CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Repositório Institucional do ISCTE-IUL

ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2018-10-15

Deposited version:
Post-print

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Conti, C., Soares, L. D. & Nunes, P. (2018). Light field coding with field of view scalability and exemplar-based inter-layer prediction. IEEE Transactions on Multimedia. 20 (11), 2905-2920

Further information on publisher's website:
10.1109/TMM.2018.2825882

Publisher's copyright statement:
This is the peer reviewed version of the following article: Conti, C., Soares, L. D. & Nunes, P. (2018). Light field coding with field of view scalability and exemplar-based inter-layer prediction. IEEE Transactions on Multimedia. 20 (11), 2905-2920, which has been published in final form at https://dx.doi.org/10.1109/TMM.2018.2825882. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Light Field Coding with Field of View Scalability and Exemplar-Based Inter-Layer Prediction

Caroline Conti, *Member IEEE*, Luís Ducla Soares, *Senior Member IEEE*, and Paulo Nunes, *Member IEEE*

***Abstract*— Light field imaging based on microlens arrays – a.k.a. holoscopic, plenoptic, and integral imaging – has currently risen up as a feasible and prospective technology for future image and video applications. However, deploying actual light field applications will require identifying more powerful representations and coding solutions that support arising new manipulation and interaction functionalities. In this context, this paper proposes a novel scalable coding solution that supports a new type of scalability, referred to as Field of View scalability. The proposed scalable coding solution comprises a base layer compliant with the High Efficiency Video Coding (HEVC) standard, complemented by one or more enhancement layers that progressively allow richer versions of the same light field content in terms of content manipulation and interaction possibilities. Additionally, for achieving high compression performance in the enhancement layers, novel exemplar-based inter-layer coding tools are also proposed, namely: i) a direct prediction based on exemplar texture samples from lower layers, and ii) an inter-layer compensated prediction using a reference picture that is built relying on an exemplar-based algorithm for texture synthesis. Experimental results demonstrate the advantages of the proposed scalable coding solution to cater for users with different preferences/requirements in terms of interaction functionalities, while providing better rate-distortion performance (independently of the optical setup used for acquisition) compared to HEVC and other scalable light field coding solutions in the literature.**

***Index Terms*— Light Field, Holoscopic, Plenoptic, Integral Imaging, Field of View Scalability, Image Compression, HEVC**

I. Introduction

THE recent advances in optical and sensor manufacturing allow having richer forms of visual data, where not only the spatial information about the three-dimensional (3D) scene is represented but also angular viewing direction – the so-called four-dimensional (4D) light field/radiance sampling [1].

In the context of Light Field (LF) imaging technologies, the approach based on a single-tier camera equipped with a Microlens Array (MLA) [2] (hereinafter referred simply to as LF camera) has become a promising approach, being applicable in many different areas of research, such as 3D television [3], richer photography capturing [4], [5], biometric recognition [6], and medical imaging [7].

Manuscript received March 24, 2018. This work was supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal) under SFRH / BD / 79480 / 2011 grant and UID / EEA / 50008 / 2013 project.

C. Conti, L.D. Soares, and P. Nunes are with Instituto Universitário de Lisboa (ISCTE-IUL) and Instituto de Telecomunicações, Lisbon, Portugal (e-mail: {caroline.conti, lds, paulo.nunes}@lx.it.pt).

Recognizing the potential of this emerging technology, as well as the new challenges that need to be overcome for successfully introducing light field media applications into the consumer market, novel standardization initiatives are also emerging. Notably, the Joint Photographic Experts Group (JPEG) committee has launched the JPEG Pleno standardization initiative [8], and the Moving Picture Experts Group (MPEG) has recently started a new work item on coded representations for immersive media (MPEG-I) [9]. The challenge to provide a LF representation with convenient spatial resolution and viewing angles requires handling a huge amount of data and, thus, efficient coding becomes of utmost importance. Another key requirement when designing an efficient LF representation and coding solution is to facilitate future interactive LF media applications with the new manipulation functionalities supported by the LF content. The advantages of enabling interactive media applications has been previously studied in the literature for a large range of media modalities, such as: i) interactive streaming of high resolution images [10]; ii) interactive multiview video streaming [11], [12]; and iii) interactive streaming of light field images captured by high density camera-arrays [13]. In this context, although standardized LF representation and coding solutions are still in an early stage of development, various LF coding solutions have been already proposed in the literature.

A. Previous Work

LF coding solutions available in the literature can be categorized in the following three main approaches, depending on the data format and prediction schemes that are adopted.

1. LF raw data-based coding

This category corresponds to encoding and transmitting the (raw) LF image in its entirety. As a result of the used optical system, the LF image corresponds to a two-dimensional (2D) array of micro-images (MIs), and a significant cross-correlation exists between these MIs in a neighborhood [14]. To exploit this inherent MI cross-correlation, a non-local spatial prediction scheme is used, which is usually integrated (but not necessarily so) on a standard 2D image coding solution, such as the state-of-the-art High Efficiency Video Coding (HEVC). Following this approach, it has been shown, in [14]–[17], that efficient LF image coding can be achieved by using the concept of Self-Similarity (SS) compensated prediction. Similarly to motion estimation, a block-based matching algorithm is used to estimate the 'best' predictor block for the current block over the previously coded and reconstructed area of the current picture. This predictor block can be generated from a single candidate block [14], [15] or

from a combination of two different candidate blocks [16], [17]. Furthermore, an alternative prediction scheme based on locally linear embedding was proposed in [18], [19], where a set of nearest neighbor patches were linearly combined to predict the current block. More recently, in [20], a prediction scheme based on Gaussian Process Regression (GPR) was also proposed for LF image coding. In this case, the prediction was modeled from a set of nearest neighbor patches as a non-linear (Gaussian) process, and GPR was then used for estimating the predictor block.

However, although these coding schemes have shown to achieve significant compression gains when compared to state-of-the-art 2D image coding solutions [14]–[20], transmitting the entire LF data without a scalable bitstream may represent a serious problem since the end-user needs to wait until the entire LF data arrives before he/she can visualize and interact with the content.

2. Multiview- and PVS-based LF coding

Some other coding schemes proposed to extract the viewpoint images (VIs) from the LF content to be represented as multiview content [21]–[24], or as a Pseudo Video Sequence (PVS) [25]–[28]. In these coding approaches, each VI is constructed by simply extracting one pixel with the same relative position from each MI. The VI-based multiview content is then encoded using a 3D video coding solution, for instance, the Multiview Video Coding (MVC) [29] (in [21]–[23]) and the multiview extension of HEVC (MV-HEVC) [30] (in [24]). An advantage of using a standard 3D video coding solution is that scalability and backward compatibility are straightforwardly supported. Differently, the PVS of VIs is encoded using a 2D video coding standard, such as H.264/AVC [29] (in [25]), or HEVC [31] (in [26]–[28]). Although conceptually different (in terms of coding architecture), both multiview- and PVS-based coding approaches have the same basic purpose of providing an efficient prediction configuration for better exploiting the correlations between the views. For this, different scanning patterns for ordering the views, as well as different inter-view prediction structures are proposed in [21]–[27] to improve the coding efficiency.

However, although these approaches can provide scalability in the coded bitstream, it is possible to observe in the literature (e.g., in [15], [25], [27]) that their coding performance may vary significantly depending on the LF optical setup that is used for acquiring the LF content. As will be further discussed in Section II.B, there are basically two LF camera setups: i) unfocused (a.k.a. plenoptic camera 1.0); and ii) focused (a.k.a. plenoptic camera 2.0). For an LF image captured using the unfocused LF camera setup, each VI represents an orthographic projection of the captured scene that is all in focus [2]. On the other hand, for an LF image captured using a focused camera setup, a VI can be interpreted as a subsampled perspective of the captured scene (as in [32]) or as a low resolution rendered view that is all out of focus (as in [33]), which, consequently, presents aliasing artifacts. Furthermore, using an MLA with larger microlenses pitch leads to greater aliasing in the extracted VI [33]. Since these aliasing artifacts are difficult to predict and to compress, multiview- and PVS-based LF coding solutions usually present a significantly worse coding performance for LF content captured using a focused LF camera setup (as shown in [15]).

It should be noticed that an alternative to the multiview representation based on these aliased VIs for focused LF cameras was proposed in [34], [35] using super-resolved rendered views. In this case, a scalable coding approach is used, which supports backward compatibility to legacy 2D and 3D multiview displays in the lower layer while the highest layer supports the entire LF content. However, with this coding architecture, the end-user still needs to receive the entire scalable bitstream to have a viewing experience with the novel interaction functionalities supported by the LF content.

3. Disparity-assisted LF coding

Other coding schemes proposed to represent the LF data by a subsampled set of MIs with their associated disparity information [36]–[38]. As firstly proposed in [39], the grid of MIs is subsampled to remove the redundancy between neighboring MIs and to achieve compression. Thus, only the remainder set of MIs and associated disparity are encoded and transmitted. At the decoder side, the LF data is reconstructed by simply applying a disparity shift (in [36], [38]) or by using a Depth Image Based Rendering (DIBR) algorithm modified to support the multiple MIs as input views (in [37]), and followed by an inpainting algorithm to fill in the missing areas. However, in real-world images, the disparity/depth information is estimated from the acquired LF raw data, which introduces some inaccuracies. Hence, the quality of the reconstructed MIs – and, consequently, the quality of rendered views – is severely affected by these inaccuracies at the encoder side. Additionally, due to occlusion problems and quantization errors when (lossy) encoding this disparity/depth maps, some synthesized MIs might present too many missing areas to be filled [37], thus introducing even further inaccuracies. Instead of uniformly selecting the MIs as in [38], the selection is performed adaptively in [36], [37], so as to obtain better view reconstruction.

However, a common characteristic of these approaches is that the quality of rendered views is negatively affected by the inaccuracies in the synthesis of the missing MIs, thus presenting a significant drop in the Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) Index, mainly for natural content. In [38], the residue between the reconstructed LF image and the original LF image is also encoded and transmitted in an enhancement coding layer so as to provide better rendered views. Nevertheless, in this case [38], the end-user still needs to decode the entire scalable bitstream to visualize these rendered views with better quality.

B. Motivations and Contributions

Among the advantages of employing a LF imaging approach is the ability to open new degrees of freedom in terms of content production and manipulation, supporting manipulation functionalities not straightforwardly available in conventional imaging systems, namely: post-production refocusing, changing depth-of-field, and changing viewing perspective. This means, for instance, that the user can receive captured LF content and interactively adjust the plane of focus and depth-of-field of the rendered content. Moreover, as part of the creative process, the content creator can define how to

organize the LF content to be sent to multiple end-users who may be using different display technologies, as well as applications, that allow different levels of interaction. In this sense, an efficient scalable LF coding architecture is desirable to accommodate in a single compressed bitstream a variety of sub-bitstreams appropriate for users with different preferences/requirements and various application scenarios: from the user who wants to have a simple 2D version of the LF content without actively interacting with it; to the user who wants full immersive and interactive LF visualization.

