R. Olsson, "Shearlet transform based prediction scheme for light field compression," in *Proc. Data Compress. Conf.*, Mar. 2018, p. 396.
[252] M. Le Pendu, C. Guillemot, and A. Smolic, "A Fourier disparity layer representation for light fields," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5740–5753, Nov. 2019.
[253] I. Viola, H. P. Maretic, P. Frossard, and T. Ebrahimi, "A graph learning approach for light field image compression," *Proc. SPIE*, vol. 10752, Sep. 2018, Art. no. 107520E.
[254] J. Hou, J. Chen, and L.-P. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 517–530, Feb. 2019.
[255] M. Rizkallah, X. Su, T. Maugey, and C. Guillemot, "Graph-based transforms for predictive light field compression based on super-pixels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1718–1722.
[256] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[257] X. Su, M. Rizkallah, T. Mauzev, and C. Guillemot, "Rate-distortion optimized super-ray merging for light field compression," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1850–1854.
[258] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, and J. Yang, "Light field image compression based on deep learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.
[259] C. Jia, X. Zhang, S. Wang, S. Wang, S. Pu, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 177–189, Mar. 2019.
[260] N. Bakir, W. Hamidouche, O. Deforges, K. Samrouth, and M. Khalil, "Light field image compression based on convolutional neural networks and linear approximation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1128–1132.
[261] ISO/IEC JTC1/SC29/WG1, "Press release of the 84th meeting," ISO/IEC JTC1/SC29/WG1, Brussels, Belgium, Tech. Rep. ISO/IECJTC1/SC29/WG1N84005, 2019.
[262] G. Lafruit, D. Bonatto, C. Tulvan, M. Preda, and L. Yu, "Understanding MPEG-I coding standardization in immersive VR/AR applications," *SMPTE Motion Imag. J.*, vol. 128, no. 10, pp. 33–39, Nov. 2019.

CAROLINE CONTI (Member, IEEE) received the B.Sc. degree in electrical engineering from the Universidade de São Paulo (USP), Brazil, in 2010, and the Ph.D. degree in information science and technology from the Instituto Universitário de Lisboa (ISCTE-IUL), Portugal, in 2017. She is currently a Postdoctoral Researcher with the Multimedia Signal Processing Group, Instituto de Telecomunicações, Portugal. She is also an Invited Assistant Professor with the Information Science and Technology Department, ISCTE-IUL. Her research interests include immersive visual technologies and image and video processing, including light field processing and coding. She has contributed more than 20 articles to international journals and conferences in these areas. In addition, she has participated in many national and international projects related to light field processing and coding. She also acts as a reviewer for various IEEE and EURASIP journals and conferences.

LUÍS DUCLA SOARES (Senior Member, IEEE) received the Licenciatura and Ph.D. degrees in electrical and computer engineering from the Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Portugal, in 1996 and 2004, respectively. He is currently a Senior Researcher with the Multimedia Signal Processing Group, Instituto de Telecomunicações, Portugal. He is also an Associate Professor with the Information Science and Technology Department, Instituto Universitário de Lisboa (ISCTE-IUL), Portugal. His research interests are centered on image and video coding/processing, including light field coding and processing and biometric recognition. He has contributed more than 65 articles to international journals and conferences in these areas (20 of which on light field coding). In addition, he has participated in the development of the MPEG-4 Visual standard, as well as in several national and international projects. He is also a member of the Editorial Board of the *EURASIP Signal Processing* (Elsevier) journal. In parallel, he acts as a reviewer of several IEEE, IET, and EURASIP journals and conferences.

PAULO NUNES (Member, IEEE) received the degree in electrical and computers engineering from the Instituto Superior Técnico (IST), Universidade Técnica de Lisboa, Portugal, in 1992, and the M.Sc. and Ph.D. degrees in electrical and computers engineering from IST, in 1996 and 2007, respectively. He is currently an Assistant Professor with the Information Science and Technology Department, University Institute of Lisbon (ISCTE-IUL), Portugal, and a Senior Research with the Instituto de Telecomunicações, Portugal. He has coordinated and participated in various national and international (EU) funded projects and has acted as project evaluator for the European Commission. He acts often as a reviewer for various conferences and journals and a member of the program and organizing committees of various international conferences. He has contributed more than 60 articles. His current research interests include 2D/3D image and video processing and coding, namely light field image and video processing and coding.