6 Light Field Image Compression

Caroline Conti[1], Luís Ducla Soares[1], Paulo Nunes[1], Cristian Perra[2], Pedro Assunção[3], Mårten Sjöström[4], Yun Li[4], Roger Olsson[4], Ulf Jennehag[4]

[1]ISCTE – Instituto Universitário de Lisboa and Instituto de Telecomunicações, Lisbon, Portugal (email: {caroline.conti, lds, paulo.nunes}@lx.it.pt)

[2]Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy (email: cperra@ieee.org)

[3]Instituto de Telecomunicações and Politécnico de Leiria, Leiria, Portugal (email: amado@co.it.pt)

[4]Mid Sweden University, Sundsvall, Sweden (email: {marten.sjostrom, yun.li, roger.olsson, ulf.jennehag}@miun.se)

Abstract

Light field imaging based on a single-tier camera equipped with a microlens array has currently risen up as a practical and prospective approach for future visual applications and services. However, successfully deploying actual light field imaging applications and services will require identifying adequate coding solutions to efficiently handle the massive amount of data involved in these systems. In this context, this chapter presents some of the most recent light field image coding solutions that have been investigated. After a brief review of the current state-of-the-art in image coding formats for light field photography, an experimental study of the rate-distortion performance for different coding formats and architectures is presented. Then, aiming at enabling faster deployment of light field applications and services in the consumer market, a scalable light field coding solution that provides backward compatibility with legacy display devices (e.g., 2D, 3D stereo, and 3D multiview) is also presented. Furthermore, a light field coding scheme based on a sparse set of micro-images and associated block-wise disparity is also presented. This coding scheme is scalable with three layers such that the rendering can be performed with the sparse micro-image set, the reconstructed light field image, and the decoded light field image.

7.1 Introduction

Light field imaging based on a single-tier camera equipped with a microlens array (MLA) – simply referred to as Light Field (LF) in this chapter – has currently risen up as a practical and prospective approach for future visual applications and services. However, successfully deploying actual LF imaging applications and services will require identifying adequate coding solutions to efficiently handle the massive amount of data involved in these systems.

In this context, this chapter overviews some relevant LF image coding solutions that have been recently proposed in the literature. For this, the chapter starts reviewing the state-of-the-art in image coding formats for LF photography in Section 7.2. Moreover, since the choice of the used data format strongly influences the LF coding performance, a comprehensive analysis of the rate-distortion performance for different coding formats and different coding architectures applied to LF image coding is presented in Section 7.3. In addition to this, aiming at allowing faster deployment of LF applications and services in the consumer market, a scalable LF coding solution that provides backward compatibility with legacy display devices (e.g., 2D, 3D stereo, and 3D multiview) is presented in Section 7.4 This display scalable solution makes use of an efficient inter-layer prediction scheme that when combined with a spatial displacement compensated prediction is able to achieve, in most of the cases, better rate-distortion performance than the non-scalable HEVC solution.

Furthermore, a LF coding scheme based on a sparse set of Micro-Images (MIs) and associated block-wise disparity is presented in Section 7.5. This coding scheme is scalable with three layers such that the rendering can be performed with the sparse MI set, the reconstructed LF image, and the decoded LF image. Moreover, it is shown that this coding scheme improves considerably the coding efficiency with respect to HEVC Intra and is slightly better than the spatial displacement compensated prediction with multiple hypotheses.

7.2 Light Field Image Representation

Since the first approach proposed by Lippman [1] to capture light rays, continuous research and technological developments led to production of LF cameras (a.k.a. plenoptic cameras) that are now available in the consumer market and also for research and scientific applications. Such cameras are mainly characterized by their ability to record not only the light intensity but also the directionality of light-rays that reach the camera. This is equivalent to sample the continuous plenoptic function in (6.1), which describes the intensity of light rays passing through any point at a 3D spatial location (x,y,z), i.e. the camera center, from any possible direction (θ,ϕ) with wavelength λ at any instant t.

$$P = P(x, y, z, \theta, \phi, \lambda, t) \tag{6.1}$$

For practical acquisition and representation of light fields, the high dimensionality of the plenoptic function is reduced by assuming that the optical spectrum is monochromatic and the light intensity does not change over the discrete acquisition time of each sample (i.e., a single shot that captures one image at each instant *t*). Moreover, the LF is not captured for all possible 3D positions in the scene space. Instead, only the light projected onto the 2D camera plane is recorded. This simplification turns the 7D plenoptic function into a 4D representation of light fields, which is commonly used by defining two parallel planes, the camera plane (*u,v*) and the image plane (*s,t*). In such 4D model, the light field *L*(*s,t,u,v*) defines the intensity of a light ray intersecting both planes [2]. This representation allows visualization of a light field as a (*u*,*v*) array of (*s*,*t*) images (i.e. different views or perspectives) or as a (*s*,*t*) array of (*u*,*v*) images (i.e. sub-aperture images of the whole captured scene [3]. Since in currently available LF cameras, 4D light fields are captured as a two-dimensional matrix of tiled 2D MIs, the latter is also the most common representation format used in many application areas and computational algorithms based on the information conveyed by light directionality.

However, such tiled representation may not enable simple and fast access to other type of implicit information embedded in a 4D light field, such as surface reflection and the 3D structure of objects in the scene, i.e., depth. For extracting and processing such type of information the Epipolar Plane Image (EPI) representation is in general more appropriate. An EPI representation of a 4D light field can be thought as a large set of views, where the viewpoints all lie in the common focal plane and the views are projected onto the same image plane *I*. If *P* is parameterized with coordinates (*s*,*t*) and *I* with coordinates (*x*,*y*), then by fixing a camera coordinate *t* and image plane coordinate *y*, the resulting cut in the (*x*,*s*) plane is the EPI image. The EPI structure captures 2D views from different viewpoints and encodes the depth information as the slope of line structures in the 2D (*x*,*s*) planes.

7.3 Light Field Image Coding Formats

Regardless of the representation format, LF images require a huge amount of data to capture and store the light intensity along with directional information. The number of samples that is necessary to capture the intensity and direction of light rays is much higher than the spatial resolution of conventional 2D images usually rendered in end user devices.