Based on the abovementioned application scenarios, the contributions of this paper are:

- **Field Of View (FOV) scalability** – To support the richer and flexible interaction functionalities that arise in LF imaging applications, a new scalability concept, named FOV scalability, and a novel Field Of View Scalable Light Field Coding (FOVS-LFC) solution are proposed. Taking advantage of the 4D radiance distribution, the FOV scalability progressively supports richer forms of the same LF content by hierarchically organizing the angular information of the captured LF data. More specifically, the base layer contains a subset of the LF raw data with narrower FOV, which can be used to render a 2D version of the content with very limited rendering functionalities. Following the base layer, one or more enhancement layers are defined to represent the necessary information to obtain more immersive LF visualization with a wider FOV. Therefore, this new type of scalability creates bitstreams adaptable to different levels of user interaction, allowing increasing degrees of freedom in content manipulation at each higher layer. This means that, for instance, a user who wants to have a simple 2D visualization will only need to extract the base layer of the bitstream, thus reducing the necessary bitrate and the required computational power. On the other hand, a user who wants to creatively decide how to interact with the LF content can promptly start visualizing and flexibly manipulating the LF content, even over limited bandwidth connections, by extracting only the adequate bitstream subsets (which fit in the available bitrate). Additionally, this coding architecture enables easy support to quality scalability and Region Of Interest (ROI) coding [40].
- **Exemplar-based Inter-Layer (IL) coding tools** – To improve the efficiency when coding an enhancement layer, two novel inter-layer prediction schemes are also proposed: i) a direct IL prediction, and ii) an IL compensated prediction. In the direct IL prediction, a set of samples from a previously coded layer is used as exemplar samples for estimating a good prediction block. Therefore, no further information about the used predictor block needs to be transmitted to the decoder side. The IL compensated prediction relies on an IL reference picture, which is constructed using samples from previously coded layers and a new exemplar-based [41] algorithm for texture synthesis.

In a nutshell, the proposed FOVS-LFC solution is able to overcome some of the limitations of previously proposed LF coding solutions by providing: i) a scalable bitstream that supports richer and flexible manipulation functionalities (such as refocusing, changing perspective and depth-of-field) and backward compatibility with the current state-of-the-art HEVC standard [31]; ii) support for quality scalability and ROI coding; iii) high compression efficiency for LF content captured using different LF camera setups; as well as iv) high quality of rendered views in all hierarchical layers.

C. Paper Outline

The remainder of this paper is organized as follows. Section II briefly reviews the LF imaging principles that are important to understand the concepts discussed in this paper. Section III presents the concept of FOV scalability, while Section IV describes the FOVS-LFC solution architecture. Section V describes the exemplar-based inter-layer coding tools that are proposed for LF enhancement layer coding. Section VI presents the test conditions and experimental results; and, finally, Section VII concludes the paper.

II. LF Imaging Technology

As illustrated in Fig. 1a, an LF camera basically comprises a main lens, and an MLA that lies at a distance b of the image sensor. Therefore, different from a conventional 2D camera that captures an image by integrating the intensities of all rays (from all directions) impinging each sensor element (hereinafter referred to as pixel[1]) at position (x,y); in an LF camera, each pixel collects the light of a single ray (or of a thin bundle of rays) from a given angular direction (θ,φ) that converges on a specific microlens at position (x,y) in the array. This means that it is possible to sample the 4D radiance and organize it in a conventional 2D image, known as the (raw) LF image.

To simplify the visualization of this 4D function (coordinates x, y, θ and φ), the flat Cartesian ray-space diagram [1], [42] shown in Fig. 2a is used in this paper, where only two dimensions – in this case, x and θ – are represented.

A. LF Camera Setups

As discussed in [2], [33], different LF camera setups can be derived from the basic elements in Fig. 1a (i.e., a main lens, and an MLA at a distance b of the image sensor), namely:

- **Unfocused LF camera** – In this setup, the main lens is focused on the microlens plane while the microlenses are focused at infinity as illustrated in Fig. 2b (top) (the sensor is placed at the MLA focal length f, i.e., $b = f$ in Fig. 1a). Moreover, since microlenses usually have a much smaller focal length than the main lens, it is reasonable to admit that the main lens is at the microlenses optical infinity. Consequently, the radiance coming from the captured scene is refracted through the main lens and then split by each microlens in the array. This can be seen in Fig. 2b (bottom), where the captured radiance is split into different columns corresponding to the bundle of rays sampled as an MI at the sensor. Afterwards, the light rays that hit a single microlens are separated into different angular directions to be projected onto the pixels in the image sensor underneath. Hence, each small rectangle in Fig. 2b, corresponds to the tiny bundle of rays (with width given by the microlens aperture, d) that is integrated into a single pixel of the MI. Examples of LF cameras using this setup are the Lytro LF cameras [5].

[1] For the sake of simplicity, a pixel is here understood as a three-dimensional variable where each dimension contains the information of one color component: Red, Green, and Blue (RGB).

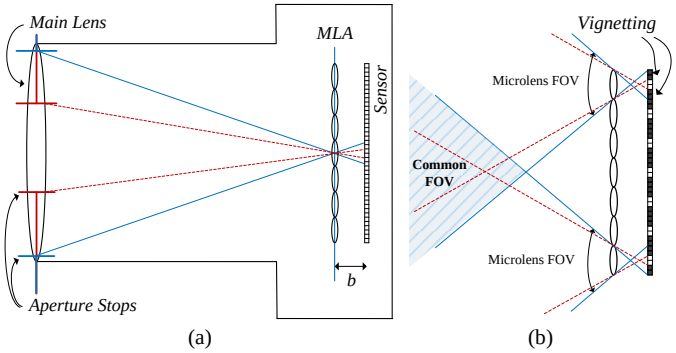


Fig. 1 LF camera: (a) Basic optical setup comprising a main lens, an MLA at a distance b of the image sensor. Two aperture sizes are shown by the blue and dashed red lines; (b) The FOV can be used as a measure of overall angular information in the LF content. If the main lens and MLA F-numbers are not adjusted, the FOV is restricted and MI vignetting is observed

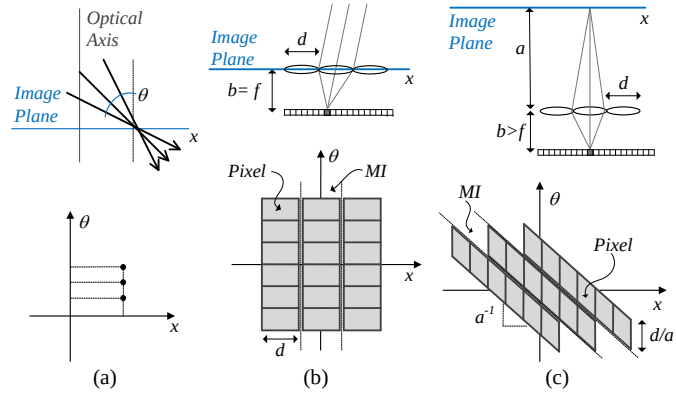


Fig. 2 Parameterization of the 4D radiance in an LF camera: (a) A single light ray is described by the position it intersects the plane x and its slope θ. Each possible ray in the ray diagram (top) corresponds to a different point in the Cartesian ray-space diagram (bottom); (b) Sampling the radiance at the main lens image plane for the unfocused LF camera (when $b = f$); and (c) Sampling the radiance at the image plane for the focused LF camera (when $b > f$)

• **Focused LF camera** – As discussed in [2], [33], the unfocused LF camera (Fig. 2b) can be generalized to an alternative camera setup that is referred to as focused LF camera [33]. Examples of focused LF cameras are the Raytrix LF cameras [4]. In this setup, the main lens and the MLA are both focused in an image plane at a distance a of the MLA plane, as illustrated in Fig. 2c (top). Thus, the main lens forms a relay system with each microlens, and the MLA works as a conventional camera array (with very low resolution and small baseline). As shown in [42], each MI will then capture what corresponds to a slanted stripe of the radiance (slope a^{-1}), depicted by the ray-space diagram in Fig. 2c (bottom). Consequently, this configuration allows an effective increase in spatial resolution at the price of a decrease in angular resolution [2]. Comparing the ray-space diagram in Fig. 2b and c (bottom), it is possible to see that an MI captured using the focused LF camera setup contains more spatial information (at x axis) than an MI captured using the unfocused LF camera setup (Fig. 2b). In a generalized LF camera, changing the distance b will change the slope a^{-1} in Fig. 2c (bottom) and, consequently, the balance between providing larger angular or spatial resolution in the captured LF image [2]. A notable limit is when $b \to f$, $a \to \infty$, and this generalized setup corresponds to the unfocused LF camera.

B. FOV in LF Cameras

The FOV of a lens (typically expressed by a measurement of area or angle) corresponds to the area of the scene over which objects can be reproduced by the lens. In a conventional 2D camera, the FOV is related to the lens focal length and the physical size of the sensor. In an LF camera, the microlens FOV is directly related to the aperture of the main lens. To illustrate this fact, Fig. 1a depicts the unfocused LF camera with two different aperture sizes (as shown by the blue and red aperture stops). As can be seen with the blue and the dashed red lines, all the rays coming from the focused subject will intersect at the MLA (at the image plane) and will then diverge until they reach the image sensor. Moreover, comparing the blue lines with the dashed red ones (in Fig. 1a), it is possible to see that the main lens aperture (or more specifically, the F-number[2] of the main lens) needs to be matched to the F-number of the MLA to guarantee that MIs receive homogeneous illumination on their entire area, as seen in the blue line case (Fig. 1a). Otherwise, in the case of the dashed red line (Fig. 1a), pronounced vignetting (with the shape of the main lens aperture) will be visible in each MI, as illustrated in Fig. 1b.

Moreover, as depicted in Fig. 1b, the common area where the FOV of all microlenses overlaps can be seen as a measure of the amount of angular information in the captured LF content. Note that, if there is MI vignetting (see dashed red lines in Fig. 1b), the microlens FOV will be further restricted and, consequently, the angular information will be narrowed. This means that it is possible to control the amount of angular information that is available in the captured LF content by adjusting the main lens aperture. This fact has motivated the FOV scalability concept that is presented in the following.

III. The FOV Scalability Concept

The basic idea of the proposed FOV scalability is to split the LF raw data into hierarchical subsets with partial angular information. Generally speaking, the FOV scalability can be thought of as a virtual increase in the main lens aperture (see Fig. 3a) from one scalable layer to the next higher layer, corresponding to a wider microlens FOV and virtual narrower vignetting inside each MI (along its border).

A. LF Data Organization for FOV Scalability

As was shown in Section II, each pixel underneath its corresponding microlens gathers light information from a given angular direction. Therefore, it is possible to split the overall angular information available in the captured LF image by properly selecting subsets of pixels from each MI. This concept is depicted in Fig. 3 for a hypothetical case in which three subsets of pixels are sampled from each MI. Therefore, the angular information is split into three hierarchical layers as seen in Fig. 3b (for the generalized focused LF camera setup). In each lower layer (from top to bottom in Fig. 3b), the

[2] In optical terminology, the F-number corresponds to the ratio between the lens focal length and its aperture diameter.

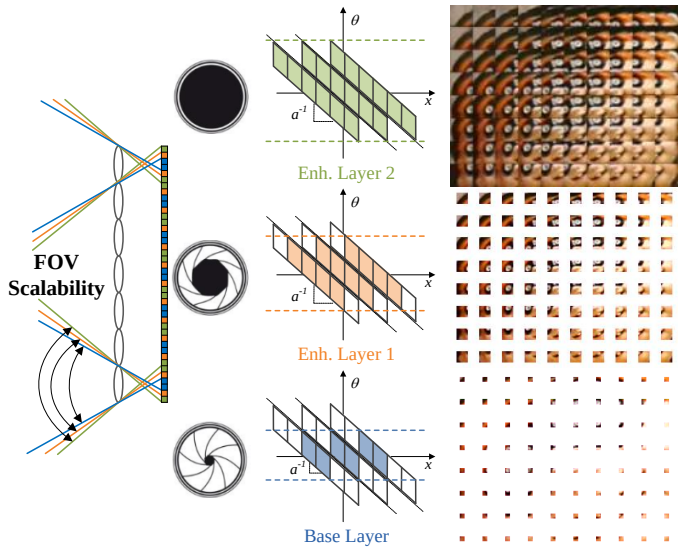


Fig. 3 The concept of FOV Scalability for a hypothetical three-layer approach: (a) Ray tracing diagram showing that three hierarchical layers of FOV Scalability can be sampled by properly selecting three subsets of pixels (with different colors) from each MI, corresponding to a virtual increase in the main lens aperture; (b) Corresponding ray-space diagram showing the angular information in each hierarchical layer; and (c) Illustrative example for gathering the three subsampled set of pixels from each MI. From the base layer (bottom) to the last enhancement layer (top), the FOV is wider and, consequently, the LF content resolution progressively grows as well

microlenses FOV will be further restricted (see Fig. 3a) and, consequently, the angular information of the system will be narrowed.

The angular information is chosen to grow from the central to the border pixels in each MI due to two essential reasons: i) the central angular direction is usually the perspective the shooter will point towards when capturing the LF image; and ii) pixels at the MI border are usually more affected by optical and geometric distortions than the central pixels. As an illustrative example, Fig. 3c shows the selection of the three subsets of pixels with different angular information from each MI to build a FOV scalable data format. For the base layer in Fig. 3c (bottom), only a set with central angular information is gathered. For the enhancement layers 1 and 2 in Fig. 3c (respectively, middle and top), the samples progressively contain wider angular information (from the center to the borders).

Due to the nature of the LF imaging technology, where angular (θ,φ) and spatial (x,y) information is spatially arranged in a 2D image (i.e., the LF image), the increased angular information in each higher FOV scalable layer implies also an increase in the mega-ray[3] resolution of the LF content in the layer. This means that resolution scalability is inherently associated to the FOV scalability (see Fig. 3c).

It is also important to notice that the FOV scalable LF data format is always restricted by the optical setup used when acquiring the original LF content. For instance, the total amount of angular information that is available to be subsampled in each MI is controlled by:

[3] Mega-ray is a measure of light field data capture that corresponds to the number of rays that are captured by the image sensor. This is numerically given by the resolution of the LF camera image sensor.

- **Real aperture size** – As seen in Section II.B, the (real) main lens aperture used in the LF content acquisition controls the amount of light angular information that is admitted through the LF camera optical system and that is sampled by the MIs.
- **Distance b in the generalized (focused) LF camera** – As discussed in Section II.A, the distance b (Fig. 1a) controls the balance between angular and spatial resolution in the captured LF image. Then, the closer b is to the MLA focal length, f, the wider angular information is sampled by the MIs (Fig. 2).

B. Application of the FOV Scalability for Flexible Interaction

The great advantage of the proposed FOV scalable data format is the increased flexibility it gives to the authoring process. This means that the content creator is able to select the number of hierarchical layers and the size of the subset of pixels to be sampled for each layer as a part of the creative process.

With this format, it is possible to define new levels of scalability, for instance, in terms of the following rendering capabilities:

- **Changing perspective** – It is straightforward to see that narrowing the FOV of each MI will limit the angular information in lower scalable layers and, consequently, the number of different viewpoint perspectives that are possible to render. Therefore, the higher the layer is, the greater the number of available viewpoints is for the user's interaction.
- **Changing focus (refocusing)** – Refocusing can be seen as virtually translating the image plane of the LF camera to another plane in front or behind it. Briefly, narrowing the FOV of the MI in each scalable layer will result in fewer depth planes that are available for refocusing. Hence, the higher the layer is, the richer the refocusing range is for the user's interaction.
- **Varying depth-of-field** – Increasing or decreasing the depth-of-field in LF images simply means to define larger or smaller (discrete) numbers of depth planes to be in focus simultaneously. Similarly to refocusing, limiting the MI angular information in each scalable layer will also limit the number of planes that are available to be in focus. Therefore, the higher the layer is, the deeper is the depth-of-field that can be selected during the user's interaction.

Therefore, the author can decide which perspective(s) and depth plane(s) need to be in focus when presenting the content in each of the hierarchical layers. Depending on his/her decision, narrower or wider angular information needs to be gathered for these layers.

C. LF Data Organization for ROI Coding

Another advantage of the proposed FOV scalable data format is the ability to enable easy integration of ROI coding [40], [43]. ROI coding can be an important functionally, especially in limited network channels [43], in applications scenarios where some portions of the visualized content are of higher importance than others. In the proposed FOV scalable data organization, this functionality would allow further flexibility in the bitstream for supporting the new interactive manipulations capabilities in the LF visualization.

For instance, for an LF image with very large resolution, the size of the compressed bitstream may be still considerably big

(a) (b)

Fig. 4 Illustrative examples for gathering three hierarchical layers from the base layer (bottom) to the last enhancement layer (top) to enable the ROI functionality. The amount of information gathered in each layer is depicted with proportional sizes: (a) The last ROI enhancement layer (top) considers the same ROI as the previous layer, but with wider FOV; (b) The last ROI enhancement layer (bottom) considers the same FOV as the previous layer but with a larger ROI.

in some LF enhancement layers to be streamed efficiently. Thus, a solution would be to send in these layers only a portion of the image which is of the most interest (i.e., the ROI) with wider FOV. Therefore, the end-user receives a coarse version of the LF content in the base layer and, if the network conditions permit, he/she has the option of interactively refining a portion or portions of the coarse received LF content with the new manipulation functionalities (such as refocusing, changing perspective and depth-of-field) by decoding further enhancement layers.

Fig. 4 illustrates this concept for a hypothetical case in which three hierarchical layers are defined. In the base layer (bottom), a coarse version of the LF content is gathered with very restricted FOV. Following this, there is a great variety of options for defining the ROI enhancement layers. Two of these possibilities are depicted in Fig. 4a and b. In Fig. 4a, the highest ROI enhancement layer (top) considers the same ROI as the previous layer, but with wider FOV. Differently, in Fig. 4b, the highest ROI enhancement layer considers the same FOV as the previous layer but increases the ROI size. In both cases, a similar amount of texture information is gathered in the highest layer.

Additionally, Fig. 5 illustrates examples of refinements, in terms of FOV manipulation functionalities, which can be allowed by using the ROI enhancement layers defined in Fig. 4a. Fig. 5a shows a central view rendered from the coarse version of the LF content in the base layer. In this case, the amount of LF information that is coded and transmitted is

(a) (b) (c)

Fig. 5 Examples of refinements in terms of refocus and perspective manipulation functionalities allowed by using the ROI enhancement layers in Fig. 4a: (a) The central view rendered from the coarse version of the LF content available in the base layer; (b) The refinement in the plane of focus when overlaying the ROI enhancement layer 2, for focusing at the object on the man's hand; and (c) Changing the perspective (to the left) inside the ROI while fixing the non-ROI area in the central view.

about 6 times less compared to the complete three-layered scalable bitstream in Fig. 4a. However, it can be seen that this significant reduction in terms of bits comes at the expense of limited FOV manipulation functionalities. For instance, it is not possible for an end-user to adjust the focus at the object in the man's hand, since it is outside the refocusing range allowed in the base layer (Fig. 4a). Differently, Fig. 5b depicts the refinement in the plane of focus that becomes available when decoding the ROI enhancement layer 2.

Moreover, Fig. 5c illustrates a possible refinement in the rendered view perspective that becomes available in the ROI enhancement layer 2. In this case, the perspective is slightly changed inside the ROI (to the left) while fixing the non-ROI

area in the central view. It should be noticed that some blending inconsistencies may appear, in this case, where the ROI and non-ROI join, (e.g., in the man's beard and knee in Fig. 5c). A possible solution to this is to use an arbitrarily shaped ROI instead of a rectangular one. This solution will be considered in the future work.

IV. Proposed FOVS-LFC Architecture

The coding architecture adopted in the proposed FOVS-LFC solution is built upon a predictive and multi-layered approach, as depicted in Fig. 6.

A. Coding Flow

The coding information flow in the proposed FOVS-LFC architecture is presented in the following:

- **LF decimation** – As illustrated in Fig. 6a, the LF data is firstly decimated into several layers, where higher layers contain LF content with wider FOV. In this process, the content creator will select the number of hierarchical layers and the size of the subset of pixels to be sampled for each layer. The decision of having narrower or wider angular information in each hierarchical layer may be made, for example, targeting a set of particular application scenarios. The base layer contains a sub-sampled portion of the LF data, which can be used to render a 2D version of the content with limited interaction capabilities (narrow FOV, limited in focus planes, and shallow depth-of-field). As shown in Fig. 6b, this base layer is coded with a conventional HEVC intra encoder to provide backward compatibility with a state-of-the-art coding solution, and the reconstructed picture is used for coding the higher layers. Following the base layer, one or more enhancement layers (enhancement layers 1 to *N* in Fig. 6a) are defined to represent the necessary information to obtain more immersive LF visualization. Each higher enhancement layer picture contains progressively richer angular information, thus increasing the LF data manipulation flexibility. Finally, the last enhancement layer represents the additional information to support full LF visualization with maximum manipulation capabilities. Each enhancement layer is encoded with the proposed LF enhancement layer codec seen in Fig. 6b, which is based on the HEVC architecture and uses the following new and modified modules:
- **Direct IL prediction** – To improve the RD performance when coding an LF enhancement layer, a new direct IL prediction is proposed, as shown in Fig. 6b. This direct IL prediction aims at exploiting the redundancy between adjacent layers to implicitly derive an IL predictor block for encoding the current block in an LF enhancement layer picture. As a result, the decoder can simply use the same process for inferring the predictor block. To avoid further signaling, only an index is transmitted together with the coded residual information, which is used to distinguish the direct IL prediction from the conventional HEVC merge mode [31]. The process to derive the direct IL predictor is presented in Section V.A.
- **IL compensated prediction** – A new IL compensated prediction can also be used to further improve the LF enhancement layer coding efficiency by removing redundancy between adjacent layers. For this, an enhanced IL reference picture is constructed and used as a new reference picture when encoding the current LF enhancement layer picture. To construct this enhanced IL reference picture, an exemplar-based texture synthesis algorithm is used, which is presented in Section V.B. If this IL prediction mode is used, the residual information and an IL vector are coded and transmitted to the decoder side.
- **SS prediction** – Since the proposed FOV scalable data organization still presents high redundancy between adjacent MIs (or decimated MI texture samples), the SS prediction (see Section I.A.1), previously proposed by the authors [15], can also be used as an alternative prediction to exploit this existing MI redundancy and to improve coding efficiency within each