Due to the inherent redundancy of LF representation, this type of visual data can be easily compressed using conventional image/video coding methods. However, such redundancy is dependent on the data structure that is used in conjunction with each specific coding scheme. For instance, when using standard image/video encoders, which are not specifically tailored for images comprising a lot of MIs with sharp boundaries and highly redundant content, there is a mismatch between such input data structure and the coding units used in most standard compression algorithms. Standard image and video encoders have been used for this purpose, but optimal exploitation of the intrinsic redundancy of LF data requires specific pre-processing. For instance, the correlation between neighboring MIs was exploited in [4] while the correlation in sub-aperture images, using three-dimensional transforms was exploited in [5]. Another method based on preprocessing the raw LF in two steps was prosed in [6]. The first step consists in partitioning the raw LF in tiles of equal size and then, in the second step, these tiles are ordered as a pseudo-temporal sequence in order to adapt the data to subsequent HEVC temporal predictive coding. The results show that, by exploiting redundancies in the spatial and view angle domain, the HEVC encoding tools are more efficient than JPEG exploiting only spatial redundancies in the whole LF image. Another result of interest is the significant difference between R-D performance of the tile-based scheme and that of JPEG, which is quite large for high compression ratios (e.g., bpp = 0.1), but much lower for small compression ratios (e.g., bpp = 1). Such results indicate that the benefits of exploiting both the spatial and view angle correlations decrease as the compression ratio also decreases. Thus, for lower compression ratios, exploiting the data redundancy in the two dimensions may result in similar coding efficiency as exploiting redundancy only in the spatial dimension.

For other applications, where the full accuracy of the originally captured LF needs to be preserved, lossless encoding must be used for the entire representation data. This is required for applications with stringent accuracy requirements, such as medical imaging, computer vision for industry, microscopy, etc. For such purpose, LF lossless coding methods have been reported in the literature. For instance, in [7], Perra encodes the non-rectified lenslet image by exploiting the correlation between micro-images, like Henlin *et al*, in [8], where the proposed method encodes the sub-aperture images extracted from the rectified lenslet data, exploiting inter-image correlations by applying different predictors to regions of the same depth. An experimental study on lossless light-field coding using standard codecs is presented in [9], using pre-processing techniques to convert the LF data to a format that enables higher lossless compression performance of current standard encoders. The study analyses the use of two types of pre-processing techniques that increase the compression efficiency of standard lossless encoders, namely lenslet data rearrangement and color transformation.

7.3.1 Light field image coding using HEVC

This section presents a performance evaluation study of the coding efficiency attained by the standard High Efficiency Video Coding (HEVC) using different LF representation formats [10]. To this aim, a data set comprising twelve LF images was captured with the Lytro-Illum camera, which stores the data on LPR files (≈ 55 MBytes each). This is basically a container format comprising several types of data (the RAW image as captured by the sensor, a thumbnail in PNG format and system settings, amongst others). The RAW image itself is a 10 bits pack, in GRBG format, with a total resolution of 7728×5368. The RAW files were processed using the "Light Field Toolbox for Matlab", which allows to decode and rectify the captured information using the camera's specific calibration data, comprising a set of white images [11]. The main output of this process is a reconstructed LF corresponding to a 625×434 matrix of MIs, each one capturing the light coming from 15×15 different directions. The data set used in this study is characterized in Table 6.1.
Five data formats were defined to evaluate the standard HEVC coding efficiency, corresponding to different data structures of the same YUV LF. Three of these are organized as still images and encoded using the HEVC Still Image Profile. For the remaining two, LF images are decomposed into sequences of different views in a pseudo-video arrangement encoded using the "Low-delay B", "Low-delay P" and "Random Access" video coding configuration. The following formats were used for the HEVC Still Image Profile.

- **Light Field (Lenslet) –** This is the LF comprising a matrix of MIs obtained with Light Field Toolbox for Matlab, as described in section II. An example can be seen in Fig. 6.1.
- **All-views –** The LF data is rearranged by first extracting the different angular

Table Error! Reference source not found.**.1.** Data set used to evaluate the HEVC light field coding performance

	Light Field Image	Visual content
1	*Euro*	A 2e coin
2	*Bottles*	Bottles (1.5L) on a table
3	*Bottle caps*	Plastic bottle caps on a grey table
4	*Corridor*	Corridor in back light
5	*WhiteFlowers*	Large green leaves and few pink flowers
6	*RedFlowers*	Small red and white flowers
7	*Park*	A few cars at a park exit
8	*Garden*	Part of a garden with medium-sized trees
9	*TrashCans*	Large (≈ 1.80m) recycling containers
10	*Twobottles*	Two plastic bottles (1.5L)
11	*People*	Four adults at building entrance
12	*SkinSpots*	Dark spots (≈ 2mm) on white skin

views which are then placed side-by-side, as seen in Fig. 6.2.

- **Light field filled –** This is similar to Lenslet but the black corner pixels of each MI is filled by extending the left-neighbour pixels (Fig. 6.3).

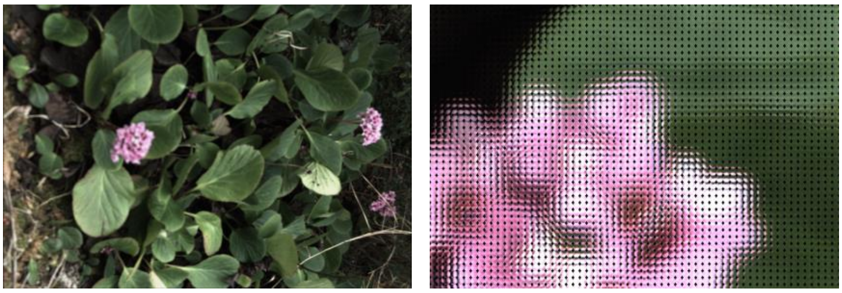


Fig. Error! Reference source not found.**.1.** Data set used to evaluate the HEVC light field coding performance

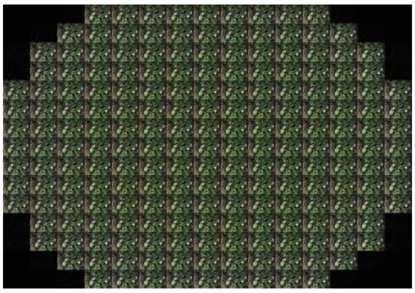


Fig. Error! Reference source not found.**.2.** Light field – all views

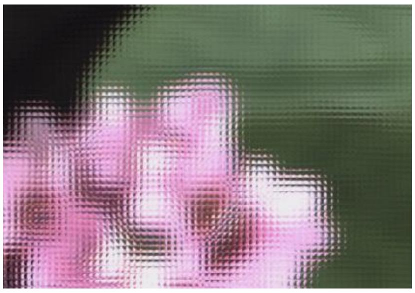


Fig. Error! Reference source not found.**.3.** Light field filled

The pseudo-video formats were obtained by using two different approaches to arrange sequences of views. Both of them result in a pseudo-video sequence such that adjacent views correspond to "temporally-adjacent" frames in order to obtain high inter-frame correlation. In general, this is observed when views have small view-angles, i.e., where disparity is smaller. The two pseudo-video formats used in this study are the following.

- **Raster –** The pseudo-video sequence is obtained by gathering the views from left to right and top-down, following the scan path shown in Fig. Error! Reference source not found.**.4** (left).
- **Spiral –** The pseudo-video sequence is obtained starting from the central view outwards, following the spiral scan path shown in Fig. Error! Reference source not found.**.4** (right).

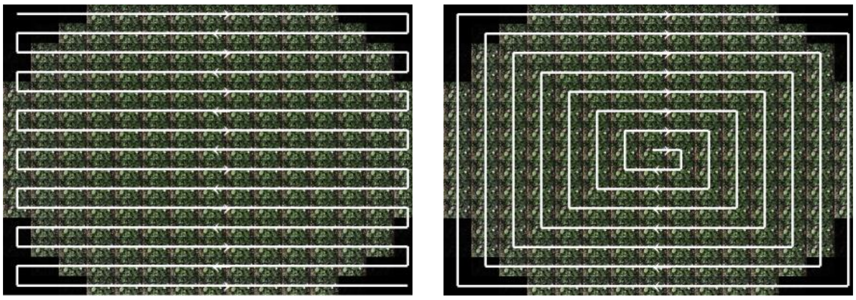


Fig. Error! Reference source not found.**.4.** Scan patterns to generate pseudo-video from *all views*: raster (left), spiral (right) [1]

For both Raster and Spiral pseudo-video formats, the HEVC configurations used for encoding the Light Field were the following: All Intra, Low Delay B, Low Delay P and Random Access. In the next section the performance of these coding configurations is evaluated, under test conditions adapted from [12].

7.3.1.1 Coding efficiency

The coding efficiency obtained from *Bottles*, *People* and *RedFlowers* is shown in Figs. Error! Reference source not found.**.5**, Error! Reference source not found.**.6**, and Error! Reference source not found.**.7** for the different configurations referred to above. From these Figures, it is quite obvious that different data formats have huge impact on the HEVC coding efficiency. For different arrangements of the LF data and HEVC coding configurations, the PSNR exhibits significant variations, which can be greater than 10dB at the same rate (bpp). It is worthwhile to note in these Figures that the pair (*data format, coding configuration*) does not correspond to a consistent relative coding performance for different visual content. Given the particular structure of the LF data, comprised of a matrix of tiny micro images with dark corners, a consistent worst performance would be expected from the *Light-Field* format. However, while this is true for *People* and *Bottles*, in the case of *RedFlowers* the *Spiral All-Intra* format is the one achieving the poorest performance. This is most likely due to the fact that the visual content of *RedFlowers* inside the MIs also contain further high frequency components corresponding to many small leaves of the flowers. Therefore, besides the high frequency nature of

[1] © 2017 IEEE. Reprinted, with permission, from [10].

the LF format itself, the coding efficiency is also greatly influenced by the characteristics of the visual content in each MI. On average, the *Bottles* and *People* LF images should not have so much high frequency content in each MI, which justifies the results shown in the Figures. Further research is necessary to find a valid threshold for the high frequency content of MIs, below which coding lenslet light field images might be better than intra coding of all views (i.e., All-Intra).

Figs. Error! Reference source not found..**5**, Error! Reference source not found..**6**, and Error! Reference source not found..**7** also show a detailed zoom of the lowest rates between 0 bpp and 0.03 bpp. For the pseudo-video formats, one can observe that very low rates are obtained for acceptable levels of PSNR. In this operational region, pseudo-video coding produces very similar results for all data formats, due to the use of high quantization parameters, which contribute to vanish the small differences between adjacent views.

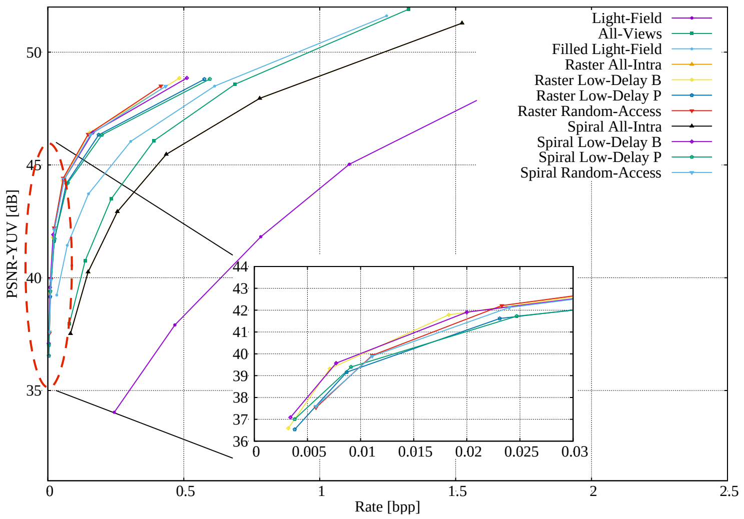


Fig. Error! Reference source not found..**5.** HEVC efficiency for LF image *Bottles*[2]

[2] © 2017 IEEE. Reprinted, with permission, from [10].

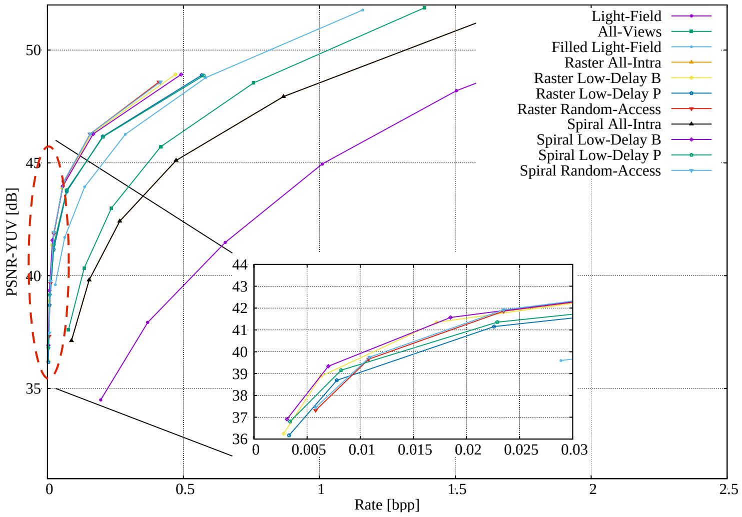


Fig. Error! Reference source not found.**.6.** HEVC efficiency for LF image *People*[3]