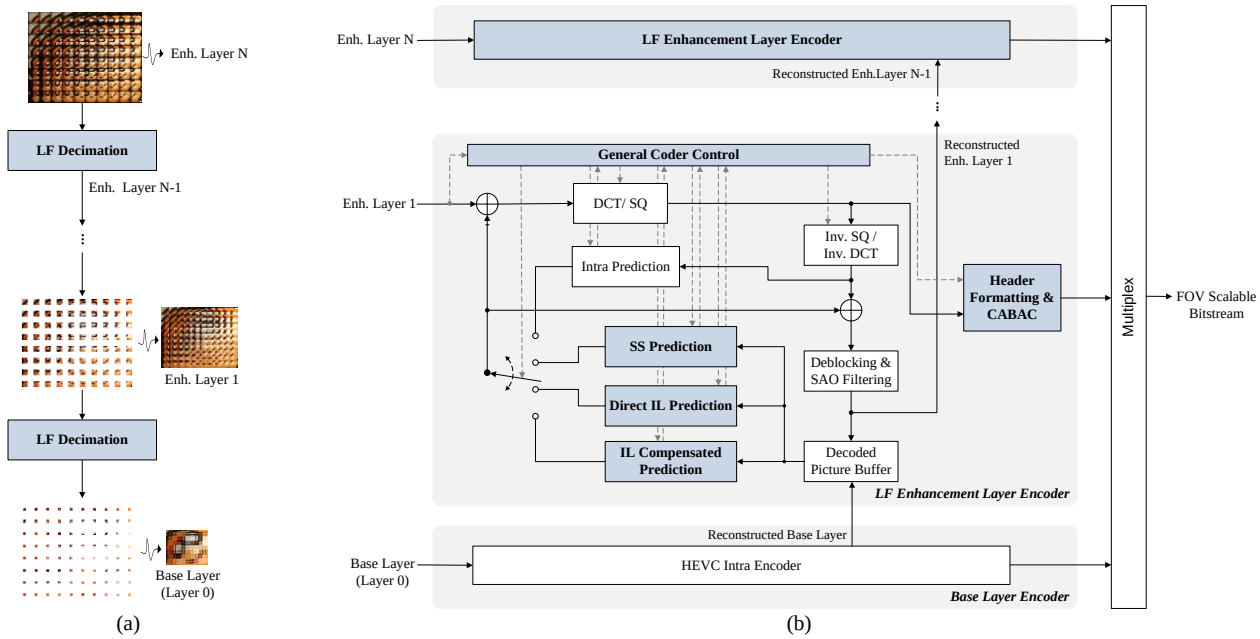


Fig. 6 The FOVS-LFC architecture (novel and modified blocks are highlighted in blue): (a) The LF decimation process to generate content for each hierarchical layer; (b) Proposed coding architecture in which one or more enhancement layers (from 1 to *N*-1) are coded with the proposed LF enhancement encoder

LF enhancement layer. As a result, the residual information and SS vector(s) are coded and sent to the decoder.

- **General coding control** –The decision among using conventional HEVC intra prediction, SS, direct IL and IL compensated prediction is made in a Rate-Distortion (RD) optimization manner [44].

Header formatting & Context-Adaptive Binary Arithmetic Coding (CABAC) – Additional syntax elements are carried through the high-level syntax bitstream to support FOV scalability. These are acquisition information (e.g., MI resolution and LF decimation information) and dependency information (for signaling the use of SS and IL prediction). Residual and prediction signaling are coded using CABAC.

B. Quality Scalability and ROI Coding Support

In addition to the FOV scalability, other functionalities are straightforwardly supported by the proposed FOVS-LFC solution, notably:

- **Quality Scalability** – Quality scalability can be achieved by quantizing the residual texture data in an LF enhancement layer with a smaller Quantization Parameter (QP) size relative to that used in the previous hierarchical layer. The QP values to be used in each layer can be adaptively adjusted to achieve the best tradeoff between quality and bitrate consumption.
- **ROI Coding** – In this case, the encoder can send, in different FOV enhancement layers, the information of the ROI with richer manipulation capabilities and better visual quality, at the expense of limited manipulation capabilities and potential lower visual quality in the background. For this, an adaptive quantization approach can also be used to proper assign reasonable bit allocation among different scalable layers.

V. EXEMPLAR-BASED IL CODING TOOLS

To achieve a high coding efficiency, the proposed FOVS-LFC solution relies on two exemplar-based IL coding tools detailed in this section: *i)* direct IL prediction, and *ii)* IL compensated prediction.

A. Direct IL Prediction

Similarly to template matching [45], the proposed direct IL prediction uses an implicit approach to avoid transmitting any information about the used predictor block. Hence, the decoder can simply use the same process for inferring the predictor block to be used for reconstructing the current block (using the decoded residual information).

The process to derive the IL predictor block can be divided in the following two steps.

1. Exemplar Block Derivation

In this first step, an exemplar-block is derived using the coded and reconstructed samples from a previous FOV scalable layer (referred to as the reference layer). This exemplar-block will then be used for implicitly finding a prediction to the current block, $I(\mathbf{x})$, at position $\mathbf{x} = (x, y)$ in the LF enhancement layer picture being coded (referred to as current layer).

Since a lower layer has narrower FOV and, consequently, a lower number of texture samples, it is firstly necessary to re-organize the texture information to align the MI samples from the reference layer according to the MI samples in the current layer. As a result, the reference layer is then represented as a picture with the same spatial resolution of the current layer picture and comprising a sparse set of known MI samples, as illustrated by the gray blocks in Fig. 7. This sparse picture is hereinafter referred to as sparse IL reference picture.

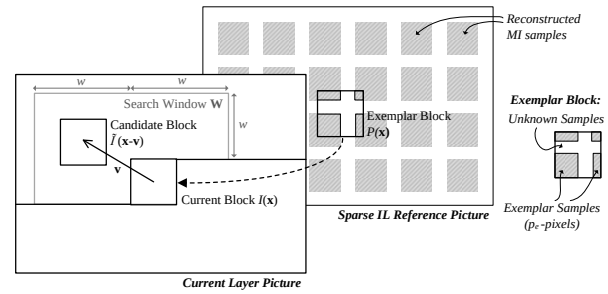


Fig. 7 Exemplar-based direct IL prediction, where an implicit predictor block for the current block is estimated by solving (1). In this process, the candidate block (within the search window **W** in the current layer picture**)** that 'best' agrees with the exemplar block is chosen as the predictor block.

As the output of this step, an exemplar block, $P(\mathbf{x})$, with the same size and co-located position to the current block, $I(\mathbf{x})$, is derived from the sparse IL reference picture (Fig. 7).

2. Direct IL Prediction Estimation

In this step, the exemplar block, $P(\mathbf{x})$, that was derived in the previous step is used as a template (similarly to template matching [45]) for estimating the 'best' predictor block to the current block, $I(\mathbf{x})$. For this, a matching algorithm is used to find the candidate block that 'best' agrees with $P(\mathbf{x})$ in the previously coded and reconstructed area of the current layer picture (Fig. 7). However, the 'best' candidate block is chosen by matching only the known samples of $P(\mathbf{x})$ (referred to as exemplar samples), since these are the only samples available at the decoding time.

Therefore, let $P(\mathbf{x})$ be a column vector containing the p-pixel samples of the exemplar block, where only the p_e-pixel exemplar samples (Fig. 7) are known at decoding time. Also, let $\tilde{I}(\mathbf{x}-\mathbf{v})$ be a column vector containing the p-pixel previously coded and reconstructed samples of a candidate predictor block in the current layer picture (Fig. 7). This candidate predictor block is displaced from $I(\mathbf{x})$ by the vector $\mathbf{v}$ (Fig. 7). Since P contains $(p - p_e)$ unknown samples, it can be modeled as $P = \mathbf{A}\,\tilde{I}$, where $\mathbf{A}$ is a binary mask in which only the corresponding known p_e sample positions are non-zero. Thus, $\mathbf{A}$ can be represented as a $p \times p$ binary diagonal matrix whose $(p - p_e)$ unknown diagonal samples are set to zero. Finally, since the mask $\mathbf{A}$ is known a priori, the 'best' candidate predictor block can be simply found by solving the matching algorithm in (1).

$$\min_{\mathbf{v},\tilde{I}(\mathbf{x}-\mathbf{v})\subset\mathbf{W}} \left\| P(\mathbf{x}) - \boldsymbol{A}\,\tilde{I}(\mathbf{x}-\mathbf{v}) \right\|_1 \qquad (1)$$

To keep the complexity low, the predictor block is searched inside a limited search window, $\mathbf{W}$, as depicted in Fig. 7 (i.e., $\tilde{I}(\mathbf{x}-\mathbf{v}) \subset \mathbf{W}$), and the ℓ_1-norm (or the sum of absolute differences), $\| \, \|_1$, is used as the matching criterion in (1).

B. IL Compensated Prediction

To further improve the LF enhancement layer coding efficiency, an IL compensated prediction is also proposed, which relies on an enhanced IL reference picture. This section describes the process for building the enhanced IL reference.

1. Input Information

The input information for this process is the coded and reconstructed samples from a reference layer picture that are properly aligned to the MI samples in the current layer picture. As seen in Section V.A.1, this re-arrangement results in the sparse IL reference picture (Fig. 7) that comprises a sparse set of known MI samples, as depicted in Fig. 8.

This sparse IL reference picture is used as the basis for building the enhanced IL reference picture. In this process, an exemplar-based texture synthesis algorithm is used to find a good estimation to fill in the unknown data in the sparse IL reference picture. This is clearly an ill-posed problem; however, it is still possible to obtain a realistic approximate solution by imposing additional constraints coming from the physics of the problem. This is done here by using the prior knowledge that neighboring MI samples present significant cross-correlation, and for this reason, it is likely to find the unknown region of a particular MI in an area of neighboring MIs. This problem is formalized as follows.

2. Problem Formulation

Firstly, the unknown pixels in the sparse IL reference picture are set to zero. Moreover, this sparse IL reference picture is divided into n-pixel non-overlapping patches, ϕ_s, to apply the texture synthesis algorithm (Fig. 8). Each patch is then given by n_s known samples – referred to as the support samples – and $(n - n_s)$ unknown samples to be synthesized (Fig. 8). Hence, each patch can be represented as the product of a texture column vector, ϕ_s, and a binary mask, $\mathbf{S}$, in which all but $(n - n_s)$ samples have value equal to one. The binary mask $\mathbf{S}$ is given by an $n \times n$ binary diagonal matrix with the respective $(n - n_s)$ unknown diagonal samples set to zero.

Accordingly, the goal of the texture synthesis algorithm is to find an n-pixel exemplar patch $\phi_e^{best}(\mathbf{x} - \boldsymbol{\omega})$ in the sparse IL reference picture – at position $(\mathbf{x} - \boldsymbol{\omega})$ – that 'best' agrees with the support samples of the patch $\phi_s(\mathbf{x})$ at position $\mathbf{x} = (x, y)$. To solve this, it can be assumed, without loss of generality, that the exemplar patch can be found in a neighborhood, $\boldsymbol{\Omega}$, of $\mathbf{x}$ (i.e., $\phi_e^{best}(\mathbf{x} - \boldsymbol{\omega}) \subset \boldsymbol{\Omega}$) comprising K neighbor MIs (i.e., $\boldsymbol{\Omega} = \{M_k\}_{k=1\ldots K}$ where M_k denotes an MI) as shown in Fig. 8. Additionally, it is assumed that a candidate exemplar patch ϕ_e comprises only n_e known pixels. Consequently, it can also be represented as the product of a texture column vector, ϕ_e, and an $n \times n$ binary diagonal matrix, $\mathbf{E}$, with $(n - n_e)$ diagonal samples set to zero.