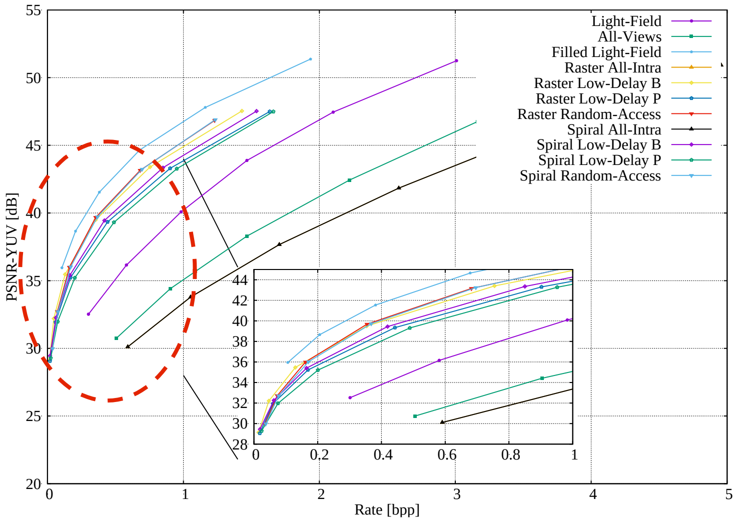


Fig. Error! Reference source not found.**.7.** HEVC efficiency for LF image *RedFlowers*[4]

[3] © 2017 IEEE. Reprinted, with permission, from [10].

[4] © 2017 IEEE. Reprinted, with permission, from [10].

In Figs. Error! Reference source not found.**.5**, Error! Reference source not found.**.6**, and Error! Reference source not found.**.7**, it is clear that organizing the LF data as pseudo video sequences provide much better performance than still images, as expected. This can also be seen in Table 6.2, where the coding efficiency

Table Error! Reference source not found.**.2** RD coding performance comparison for the et of LF images in Table Error! Reference source not found.**.1**

	Light Field (Lenslet)		AllViews		LF Filled		Raster Video (LowDelay B)		Spiral Video LowDelay B)	
Sequences	**bpp**	**PSNR**	**bpp**	**PSNR**	**bpp**	**PSNR**	**bpp**	**PSNR**	**bpp**	**PSNR**
					QP = 12					
BottleCaps	3,003	51,03	1,556	51,53	1,732	50,97	0,787	48,00	0,814	48,03
Bottles	2,406	51,58	1,326	51,89	1,246	51,61	0,484	48,85	0,511	48,85
Corridor	2,524	51,62	1,699	51,74	1,506	51,62	0,637	48,57	0,670	48,59
Euro	2,868	51,37	2,254	51,43	1,888	51,41	0,843	48,00	0,906	48,02
Garden	2,688	51,65	2,372	51,99	1,523	51,61	0,902	48,32	0,970	48,33
Park	2,967	51,31	2,155	51,60	1,826	51,35	0,818	48,03	0,891	48,02
People	2,263	51,69	1,388	51,88	1,160	51,77	0,471	48,92	0,492	48,92
RedFlowers	3,009	51,25	4,286	51,43	1,935	51,37	1,431	47,54	1,538	47,53
SkinSpots	3,315	50,94	2,225	51,34	2,074	51,04	0,895	47,60	0,945	47,63
TrashCans	3,394	51,08	1,518	51,60	2,477	50,99	0,750	48,21	0,750	48,22
TwoBottles	2,811	51,24	1,382	51,74	1,391	51,33	0,584	48,42	0,610	48,45
WhiteFlowers	3,250	51,06	3,248	51,25	2,129	51,13	1,167	47,60	1,252	47,62
					QP = 37					
BottleCaps	0,309	32,99	0,020	40,43	0,022	39,30	0,002	39,65	0,002	39,80
Bottles	0,244	34,03	0,079	37,95	0,033	39,23	0,003	36,59	0,003	37,09
Corridor	0,239	33,92	0,094	36,20	0,045	37,79	0,004	34,82	0,005	35,43
Euro	0,197	33,25	0,075	34,77	0,054	36,74	0,005	33,57	0,006	33,80
Garden	0,249	33,86	0,141	34,65	0,048	37,89	0,004	33,40	0,004	33,78
Park	0,352	33,16	0,117	35,69	0,039	36,95	0,004	34,27	0,005	34,83
People	0,196	34,49	0,077	37,60	0,029	39,60	0,003	36,25	0,003	36,91
RedFlowers	0,302	32,52	0,506	30,74	0,106	35,95	0,014	29,14	0,019	29,45
SkinSpots	0,402	32,27	0,025	36,39	0,044	36,68	0,002	35,59	0,002	35,77
TrashCans	0,341	32,18	0,064	38,53	0,106	35,02	0,005	37,03	0,006	37,48
TwoBottles	0,382	34,13	0,043	38,96	0,018	40,31	0,002	37,72	0,002	38,18
WhiteFlowers	0,343	32,68	0,243	32,89	0,076	36,23	0,006	31,46	0,007	32,05

is shown for the whole set of LF images. Table 6.2 reports the results obtained with quantization parameter QP=12 and QP = 37. As expected from the results above, the pseudo-video formats (Raster and Spiral) achieve lower bitrates in comparison with the other formats using the Still Image Profile. The Raster and Spiral scan patterns produce very similar results, which suggests that either one can be used without significant differences in performance.

The results of this simulation study lead to the conclusion that high efficiency coding of LFs is not only dependent of the encoder configuration but also requires appropriate data re-arrangement in order to obtain the best performance. The same coding configuration produce quite different results when using the same input data arranged in a different format. There are also intrinsic signal characteristics of each microlens, such as the amount of high frequency content, that influence the relative coding performance of the various methods. Further research is necessary to find the best LF pre-processing algorithms that are capable of guaranteeing a consistent relative performance across all coding configurations, for any type of content.

7.4 Scalable Light Field Coding for Backward Display Compatibility

In addition to the challenge of proposing efficient coding solutions for handling the huge amount of data involved in LF application systems, another important issue when trying to deliver LF content to end-users is to provide backward compatibility with existing legacy receivers (either 2D, or current stereo or multiview). Dealing with this specific concern is an essential requirement for enabling faster deployment of new LF imaging application services in the consumer market. For enabling this, an efficient scalable LF coding approach is then desirable where by decoding only the adequate subsets of the scalable stream, 2D or 3D compatible video decoders can present an appropriate version of the LF content. Regarding the scalable coding solution, although simulcast is a possible approach, the bandwidth consumption may not be acceptable, thus demanding a more efficient scalable coding solution.