Therefore, the best exemplar patch, ϕ_e^{best}, can then be found by solving the optimization problem in (2),

$$\min_{\phi_e(\mathbf{x}-\boldsymbol{\omega}) \subset \boldsymbol{\Omega}, \mathbf{A}} \left\| \mathbf{B} \cdot \left(\phi_s(\mathbf{x}) - \phi_e(\mathbf{x} - \boldsymbol{\omega}) \right) \right\|_1 + \lambda \times \| diag(\mathbf{I}_n - \mathbf{B}) \|_0 \qquad (2)$$

where $\mathbf{B}$ corresponds to a binary diagonal matrix that

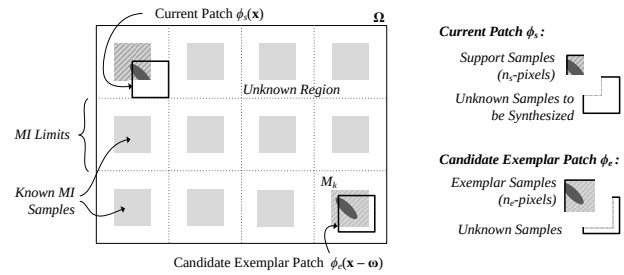


Fig. 8 Exemplar-based texture synthesis algorithm for building an enhanced IL reference picture. For each patch ϕ_s in the sparse IL reference picture, the 'best' candidate exemplar patch, ϕ_e^{best}, is derived by solving the optimization problem in (2).

represents the samples from ϕ_s and ϕ_e that overlap (i.e., $\mathbf{B} = \mathbf{S} \cdot \mathbf{E}$); $\mathbf{I}_n$ corresponds to an $n \times n$ identity matrix; $diag(\,)$ denotes a vector of the diagonal elements of a matrix; and $\| \, \|_1$ e $\| \, \|_0$ denote ℓ_1 and ℓ_0 norms, respectively.

The problem in (2) comprises a data-fitting term and a sparseness prior function, respectively. The former term tries to find the best match within the region where ϕ_s and ϕ_e overlap, while the latter term penalizes candidate patches whose n_e-pixel region is too small. In addition to this, since the border of the MIs typically exhibits high intensity variations (mainly due to the vignetting), a further constraint is imposed to the problem formulated in (2) to guarantee that these high frequency samples from the borders of an MI sample, $M_\mathrm{k} \subset \boldsymbol{\Omega}$, do not affect negatively the synthesized patterns, which is to solve the problem in (2), subjected to: $(\phi_e(\mathbf{x} - \boldsymbol{\omega}) \in M_k) \cap (\phi_e(\mathbf{x} + \boldsymbol{\omega}) \notin M_{m \neq k}) = \{\,\}$.

In the experimental results presented Section VI, the λ value (2) is selected empirically, and the patch size is selected to be a quarter of the size of an MI sample in the current layer.

The presented exemplar-based solution is chosen due to its simplicity and effectiveness for tackling the proposed problem. However, better solutions might still be formulated, for instance, by adding an edge-preserving regularizer in (2), or by using superpixel-based inpainting [46]. Moreover, although the angular information is limited in each subsampled MI in a lower layer, it is still possible to derive disparity or ray-space information to reconstruct the discarded 4D radiance samples at the receiver side. These solutions are left for future work.

3. Texture synthesis

Once the best patch ϕ_e^{best} is obtained by solving (2), the synthesized region is derived by simply copying the samples of the region defined by $\mathbf{E} \setminus \mathbf{B}$. This optimization process is iteratively repeated until all unknown samples are filled in or until the number of unknown samples stabilizes (i.e., the number of unknown samples remains the same between two iterations). Thus, at each iteration, the values of ϕ_e and $\mathbf{B}$ are updated from the values found in the previous iteration.

As an illustrative example, Fig. 9 shows a portion of the constructed enhanced IL reference picture for the LF enhancement layer 2 illustrated in Fig. 3c (top).

(a) (b)

Fig. 9 A portion of the enhanced IL reference picture built for the enhancement layer 2 in Fig. 3c: (a) the sparse IL reference picture; (b) the enhanced IL reference picture built by solving (2); and, (c) the difference to the original LF enhancement picture

VI. Performance Assessment

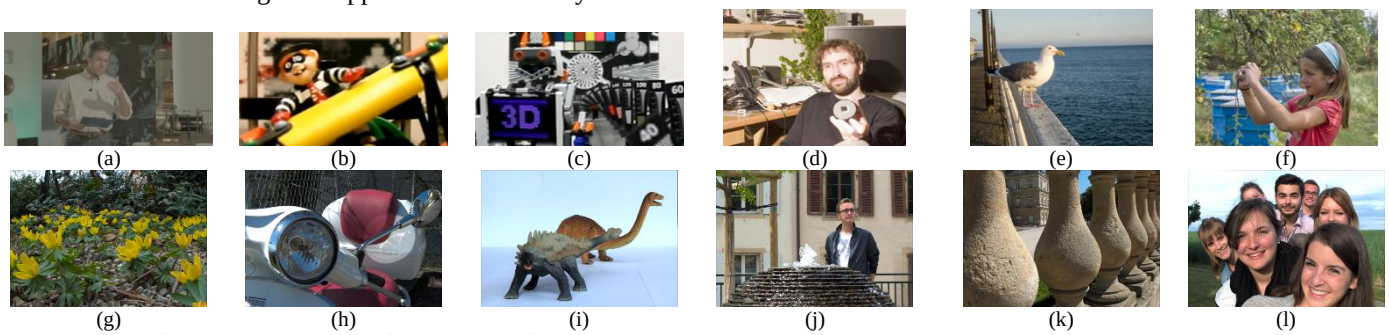
This section assesses the performance of the proposed FOVS-LFC solution. For this purpose, the test conditions are firstly introduced and, then, the obtained experimental results are presented and discussed.

A. Test Conditions

The performance assessment considered the following test conditions:

- **LF Test Images** – Twelve LF test images captured using different optical acquisition setups and with different scene characteristics are used, as shown in Fig. 10 and Table I. Before being coding, the raw LF images were pre-processed in order to: i) align and center the microlens grid to the pixels grid; ii) discard incomplete MIs (at the border of the LF image) iii) transform from hexagonal to rectangular microlens grid (only if necessary, see Table I); and iv) correcting color and gamma. As suggested in [47], only after this pre-process the LF image was convert to Y'CbCr 4:2:0 color format (8 bits) to avoid decreasing the visual quality after coding [47]. For LF test images whose calibration information was available (i.e., Fig. 10g to l), the Matlab LF Toolbox [48] was used for the pre-processing. It is worth noting that transforming from hexagonal to rectangular grid with the Matlab LF Toolbox results in an LF image with multiple black pixels at the MIs border. These black pixels were also discarded when pre-processing the LF test images. For the remaining LF test images (i.e., for the LF images in Fig. 10a to Fig. 10f, whose calibration information was not available), a DCT-based interpolation filter [49] was used for aligning and centering the microlens grid.
- **LF Decimation** – To generate the content for each hierarchical layer, *l*, in the FOVS-LFC solution, a central texture sample block with size $\left(2^{l+2}\times 2^{l+2}\right)$ is selected from each MI in the LF image to support FOV scalability. These squared texture sample blocks with a power of two size were here chosen to better fit into the CTU and PU partition patterns of HEVC [31]. However, the proposed scalable codec can be generalized for any texture sample block size and aspect ratio. The number of hierarchical layers varies for each LF test image and is given by $\lceil 1/2 \log_2 M \rceil$ for a squared MI with $M \times M$ pixels size (see Table I). Finally, the highest layer contains the entire LF image, whose resolution is shown in Table I. No ROI coding is considered in order to analyze the worst-case scenario in terms of the amount of texture information that is coded in each hierarchical layer.
- **Codec Software Implementation** – The MV-HEVC reference software version 12.0 [30] is used as the base software for implementing the proposed FOVS-LFC codec.
- **Coding Configuration** – Each LF test image is encoded using four different Quantization Parameter (QP) values: 22, 27, 32, and 37, according to the HEVC common test conditions defined in [50]. The same QP value is used for coding all hierarchical layers in order to analyze the worst-case scenario in terms of bitrate allocation. For both exemplar-based IL prediction and for the SS prediction, a search window with *w*= 128 (see Fig. 7) is adopted.
- **RD Evaluation** – For evaluating the overall RD performance of the proposed FOVS-LFC codec, two different objective quality metrics are considered, which are referred to as: *i)* Overall $\mathrm{PSNR_Y}$; and *ii)* Rendering-dependent $\mathrm{PSNR_Y}$. The overall $\mathrm{PSNR_Y}$ is calculated by taking the average luma Mean Squared Error ($\overline{\mathrm{MSE}}$) over the pictures in each hierarchical layer, and, then, converting it to the PSNR. Differently, the rendering-dependent $\mathrm{PSNR_Y}$ is measured in terms of the average luma PSNR calculated from a set of views rendered from the reconstructed LF content, similarly to the metrics proposed in [8]. To have a representative number of rendered views, a set of 11×11 views was rendered from uniformly distributed directional positions. For rendering the views from LF images captured using a focused LF camera setup, the algorithm proposed in [33] and referred to as Basic Rendering algorithm was used. In this case, the plane of focus was chosen to represent the case where the main object of the scene is in focus. For LF images captured using the unfocused LF camera setup, 11×11 VIs were extracted. The rate is calculated as the total number of bits needed for encoding all scalable layers divided by the number of pixels in the LF image given in Table I (bpp).