In this context, a display scalable architecture for LF coding is presented in Section 7.4.1 (as firstly proposed in [13]) using a three hierarchical layer approach so as to accommodate from the end-user who wants to have a simple 2D version of the LF content to be visualized in a conventional 2D display; to the end-user who wants have a more immersive and interactive visualization by using a more advanced LF display technology, such as an integral imaging display [14–17] or a head mounted display for augmented and virtual reality [18, 19]. As discussed in Section 7.4.2, a pre-processing is necessary for generating the content for each hierarchical layer before coding. Based on this hierarchical coding architecture, Section 7.4.3 presents an Light Field (LF) enhancement codec to efficiently encode the LF content in the highest layer [20]. Finally, Section 7.4.4 performs the evaluation of the presented display scalable codec.

7.4.1 Display Scalable Coding Architecture

A Display Scalable Architecture for Light Field Coding (DS-LFC) with a three-layer approach is used here as illustrated in Fig. 6.8. As can be seen, each layer of this scalable coding architecture represents a different level of display scalability:

- **Base Layer (2D Layer)** – The base layer represents a single 2D view, which can be used to deliver a 2D version of the LF content to 2D displays devices. This 2D view is then coded with conventional HEVC [21] intra coder to provide backward compatibility with a state-of-the-art coding solution. Then, the reconstructed 2D view is used for coding the higher layers, as illustrated in Fig. 6.8.
- **First Enhancement Layer (Stereo or Multiview Layer)** – This layer represents the necessary information to obtain an additional view (representing a stereo pair) or various additional views (representing multiview content). This is to allow stereo and autostereoscopic devices to play versions of the same LF content. The content in this layer can be then encoded by using a standard stereo or multiview coding solution [22–25], and the reconstructed 2D views are then made available to be used for coding of the LF enhancement layer (Fig. 6.8). In this work, the multiview extension of HEVC, MV-HEVC [25], is adopted. With these solutions [22–25], inter-view prediction can be used to improve the coding efficiency between the base layer and the first enhancement

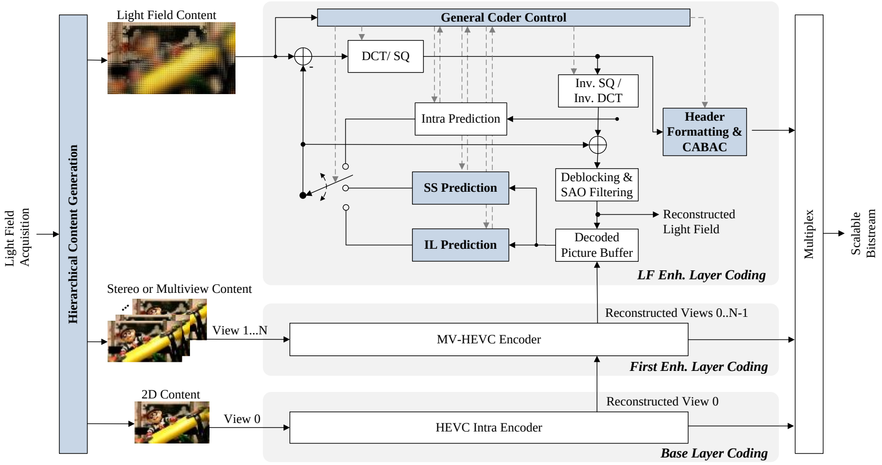


Fig. Error! Reference source not found.**.8** Scalable light field coding architecture using three hierarchical layers for backward display compatibility. The novel and modified blocks are highlighted in blue shaded blocks.

layer, as well as within the views in the first enhancement layer. However, it should be noticed that efficient prediction mechanisms between the base layer and the first enhancement layer and within the first enhancement layer are not addressed in this chapter since these cases have been extensively studied in the context of MVC [22], and in the 3D video coding extensions of the HEVC [25]. For a good review of these 3D video coding solutions, the reader can refer to [22–25] as well as Chapters 3 and 4.

- **Second Enhancement Layer (LF Enhancement Layer)** – This layer represents the additional data needed to support full LF display. The content in the LF enhancement layer is then encoded by using the LF enhancement coding solution presented in Section 7.4.3, as depicted in Fig. 6.8.

High compression efficiency is still an important requirement for the scalable coding architecture presented in this section. In this context, the scalable coding solution should be able to improve the Rate Distortion (RD) coding performance compared to independent compression of the three different layers (the simulcast case).

7.4.2 Hierarchical Content Generation

Generating 2D and 3D multiview content from LF content basically means producing various 2D views with different viewing angles. For this, a particular rendering algorithm needs to be chosen and some information about the acquisition process – such as the MI resolution and MLA structure (i.e., the array packing scheme and the microlens shape) – needs to be known at both encoder and decoder sides.

In the work presented in this section, the rendering algorithm proposed in [26] and referred to as Basic Rendering is adopted for this hierarchical content generation. The idea behind these algorithms is to combine suitable patches from each MI to properly compose a 2D view image. Then, as explained in [26], the process of generating a 2D view image can be controlled by the following two main parameters:

- **Patch Size** – It is possible to control the plane of focus in the generated 2D view image (i.e., which objects will appear in sharp focus) by choosing a suitable patch size to be extracted from each MI. Therefore, during a creative post-production process, a proper patch size will be selected for generating the content for the first two hierarchical layers. It is worth noting that this decision is limited to the available depth range in the captured LF image.
- **Patch Position** – By varying the relative position of the patch in the MI, it is possible to generate multiple 2D views with different horizontal and vertical viewing angles (i.e., different scene perspectives). It is also worthwhile to note that this choice is also made in a creative manner, and the number of views and

their corresponding positions may be based on a target type of display device that will be used for visualization.

In other words, there is a large degree of freedom when defining how to generate the content for the base and first enhancement layers. Therefore, the performance of the scalable coding solution shall be analyzed while taking into account the parameters that control this process.

7.4.3 Efficient LF Enhancement Layer Coding Solution

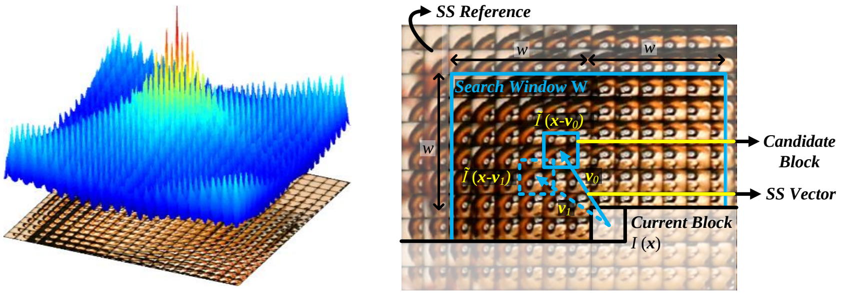
Since the lower layers of the proposed DS-LFC codec presented in Section 7.4.1 are based on the HEVC [21] standard (or on its extension for multiview coding MV-HEVC), the LF enhancement encoder proposed in this section is also based on the hybrid coding techniques of HEVC, as illustrated in **Fig. 6.8**, so as to modify as few aspects of the underlying architecture as possible. Notice that, although the LF enhancement layer encoder presented in **Fig. 6.8** targets LF still image coding, it can be also extended for scalable LF video coding by including also the HEVC inter-frame coding.