In addition, the performance of the proposed FOVS-LFC solution is compared to the following solutions:



(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k) (l)

Fig. 10 Example of a central view rendered from each light field test image: (a) *Demichelis Spark* [53], (b) *Plane and Toy* [53]; (c) *Robot 3D* [53]; (d) *Fredo* [54], (e) *Seagull* [54], (f) *Laura* [54], (g) *Flowers* [55], (h) *Vespa* [55], (i) *Ankylosaurus_&_Diplodocus_1* [55], (j) *Fountain_&_Vincent_2* [55], (k) *Stone_Pillars_Outside* [55], and (l) *Friends_1* [55]

TABLE I DESCRIPTION OF EACH LF TEST IMAGE IN FIG. 10

LF Image	Resolution* (mega-ray)	Camera Setup	MLA packing	MLA Pitch	MI Size* ($M \times M$)
(a)	2812×1520	Focused	Rectangular grid	300 µm	38×38
(b)	1904×1064	Focused	Rectangular grid	250 µm	28×28
(c)	1904×1064	Focused	Rectangular grid	250 µm	28×28
(d)	7104 ×5328	Focused	Rectangular grid	500 µm	74×74
(e)	7104 ×5328	Focused	Rectangular grid	500 µm	74×74
(f)	7104 ×5328	Focused	Rectangular grid	500 µm	74×74
(g)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11
(h)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11
(i)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11
(j)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11
(k)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11
(l)	6864×4774	Unfocused	Hexagonal grid	20 µm	11×11

*Values after pre-processing

- **HEVC (Single Layer)** – In this case, the entire LF raw data is encoded into a single layer with HEVC using the Main Still Picture profile [31].Since the proposed FOVS-LFC codec provides an HEVC-compliant base layer, this solution is used as the benchmark for non-scalable LF coding to compare the bit savings with the proposed scalable LF coding solution. Thus, it would correspond to the ideal RD performance if scalability was supported without any rate penalty.
- **FOVS-LFC (Simulcast)** – This solution corresponds to the benchmark for the simulcast case, where all pictures from each hierarchical layer are coded independently with HEVC intra coding. For this, the MV-HEVC reference software version 12.0 is used with "All Intra, Main" configuration [50].
- **FOVS-LFC (SS Simulcast)** – In this case, each picture from each hierarchical layer is coded with the FOVS-LFC codec but only enabling the SS prediction and conventional HEVC intra prediction. Hence, not only local spatial prediction is exploited (with conventional intra prediction) but also the non-local spatial correlation between neighboring MIs (with SS prediction [15]). Since each scalable layer is still coded independently (from each other) when using the SS prediction, the proposed FOVS-LFC (SS Simulcast) can be seen as an alternative simulcast coding solution.
- **VI-Based PVS (Low Delay P)** – This PVS-based solution represents a benchmark coding approach for providing scalability in the bitstream (as discussed in Section I.A.2). Similarly to what has been proposed in [25], [26], a PVS of VIs is coded using HEVC with the Low Delay P [50] configuration. However, to fairly compare this solution with the proposed FOVS-LFC solution, the QP values are kept the same for all VIs in the PVS. The VIs are scanned in outward clock-wise direction (referred to as spiral order) to form the PVS.
- **VI-Based PVS (Random Access)** – In this case, the PVS of VIs scanned in spiral order is encoded using HEVC using the Random Access [50] configuration. Similarly to the previous solution – VI-based PVS (Low Delay P), the QP values are kept the same for all VIs in the PVS.

For the FOVS-LFC (Proposed) solution, the base layer is encoded as an intra frame and the remaining LF enhancement layers are coded as inter B frames so as to allow bi-prediction.

B. Analysis of Coding Efficiency and FOV Scalability

Tables II and III present the RD performance of the proposed FOVS-LFC solution and the benchmark scalable solutions in terms of the Bjøntegaard Delta in PSNR (BD-PSNR) and bitrate (BD-BR) [51] with respect to (w.r.t) HEVC (Single Layer) for all test images in Fig. 10. For the PSNR results, Table II considers the $\overline{\text{MSE}}$ over the pictures in each hierarchical layer, while Table III considers the rendering-dependent PSNR metric (see Section VI.A).

As shown in Table II, the proposed FOVS-LFC solution presents better RD performance (0.38 dB or 7.92 % of bit savings in average) than the non-scalable HEVC (Single Layer) for most of the LF test images, independently of the used LF camera setup (i.e., focused versus unfocused). Moreover, significant coding gains of up to 3.87 dB or 82.66 % of bit savings can be achieved for LF images that present more homogeneous texture areas (e.g., for the LF image in Fig. 10i).

Considering the rendering-dependent PSNR metric in Table III, the proposed FOVS-LFC solution presents significant RD gains for focused LF images (0.72 dB or -12.66 % in average). For unfocused LF images, it is possible to support the FOV scalability with no performance loss in average. However, a comparison of the results in Tables II and III shows that the proposed FOVS-LFC is the solution with the best overall RD coding performance independently of the adopted quality metric. In terms of the rendering-dependent metric, the FOVS-LFC (Proposed) is able to achieve in average for all LF images (in terms of BD metrics): 2.59 dB (or -42.0 %) w.r.t the FOVS-LFC (Simulcast); 1.40 dB (or -28.18 %) w.r.t. FOVS-LFC (Simulcast SS); 2.83 dB (or -12.74 dB) w.r.t. PVS-based (Low Delay P); and 1.86 dB (or -10.54 %) w.r.t. PVS-based (Random Access).

Similar conclusions were observed when considering the objective quality metrics computed on all Y'CbCr components. For this reason, these results are omitted to avoid significantly increasing the size of the paper.

These results show that it is possible to support a FOV scalable bitstream with high coding efficiency for most of the LF test images (in comparison to the state-of-the-art HEVC). To complete this discussion, Section VI.D will analyze in more detail one of the worst-cases highlighted in Table II (i.e., for the LF image in Fig. 10c) where the overall RD performance of the proposed FOVS-LFC is worse than HEVC (Single Layer). This analysis will show that this RD performance penalty may be a negligible cost in some application scenarios considering the flexibility that is provided by the scalable bitstream in terms of LF interaction functionalities and bandwidth consumption.

As usually observed, the significantly better performance of the FOVS-LFC solution comes at the price of additional computational load compared to HEVC (Single Layer). Regarding the SS and the IL compensated predictions, the encoder and decoder computational complexity is conceptually the same as for HEVC inter prediction [52]. Concerning the direct IL prediction, the encoder complexity is similar to HEVC inter prediction, but the decoder complexity is increased since for coded blocks that use this type of prediction the decoder must estimate the direct IL predictor block. Regarding the exemplar-based IL texture synthesis algorithm, encoder and decoder complexities are similar, and the algorithm is employed only once for each LF enhancement layer. A careful analysis of the execution time for encoding

TABLE II RD PERFORMANCE OF THE PROPOSED FOVS-LFC SOLUTION AND THE BENCHMARK SOLUTIONS W.R.T. HEVC (SINGLE LAYER) (IN TERMS OF THE OVERALL $PSNR_Y$ METRIC AND TOTAL NUMBER OF BITS FOR THE SCALABLE BITSTREAM) FOR ALL LF TEST IMAGES IN FIG. 10

LF Image	FOVS-LFC (Proposed)		FOVS-LFC (Simulcast)		FOVS-LFC (SS Simulcast)		PVS-Based (Low Delay P)		PVS-Based (Random Access)	
	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]
(a)	0.56	-13.98	-1.99	61.64	-0.39	11.86	-3.98	162.97	-2.49	83.77
(b)	-0.08	1.12	-2.54	40.36	-0.92	15.38	-5.31	99.11	-3.39	62.76
(c)	**-1.27**	**17.34**	-3.11	43.97	-2.09	30.13	-7.38	128.07	-5.79	101.96
(d)	0.79	-13.86	-4.93	111.29	-1.30	27.49	-6.68	184.50	-6.19	153.00
(e)	1.57	-25.73	-4.49	108.06	-0.83	21.20	-3.55	86.02	-3.15	73.38
(f)	0.73	-11.47	-5.59	115.79	-2.14	46.77	-4.86	88.95	-4.02	73.60
(g)	-0.89	22.37	-2.28	63.70	-2.04	55.54	0.91	-20.39	2.00	-38.10
(h)	0.85	-22.92	-0.94	32.15	-0.01	0.86	2.49	-56.42	3.35	-65.74
(i)	**3.17**	**-82.66**	1.10	-45.34	2.41	-74.53	3.75	-86.20	4.03	-87.87
(j)	0.87	-18.61	-0.99	26.73	-0.09	2.51	2.96	-56.26	3.94	-66.25
(k)	-1.22	33.77	-2.23	65.07	-1.98	56.56	-1.02	39.29	0.47	-11.82
(l)	-0.56	19.56	-1.46	54.91	-1.18	43.10	-0.66	25.08	0.34	-10.15
Avg. Foc.	**0.38**	**-7.76**	**-3.78**	**80.19**	**-1.28**	**25.47**	**-5.29**	**124.94**	**-4.17**	**91.41**
Avg. Unf.	**0.37**	**-8.08**	**-1.13**	**32.87**	**-0.48**	**14.01**	**1.41**	**-25.82**	**2.36**	**-46.66**
Avg. All	**0.38**	**-7.92**	**-2.45**	**56.53**	**-0.88**	**19.74**	**-1.94**	**49.56**	**-0.91**	**22.38**

TABLE III RD PERFORMANCE OF THE PROPOSED FOVS-LFC CODEC AGAINST THE BENCHMARK SOLUTIONS W.R.T. HEVC (SINGLE LAYER) (IN TERMS OF THE RENDERING-DEPENDENT $PSNR_Y$ METRIC AND TOTAL NUMBER OF BITS FOR THE SCALABLE BITSTREAM) FOR ALL LF TEST IMAGES IN FIG. 10

LF Image	FOVS-LFC (Proposed)		FOVS-LFC (Simulcast)		FOVS-LFC (SS Simulcast)		PVS-Based (Low Delay P)		PVS-Based (Random Access)	
	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]	BD-PSNR [dB]	BD-BR [%]
(a)	0.33	-8.41	-3.20	129.44	-1.80	69.68	-6.40	341.90	-4.82	212.55
(b)	0.01	0.07	-2.79	50.44	-1.50	28.04	-6.18	122.29	-4.33	86.53
(c)	-1.04	14.80	-4.61	77.97	-3.80	64.99	-9.47	177.23	-7.98	152.27
(d)	1.58	-27.20	-4.90	125.64	-1.93	46.73	-8.02	250.54	-7.59	214.51
(e)	2.21	-35.90	-5.46	158.99	-2.27	65.09	-5.51	147.50	-5.20	139.95
(f)	1.24	-19.30	-6.11	143.73	-3.14	76.33	-6.14	115.39	-5.40	103.65
(g)	-0.85	21.53	-2.22	61.81	-2.06	56.78	0.43	-10.36	1.53	-30.63
(h)	0.47	-13.91	-1.40	51.78	-0.59	19.99	2.02	-49.76	2.87	-60.91
(i)	1.70	-64.44	-0.44	30.99	0.72	-35.75	2.60	-79.93	2.87	-81.77
(j)	0.32	-7.58	-1.60	45.97	-0.81	21.51	2.41	-49.41	3.38	-61.22
(k)	-1.18	33.40	-2.12	63.00	-1.96	57.04	-1.36	54.71	0.10	-2.30
(l)	-0.69	24.94	-1.57	61.24	-1.38	52.30	-1.04	42.58	-0.05	1.55
Avg. Foc.	**0.72**	**-12.66**	**-3.46**	**75.09**	**-1.19**	**25.17**	**-5.73**	**140.70**	**-4.65**	**105.05**
Avg. Unf.	**0.04**	**-1.01**	**-1.56**	**52.47**	**-1.01**	**28.65**	**0.84**	**-15.36**	**1.78**	**-39.21**
Avg. All	**0.34**	**-6.83**	**-2.51**	**63.78**	**-1.10**	**26.91**	**-2.44**	**62.67**	**-1.44**	**32.92**

and decoding each hierarchical layer using the proposed FOVS-LFC solution (according to the test conditions in Section VI.A) has shown that, in the worst case, the complexity load of the FOVS-LFC solution becomes larger than the HEVC (Single Layer) after coding/decoding two complete hierarchical layers. As will be seen in Section VI.D, scaling the complexity load may be advantageous since the user may not need to decode the complete bitstream to start visualizing and interacting with the LF content.