The main blocks of the proposed HEVC-based LF enhancement encoder (highlighted in Fig. 6.8) are explained in the following.

7.4.3.1 Self-Similarity (SS) Prediction

Since the LF content in the highest enhancement layer presents a significant
The SS prediction [27–29] (Fig. 6.8) is used to exploit the redundancy within the highest enhancement layer and to improve coding efficiency. As can be seen in Fig. Error! Reference source not found.**.9**a, a significant cross-correlation exists between neighbor MIs in the LF image captured with a LF camera.

Hence, the SS prediction is a spatial displacement compensated prediction [31] which makes use of a block-based matching algorithm to estimate the prediction block with the highest similarity (according to appropriate criteria) to the current block in the previously coded and reconstructed area of the current picture itself (the SS reference, as seen in Fig. Error! Reference source not found.**.9**b). This predictor block can be generated from a single candidate block [27, 28] or from a combination of two different candidate blocks [29, 31] (Fig. Error! Reference source not found.**.9**b). Hence, the relative position between the current and the 'best' candidate block(s) is signaled by one of two SS vector(s), $\boldsymbol{v}_i$, (Fig. Error! Reference source not found.**.9**b).



(a)

(b)

Fig. Error! Reference source not found.**.9** SS prediction: (a) inherent MI cross-correlation in a light field image neighborhood; and (b) SS estimation process (example of a second candidate block and SS vector for bi-prediction is shown in dashed blue line). (From [30][5].)

As a result of the SS prediction, the residual information and the SS vector(s) are coded and sent to the decoder.

7.4.3.2 Inter-Layer (IL) Prediction

An IL prediction mode can also be used to further improve the LF enhancement coding efficiency by removing redundancy between the LF content and its stereo or multiview version from the enhancement layer underneath.

For this, an Inter-Layer Reference (ILR) is constructed by using information from the lower layers. This ILR picture can be then used as new a reference frame for employing an IL compensated prediction (see Fig. 6.8) when encoding the LF image. To build an ILR picture, the following information is needed:

- **Set of 2D Views** – The set of reconstructed 2D views obtained by decoding the bitstream in the lower layers is available in the decoded picture buffer, as depicted in Fig. 6.8;
- **Acquisition Parameters** – These parameters comprise information from the LF capturing process (such as the MI resolution and the MLA structure) and also information from the 2D view generation process (i.e., size and position of the patches). As explained in Section 7.4.3.4, this information has to be conveyed along with the bitstream to be available at the decoding side.

Therefore, two steps are distinguished when generating an ILR picture, which are explained in the following.

[5] Reprinted from [30], Copyright (2017), with permission from Elsevier.

7.4.3.2.1 Patch Remapping

Although most of the LF information is discarded when rendering each view in the hierarchical layer generation block in Fig. 6.8 (see Section 7.4.2), it is still possible to re-organize the reconstructed view texture information into its original positions in the LF image. This is the purpose of the patch remapping step. The input for this step is the coded and reconstructed views from the two lower layers, as well as the acquisition parameters used for acquiring these views at the encoder side.

The patch remapping simply corresponds to an inverse process of the rendering algorithm used Section 7.4.2. More specifically, it corresponds to an inverse mapping (referred to here as remapping) of the patches from all rendered and reconstructed views to their original positions in the LF image, as illustrated in Fig. 6.10a. A template for the LF image assembles all patches, and the output is referred to as the sparse ILR picture, as seen in Fig. 6.11a.

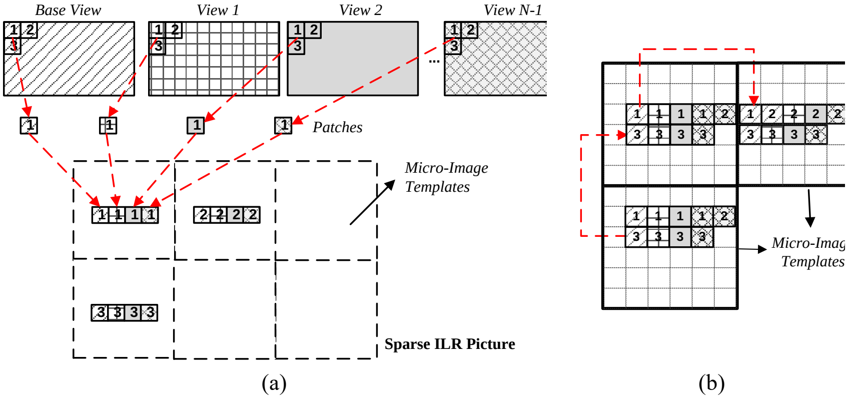


Fig. Error! Reference source not found.**.10** The process to generate an ILR picture to be used in the proposed IL prediction: (a) Patch remapping step; and (b) MI refilling step

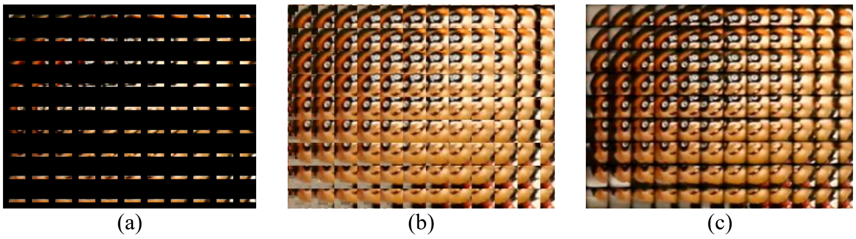


Fig. Error! Reference source not found.**.11** Illustrative example of a portion of an ILR built for the LF image Plane and Toy (frame 123): (a) the sparse ILR picture; (b) the corresponding complete ILR constructed using the MI refilling algorithm; and (c) the corresponding portion of the original LF image (which is coded in the LF enhancement layer)

7.4.3.2.2 MI Refilling

This step aims at emulating the significant cross-correlation existing between neighboring MIs so as to fill the holes in the sparse ILR picture (built in the previous step) as much as possible.

Since there is no information about the disparity/depth between objects in neighboring MIs, the disparity is defined in a patch-based manner, by using the patch size parameter that was used in the hierarchical layer generation block (see Section 7.4.2). An illustrative example of this process is shown in Fig. 6.10a for only three neighboring MIs in the sparse ILR picture. As can be seen, for each MI in the sparse ILR picture, an available set of pixels (see Fig. 6.10a) is copied to a suitable position in a neighboring MI that is shifted by the patch size. Additionally, the number of neighboring MIs where the patch may be copied to depends on the size of the MI and the patch size. Finally, the output of the process is the ILR picture (see Fig. 6.11b).