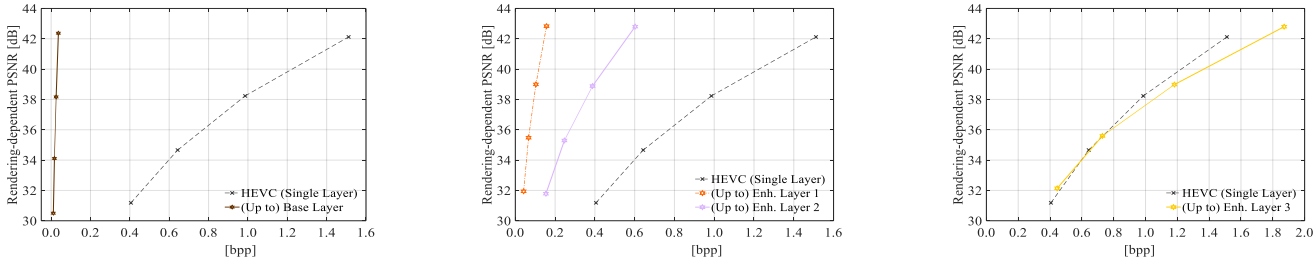
C. Analysis of the Exemplar-Based Coding Tools Efficiency

Comparing the results of the proposed FOVS-LFC solution with the FOVS-LFC (Simulcast) in Tables II and III, it can be seen that the FOVS-LFC (Proposed) outperforms this simulcast case with significant RD gains. These RD gains confirm the efficiency of the proposed FOVS-LFC in exploiting the redundancy in all domains, notably: *i)* local (using the HEVC intra prediction) and non-local (using the SS prediction) spatial redundancy within a single LF enhancement layer; and *ii)* the redundancy between the FOV scalable layers (using the proposed exemplar-based IL coding tools). Moreover, comparing these results with the FOVS-LFC (SS Simulcast), where the SS prediction is also available to be used in all LF enhancement layers, it can be seen that a considerable portion of the RD gains in the proposed FOVS-LFC solution is due to the proposed exemplar-based IL coding tools (i.e., the direct IL, and the IL compensated predictions).

D. RD Performance for Different Application Scenarios

To further discuss the usability of the proposed scalable architecture, the RD coding performance is here analyzed for three possible application scenarios, for which the use of LF imaging can be advantageous and likely to happen in the future. For each of the considered scenarios, the corresponding RD performance of proposed FOVS-LFC is compared to HEVC (Single Layer), in which scalability is not supported, to analyze the advantages of the proposed FOVS-LFC solution in terms of the flexibility enabled in the bitstream.

This analysis will consider one of the worst-case scenario highlighted in Table II (i.e., for the LF image *Robot 3D* in Fig. 10c), where the FOVS-LFC solution overall RD performance is worse than HEVC (Single Layer), so as to show the advantageous flexibility of the proposed coding architecture in terms of interaction capabilities and compression efficiency in each layer. For this, Fig. 11 shows the RD performance for the LF image *Robot 3D* (Fig. 10c), in terms of PSNR of a central rendered view and the corresponding bpp in each of the following scenarios:



(a) (b) (c)

Fig. 11 RD efficiency for *Robot 3D* regarding three different streaming scenarios for different user preferences and/or network conditions: (a) Scenario 1 – support of a 2D version of the LF content; (b) Scenario 2 – flexible support for LF applications with limited angular information; and (c) Scenario 3 – support for LF applications with full functionalities and angular information

(a) (b) (c) (d)

Fig. 12 Example of a portion from rendered views (for test image *Robot 3D* in Fig. 10c) when using the proposed FOVS-LFC solution (with QP 22). Each image corresponds to a different hierarchical layer: (a) base layer; (b) enhancement layer 1; (c) enhancement layer 2; and (d) enhancement layer 3. It is possible to observe how the larger angular information in higher layer allows having richer depth-of-field effects when manipulating the rendered views. This can be noticeable mainly by the blur at the out-of-focus areas.

- **Scenario 1 (no interaction capabilities)** – This first scenario supports the simplest LF visualization, in which the user only wants to visualize a simple 2D version of the LF content, possibly due to a limited bandwidth connection. In this case, the user would access (or start accessing) the LF content by decoding only the subset of the bitstream that corresponds to the base layer. As can be seen in Fig. 11a, the base layer corresponds to a very small percentage of the complete scalable bitstream and the RD efficiency of the proposed FOVS-LFC solution would greatly increase.
- **Scenario 2 (limited interaction capabilities)** – This scenario supports applications in which the user can select different viewpoints or can interact with the content with a larger degree of freedom. Additionally, it would also support 3D visualization of the LF content with horizontal and vertical motion parallax, but with narrower angular information. In this case, depending on the user's demand and the network conditions, a different number of scalable layers would have to be decoded. Consider, for instance, that for two different users it is necessary to decode the bitstream up to enhancement layer 1 (for user 1) and up to enhancement layer 2 (for user 2). The corresponding RD performance is illustrated in Fig. 11b. In both cases, it is still possible to significantly improve the coding efficiency compared to the HEVC (Single Layer). Fig. 12 illustrates a portion of the central views rendered from reconstructed frames in each scalable layer for the tested image *Robot 3D*. As expected, the richer angular information in higher layers (from Fig. 12a to Fig. 12d) allows the user to have larger degrees of freedom in manipulation (e.g., enabling a shallow depth-of-field). However, comparing Fig. 12b and Fig. 12c with Fig. 12d, it can be seen that in Fig. 12c the user may not need to decode the complete bitstream to have rendered views with similar perceived results to Fig. 12d.
- **Scenario 3 (full interaction capabilities)** – This scenario supports LF applications in which the user demands full interaction capabilities and visualization with maximum angular information. This corresponds to the lower bound case of the RD performance when FOV scalability is provided to a user without limitations in the network bandwidth. Fig. 11c shows that this is the only case where the scalable solution proposed FOVS-LFC presents worse RD performance compared to the HEVC (Single Layer). However, Table III shows that for most of the LF test images the proposed FOVS-LFC outperforms HEVC (Single Layer) with bit savings of 17.19 % in average. Hence, comparing this worst-case scenario with the average case, this bit saving loss for allowing the scalable coding architecture may be a considerably small cost to pay for the increased flexibility.

Moreover, differently from what happens in scalable LF solutions in the literature that rely on the accuracy of the depth estimation (as discussed in Section I.A.3), there is no significant discrepancy between the quality (in terms of PSNR) of a view rendered from the entire LF image coded in the latest layer (see Fig. 11c) and a view rendered from the LF content in a lower layer (see Fig. 11a and b).

To complete this analysis, Fig. 13 illustrates the needed bits for encoding each of the scalable layers using the proposed FOVS-LFC solution compared to the bits needed for the non-scalable HEVC (Single Layer) solution for all LF test images. From these results, it is possible to see that, in most cases, the rate cost to have the complete proposed FOVS-LFC solution does not exceed the cost of encoding the LF content in a single layer with HEVC.

E. Comparison against PVS-based Coding Approaches

It can be seen (Tables II and III) that the proposed FOVS-LFC solution architecture presents better overall RD performance than the PVS-based arrangement of VIs for both tested configurations (Low Delay P and Random Access).

Moreover, it can be seen that the RD performance of these PVS-based coding solutions varies significantly depending on the LF camera setup that is used for capturing the LF test

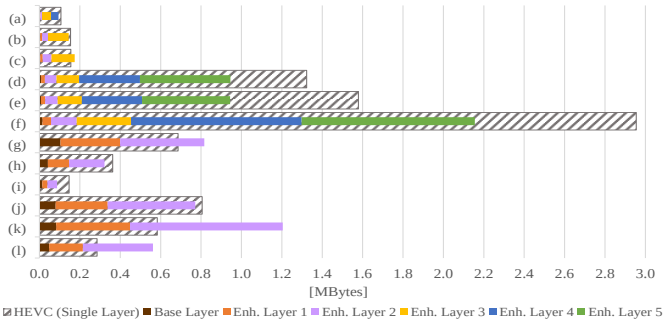


Fig. 13 Coding bits (in Mbytes) for each scalable layer using the FOVS-LFC solution w.r.t the non-scalable benchmark solution HEVC (Single Layer) for QP value 32

image. A significantly worse RD performance of PVS-based approach is observed for LF images captured using a focused LF camera setup. In these cases, the extracted VIs correspond to subsampled views with very low resolution and with significant aliasing artifacts (as discussed in Section I.A.2). Alternatives to deal with these aliased views are still possible, but would involve to work with a super-resolved LF image and/or to make use of depth information for improving the quality of these rendered views [32], [33]. In both cases, this would mean to increase the amount of information that is coded and transmitted to the decoder side. Assessing the RD performance of these alternative PVS-based coding approaches for LF images captured using a focused LF camera is out of the scope of this paper, but will be considered in future work. On the other hand, Tables II and III also show that the PVS-based arrangements is advantageous in terms of RD performance for coding LF images captured with unfocused LF cameras, outperforming the proposed FOVS-LFC solution. However, it is important to highlight that the proposed FOV scalability may be still advantageous in this case, in terms of the flexibility for supporting ROI enhancement layers, as discussed in Sections III.C and IV.B. Considering ROI enhancement layers, the proposed FOVS-LFC solution may also achieve a more competitive RD performance since less texture information is coded and transmitted in LF enhancement layers. This solution will be further studied in future work.

Comparing the results of the PVS-based approaches for the different coding configurations (i.e., Low Delay P versus Random Access) shows that it is possible to improve the RD performance of the PVS-based approach by selecting enhanced inter-view prediction structures. In fact, it has been shown in the literature that a 2D inter-view prediction structure [23], [27] may lead to further RD gains for LF images captured using a unfocused LF camera setup. However, it should be noticed that these solutions have not addressed the problem of coding aliased VIs yet (as discussed in Section I.A.2). These solutions were not evaluated in this paper due to difficulties to implement them for the very high number of VIs in the LF test images in Table I (in order to avoid making decisions and modifications that might not perfectly reflect the original solutions in [23], [27]).

VII. Final Remarks

This paper has proposed a flexible and efficient scalable coding framework for emerging LF applications that provides a novel type of scalability, here referred to as FOV scalability. The proposed FOVS-LFC solution comprises an HEVC backward compatible base layer and a flexible number of enhancement layers, which are coded using two new exemplar-based IL prediction schemes for improving the RD compression performance. The proposed scalable coding architecture satisfies many of the current requirements for the emerging image and video technologies, being easily adaptable to various user case scenarios demanding richer and immersive visualization. Experimental results have shown that the proposed FOVS-LFC solution can achieve significantly better RD performance compared to the tested benchmark scalable solutions, independently of the LF camera setup used for acquiring the content. Furthermore, the proposed scalable design provides flexibility in the rendering functionalities that emerge from LF imaging applications at no rate cost (in average) compared to the non-scalable benchmark HEVC. Additionally, it is shown that the compressed rendered views presented high quality in all hierarchical layers.

References

[1] M. Levoy and P. Hanrahan, “Light Field Rendering,” in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH ’96*, New Orleans, LA, US, 1996, pp. 31–42.

[2] R. Ng, “Digital Light Field Photography,” Ph.D Thesis, Stanford University, Stanford, CA, US, 2006.

[3] J. Arai, “Integral Three-Dimensional Television (FTV Seminar),” ISO/IEC JTC1/SC29/WG11 M34199, Sapporo, Japan, Jul. 2014.

[4] Raytrix, “Raytrix Website,” 2012. [Online]. Available: http://www.raytrix.de/. [Accessed: 07-Jul-2014].

[5] “Lytro Inc.,” 2012. [Online]. Available: https://www.lytro.com/. [Accessed: 07-Jul-2016].

[6] 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. 1060–75, Mar. 2015.

[7] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, “Advances in Three-Dimensional Integral Imaging: Sensing, Display, and Applications [Invited],” *Appl. Opt.*, vol. 52, no. 4, pp. 546–560, Feb. 2013.

[8] “JPEG Pleno Call for Proposals on Light Field Coding,” ISO/IEC JTC 1/ SC29/WG1 N74014, Geneva, Switzerland, Jan. 2017.

[9] K. Wegner and G. Lafruit, Eds., “Call for Immersive Visual Test Material,” ISO/IEC JTC1/SC29/WG11 N16766, Hobart, Australia, Apr. 2017.

[10] J. J. Sanchez-Hernandez, J. P. Garcia-Ortiz, V. Gonzalez-Ruiz, and D. Muller, “Interactive Streaming of Sequences of High Resolution JPEG2000 Images,” *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1829–1838, Oct. 2015.

[11] L. Toni, G. Cheung, and P. Frossard, “In-Network View Synthesis for Interactive Multiview Video Systems,” *IEEE Trans. Multimed.*, vol. 18, no. 5, pp. 852–864, May 2016.

[12] L. Toni and P. Frossard, “Optimal Representations for Adaptive Streaming in Interactive Multi-View Video Systems,” *IEEE Trans. Multimed.*, pp. 1–1, 2017.

[13] Prashant Ramanathan, M. Kalman, and B. Girod, “Rate-Distortion Optimized Interactive Light Field Streaming,” *IEEE Trans. Multimed.*, vol. 9, no. 4, pp. 813–825, Jun. 2007.

[14] C. Conti, P. Nunes, and L. D. Soares, “New HEVC Prediction Modes for 3D Holoscopic Video Coding,” in *2012 19th IEEE International Conference on Image Processing*, Orlando, FL, US, 2012, pp. 1325–1328.

[15] 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.

[16] 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.

[17] C. Conti, P. Nunes, and L. D. Soares, "HEVC-Based Light Field Image Coding with Bi-Predicted Self-Similarity Compensation," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, WA, US, 2016, pp. 1–4.

[18] L. F. R. Lucas *et al.*, "Locally Linear Embedding-Based Prediction for 3D Holoscopic Image Coding using HEVC," in *2014 Proc. of the 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 11–15.

[19] R. Monteiro *et al.*, "Light Field HEVC-Based Image Coding using Locally Linear Embedding and Self-Similarity Compensated Prediction," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, WA, US, 2016, pp. 1–4.

[20] 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.

[21] S. Shi, P. Gioia, and G. Madec, "Efficient Compression Method for Integral Images using Multi-View Video Coding," in *2011 18th IEEE International Conference on Image Processing*, Brussels, Belgium, 2011, pp. 137–140.

[22] J. Dick, H. Almeida, L. D. Soares, and P. Nunes, "3D Holoscopic Video Coding Using MVC," in *2011 IEEE EUROCON - International Conference on Computer as a Tool*, Lisbon, Portugal, 2011, pp. 1–4.

[23] G. Wang, W. Xiang, M. Pickering, and C. W. Chen, "Light Field Multi-View Video Coding With Two-Directional Parallel Inter-View Prediction," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5104–5117, Nov. 2016.

[24] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Full Parallax 3D Video Content Compression," in *Novel 3D Media Technologies*, A. Kondoz and T. Dagiuklas, Eds. New York, NY: Springer New York, 2015, pp. 49–70.

[25] R. Olsson, M. Sjostrom, and Y. Xu, "A Combined Pre-Processing and H.264-Compression Scheme for 3D Integral Images," in *2006 International Conference on Image Processing*, Atlanta, GA, US, 2006, pp. 513–516.

[26] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data Formats for High Efficiency Coding of Lytro-Illum Light Fields," in *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Orleans, France, 2015, pp. 494–497.

[27] D. Liu, L. Wang, L. Li, Zhiwei Xiong, Feng Wu, and Wenjun Zeng, "Pseudo-Sequence-Based Light Field Image Compression," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, WA, US, 2016, pp. 1–4.

[28] C. Perra and P. Assuncao, "High Efficiency Coding of Light Field Images based on Tiling and Pseudo-Temporal Data Arrangement," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Seattle, WA, US, 2016, pp. 1–4.

[29] 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.

[30] "MV-HEVC Reference Software HTM-12.0." [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-12.0/. [Accessed: 22-Dec-2014].

[31] 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. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[32] 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.

[33] T. Georgiev and A. Lumsdaine, "Focused Plenoptic Camera and Rendering," *J. Electron. Imaging*, vol. 19, no. 2, pp. 021106–021106, Apr. 2010.

[34] 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.

[35] A. Dricot, J. Jung, M. Cagnazzo, B. Pesquet, and F. Dufaux, "Integral Images Compression Scheme Based On View Extraction," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015, pp. 101–105.

[36] C. Choudhury and S. Chaudhuri, "Disparity Based Compression Technique for Focused Plenoptic Images," in *Proc. of the 2014 Indian Conference on Computer Vision Graphics and Image Processing - ICVGIP '14*, Bangalore, India, 2014, pp. 1–6.

[37] D. B. Graziosi, Z. Y. Alpaslan, and H. S. El-Ghoroury, "Depth Assisted Compression of Full Parallax Light Fields," in *Proc. SPIE 9391, Stereoscopic Displays and Applications XXVI*, San Francisco, CA, US, 2015, p. 93910Y.

[38] Y. Li, M. Sjöström, 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.

[39] Y. Piao and X. Yan, "Sub-Sampling Elemental Images for Integral Imaging Compression," in *2010 International Conference on Audio, Language and Image Processing*, Shanghai, China, 2010, pp. 1164–1168.

[40] A. Ebrahimi-Moghadam and S. Shirani, "Progressive scalable interactive region-of-interest image coding using vector quantization," *IEEE Trans. Multimed.*, vol. 7, no. 4, pp. 680–687, Aug. 2005.

[41] A. Criminisi, P. Perez, and K. Toyama, "Region Filling and Object Removal by Exemplar-Based Image Inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[42] A. Lumsdaine, T. G. Georgiev, and G. Chunev, "Spatial Analysis of Discrete Plenoptic Sampling," in *Proc. SPIE 8299, Digital Photography VIII*, Burlingame, CA, US, 2012, p. 829909.

[43] J. Park and B. Jeon, "Rate-Constrained Region of Interest Coding Using Adaptive Quantization in Transform Domain Wyner–Ziv Video Coding," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 685–699, Sep. 2016.

[44] G. J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, Nov-1998.

[45] T. Tan, C. Boon, and Y. Suzuki, "Intra Prediction by Template Matching," in *2006 International Conference on Image Processing*, Atlanta, GA, US, 2006, pp. 1693–1696.

[46] M. Schmeing and X. Jiang, "Faithful Disocclusion Filling in Depth Image Based Rendering Using Superpixel-Based Inpainting," *IEEE Trans. Multimed.*, vol. 17, no. 12, pp. 2160–2173, Dec. 2015.

[47] I. Viola, M. Rerabek, and T. Ebrahimi, "Comparison and Evaluation of Light Field Image Coding Approaches," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 7, pp. 1092–1106, Oct. 2017.

[48] D. Dansereau, "Light Field Toolbox v0.4," *MathWorks*, 25-Feb-2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/49683. [Accessed: 10-Feb-2016].

[49] H. Lv, R. Wang, X. Xie, H. Jia, and W. Gao, "A Comparison of Fractional-Pel Interpolation Filters in HEVC and H.264/AVC," in *2012 Visual Communications and Image Processing*, San Diego, CA, US, 2012, pp. 1–6.

[50] F. Bossen, "Common HM Test Conditions and Software Reference Configurations," JCTVC-L1100, Geneva, Switzerland, 2013.

[51] G. Bjøntegaard, "Calculation of Average PSNR Differences between RD Curves," VCEG-M33, Austin, TX, US, Apr. 2001.

[52] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC Complexity and Implementation Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1685–1696, Dec. 2012.

[53] "3D Holoscopic Sequences (Download Link)," 2013. [Online]. Available: http://3dholoscopicsequences.4shared.com/. [Accessed: 30-Oct-2016].

[54] T. Geogiev, "Todor Georgiev Gallery of Light Field Data." [Online]. Available: http://www.tgeorgiev.net/Gallery/. [Accessed: 17-Sep-2016].

[55] M. Řeřábek and T. Ebrahimi, "New Light Field Image Dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal, 2016.

Caroline Conti (S'11-M'18) received her B.Sc in Electrical Engineering from *Universidade de São Paulo (USP)*, Brazil, in 2010 and her Ph.D in Information Science and Technology from *Instituto Universitário de Lisboa* (ISCTE-IUL), Portugal, in 2017. Currently, she is a Postdoctoral Researcher with the Multimedia Signal Processing Group of the *Instituto de Telecomunicações*, Portugal. In addition, she is also an Invited Assistant Professor with ISCTE-IUL at the Information Science and Technology Department. Her research interests include immersive visual technologies and image and video processing and coding, including light field processing/coding. She has contributed more than 20 papers 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. In parallel, she acts as reviewer for various IEEE and EURASIP journals and conferences.

Luís Ducla Soares (S'98-M'04-SM'15) received the *Licenciatura* and Ph.D. degrees in Electrical and Computer Engineering from *Instituto Superior Técnico* (IST), *Universidade Técnica de Lisboa*, Portugal, in 1996 and 2004, respectively. Currently, he is a Senior Researcher with the Multimedia Signal Processing Group of the *Instituto de Telecomunicações*, Portugal. In addition, he is also an Assistant Professor with the *Instituto Universitário de Lisboa* (ISCTE-IUL), Portugal, at the Information Science and Technology Department. His research interests are centred around image and video coding/processing, including light field coding and processing as well as biometric recognition. He has contributed more than 65 papers 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 a member of the Editorial Board of the EURASIP Signal Processing (Elsevier) journal. In parallel, he acts as reviewer for several IEEE, IET and EURASIP journals and conferences. He is a Senior Member of the IEEE and EURASIP National Representative.

Paulo Nunes (S'98-M'07) graduated in Electrical and Computer Engineering from *Instituto Superior Técnico* (IST), *Universidade Técnica de Lisboa*, Portugal, in 1992 and he received the M.Sc. and Ph.D. degrees in Electrical and Computers Engineering from IST in 1996 and 2007, respectively. Currently, he is a Senior Researcher with the Multimedia Signal Processing Group of the *Instituto de Telecomunicações*, Portugal. In addition, he is also an Assistant Professor with the *Instituto Universitário de Lisboa* (ISCTE-IUL), Portugal, at the Information Science and Technology Department. His current research interests include 2D/3D image and video processing and coding, namely light field image and video processing and coding. He has contributed more than 60 papers to international journals and conferences in these areas (20 of which on light field coding). 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 reviewer for several IEEE, IET, EURASIP and SPIE conferences and journals, and as a member of the technical program and organizing committees of various international conferences.