It is worthwhile to notice that there are still opportunities to enhance the proposed IL prediction (notably, the MI refilling step) and to enlarge the applicability of the proposed DS-LFC solution. A possibility is to incorporate supplementary data (such as depth, ray-space, and 3D model data) into the scalable bitstream. This solution will be further studied in future work.

7.4.3.3 Intra Prediction

HEVC Intra prediction is available as an alternative prediction when selecting the most efficient mode for encoding a CB in the LF enhancement layer (Fig. 6.8). The decision between the different available prediction modes is made in an Rate Distortion Optimization(RDO) manner [32] as in conventional HEVC [21].

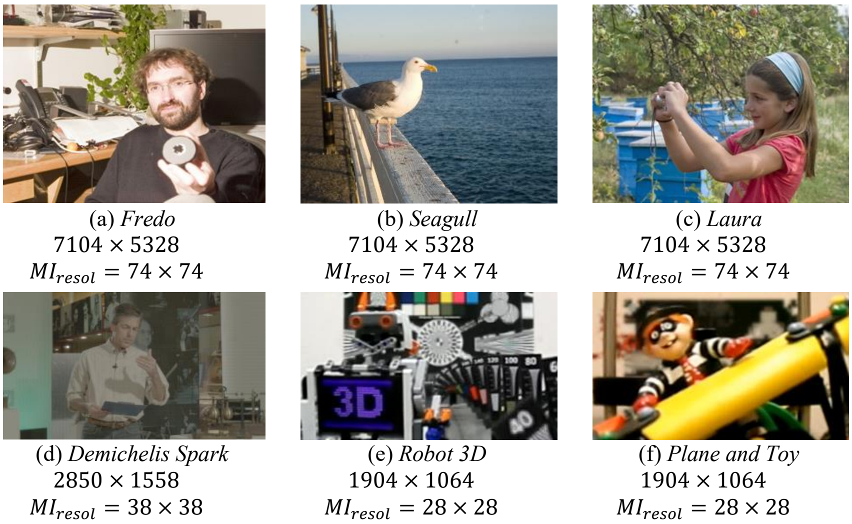
7.4.3.4 Header Formatting & CABAC

Additional high-level syntax elements are carried through the scalable bitstream to support this new type of scalability. These are basically: i) acquisition parameters that are used to generate the content for the lower layers and are also necessary to build the ILR picture (i.e., MI resolution, MLA structure, size and position of the patches); and ii) dependency information for signaling the use of the novel reference pictures (SS reference and ILR). Finally, residual and prediction mode signaling data are entropy coded using CABAC.

7.4.4 Performance Assessment

To evaluate the performance of the proposed DS-LFC codec, the following test conditions were considered:

- **Light field test images** – Six LF images with different spatial and MI resolutions are considered to achieve representative RD results. These are (see Fig. 6.12): *Fredo*, *Seagull*, *Laura*, *Demichelis Spark* (first frame of a sequence with identical name), *Robot 3D*, and *Plane and Toy* (frame number 123 from a sequence with identical name). The first three images are available in [33] and the remaining images in [34]. The original tested images were rectified to have all MIs with integer number of pixels, and they were then converted to the Y'CbCr 4:2:0 color format.



(a) *Fredo*
7104 × 5328
$MI_{resol} = 74 \times 74$

(b) *Seagull*
7104 × 5328
$MI_{resol} = 74 \times 74$

(c) *Laura*
7104 × 5328
$MI_{resol} = 74 \times 74$

(d) *Demichelis Spark*
2850 × 1558
$MI_{resol} = 38 \times 38$

(e) *Robot 3D*
1904 × 1064
$MI_{resol} = 28 \times 28$

(f) *Plane and Toy*
1904 × 1064
$MI_{resol} = 28 \times 28$

Fig. Error! Reference source not found.**.12** Example of a central view rendered from each light field test image (with the corresponding characteristics below each image)

- **Hierarchical Content Generation** - To generate the content for the 2D, stereo or multiview layers, the six LF test images were processed using the algorithm Basic Rendering [26] (Section 7.4.2). In this process, a set of 9×1 regularly spaced 2D views were generated – one for the base layer and the remainder for the first enhancement layer. Additionally, the patch size was chosen to represent the case where the main object of the scene is in focus. Based on the above decisions, the chosen patch sizes and positions for each LF test image are summarized in Table 6.3.
- **Codec Software Implementation** – For these tests, the reference software for the MV-HEVC extension version 12.0 [35] is used as the base software for implementing the proposed DS-LFC codec.
- **Coding Configuration** – The results are presented for four QP values (22, 27, 32, and 37). The same QP value was used for coding all hierarchical layers. In the proposed DS-LFC codec, all the views in the lower layers are independently encoded as intra frames. Notice that, other configurations for encoding the content in the first layer are still possible, notably, by enabling inter-view prediction (coding as P or B frames). However, due to the large number of possible test condition combinations, the following sections will focus on analyzing the influence of varying the parameters for generating the content for the lower layers in the performance of the proposed IL prediction. Following this, the LF enhancement layer is encoded as an inter B frame.
- **Search Strategy** – Considering both IL and SS prediction, a search range value of 128 is adopted for all tested LF images. The full search algorithm with the HEVC quarter-pixel accuracy is also used.

Table Error! Reference source not found.**.3** Test Conditions – Patch sizes and positions (in pixels) for generating content for the lower hierarchical layers using the DS-LFC solution (for each light field test image in Fig. Error! Reference source not found.**.12**)

Test Image	Patch Size (Focus Plane)	Patch Positions (View's Perspectives)
(a)	10	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
(b)	9	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
(c)	10	{(-24,0), (-18,0), (-12,0), (-6,0), (0,0), (6,0), (12,0), (18,0), (24,0)}
(d)	12	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}
(e)	4	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}
(f)	4	{(-8,0), (-6,0), (-4,0), (-2,0), (0,0), (2,0), (4,0), (6,0), (8,0)}

- **RD Evaluation** – For evaluating the RD performance of the proposed LF enhancement layer encoder, the distortion, in terms of PSNR, of the reconstructed LF image in the LF enhancement layer is considered. The rate is presented in bits per pixel (bpp), which is calculated as the total number of bits needed for encoding all scalable layers, divided by the number of pixels in the LF raw image. Therefore, the BD [36] results are presented in terms of the luma PSNR of the reconstructed LF image in the LF enhancement layer and the corresponding rate in terms of bpp values.
- **Additional Objective Quality Metrics** – Additionally, to analyze the performance in terms of the quality for views synthesized from the reconstructed content in the LF enhancement layer, the distortion is also measured in terms of average PSNR and SSIM values calculated for a set of 3×3 views rendered from viewpoint positions equally distributed in horizontal and vertical directions. This metric is referred to here as $\text{PSNR}_{3\times3\text{Views}}$ and $\text{SSIM}_{3\times3\text{Views}}$. These views are different than the views rendered for the lower layers (except for the central view). The standard deviation for each of these metrics is also used as a dispersion evaluation of the presented average values. For rendering the views, the same algorithm used for generating content for each hierarchical layer is used (i.e., Basic Rendering or Weighted Blending [26]).

The next subsections present and analyze the performance of the proposed DS-LFC solution and compare it to the following solutions:

- **DS-LFC (Simulcast)** – This scalable codec corresponds to the benchmark for the simulcast case, where the content from each hierarchical layer is coded independently with the MV-HEVC standard using "All Intra, Main" configuration [37].
- **DS-LFC (SS Simulcast)** – In this case, the content from the LF enhancement Layer was coded with the DS-LFC codec but only enabling the SS prediction and conventional HEVC Intra prediction (without IL prediction). Hence, not only local spatial prediction is exploited (with intra prediction) but also the

non-local spatial correlation between neighbor MIs (with SS prediction). Since when using the SS prediction each scalable layer is still coded independently (from each other), the proposed DS-LFC (SS) can be seen as an alternative simulcast coding solution.

- **HEVC (Single Layer)** – In this case, the entire LF image is encoded into a single layer with HEVC using the Main Still Picture profile [21]. Since the proposed DS-LFC codec provides an HEVC-compliant base layer, this solution is used as the benchmark for non-scalable LF coding, and the resulting bit savings are compared to the proposed scalable LF coding solution so as to analyze the cost (in terms of RD performance) of supporting display scalability in the bitstream.

7.4.4.1 Overall DS-LFC RD Performance

To assess the performance of the proposed DS-LFC codec, Table 6.4 presents the RD performance in terms of the Bjøntegaard Delta in PSNR (BD-PSNR) and rate (BD-BR) [36] regarding the benchmarks solutions for all test images in Fig. 6.12.

From these results, the following conclusions can be derived:

- **Comparison with simulcast cases** – The RD performance of the proposed DS-LFC is significantly better than the DS LFC (Simulcast) for all tested images, with average BD gains of 2.05 dB or 33.71 % of bit savings (see Table 6.4). The gains are much more expressive for test images with higher MI resolution, where the BD gain goes up to 3.00 dB with 44.56 % of bit savings (for *Seagull*). These gains show the efficiency of the predictive coding tools used in the LF enhancement encoder. Moreover, comparing the DS-LFC (Proposed) solution with the DS-LFC (SS Simulcast), improved RD performance can be attained by taking advantage of the redundancy in all domains (local and non-local spatial domain, and inter-layer domain), leading to average BD gains of 0.35 dB or -6.56 %.

Table Error! Reference source not found.**.4** BD-PSNR and BD-BR performance of the proposed DS-LFC codec against the benchmarks (for each test image)

Test Image	**DS LFC (Simulcast)**		**DS LFC (SS Simulcast)**		**HEVC (Single Layer)**	
	PSNR [dB]	**BR [%]**	**PSNR [dB]**	**BR [%]**	**PSNR [dB]**	**BR [%]**
(a)	2.85	-41.32	0.44	-8.52	2.08	-32.27
(b)	3.00	-44.56	0.43	-9.08	2.40	-37.90
(c)	2.59	-33.05	0.35	-5.86	1.32	-19.99
(d)	1.14	-29.04	0.26	-7.56	-0.19	6.80
(e)	1.18	-13.02	0.26	-3.12	-0.56	7.54
(f)	1.53	-20.58	0.34	-5.22	0.32	-5.13
Average	**2.05**	**-33.71**	**0.35**	**-6.56**	**0.90**	**-13.49**

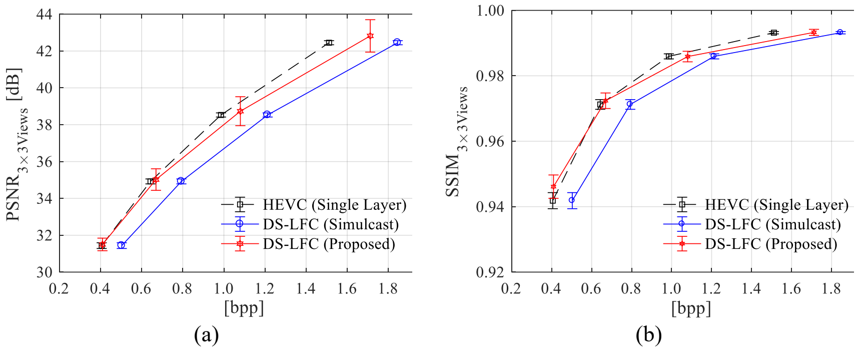
- **Comparison with HEVC (Single Layer)** – As shown in Table 6.4, the proposed DS-LFC solution presents better RD performance, in terms of average BD gains (0.90 dB and 13.49 %), than the non-scalable HEVC (Single Layer), showing that it is possible to support a display scalable bitstream with no additional bitrate cost. Moreover, for LF images with larger resolution and MI sizes, it is even possible to achieve significant better RD performance with the proposed DS-LFC (with BD gains of up to 2.40 dB and 37.90 % of bit savings). On the other hand, for some LF images with smaller resolutions and MI sizes, the scalability is allowed at a cost of some compression efficiency penalty (up to -0.56 dB and 7.54 % of penalty). However, it is important to notice that the worse RD performance of the proposed DS-LFC solution is, in this case, also due to the set of 9×1 views that are independently encoded as intra frames in the lower layers, instead of enabling the inter-view prediction to improve the RD performance.

7.4.4.2 Quality of Rendered Views

To assess the performance of the proposed scalable coding architecture regarding the quality of rendered views, the RD performance of the DS-LFC (Proposed) is here presented in terms of the $\mathrm{PSNR}_{3\times3\mathrm{Views}}$ and $\mathrm{SSIM}_{3\times3\mathrm{Views}}$ metrics and compared to the DS-LFC (Simulcast) and HEVC (Single Layer) solutions. The results are illustrated in Fig. 6.13 and Fig. 6.14, respectively, for the worst (i.e., for test image Robot 3D) and best case (i.e., for test image Seagull) in terms of DS-LFC (Proposed) RD coding gains.

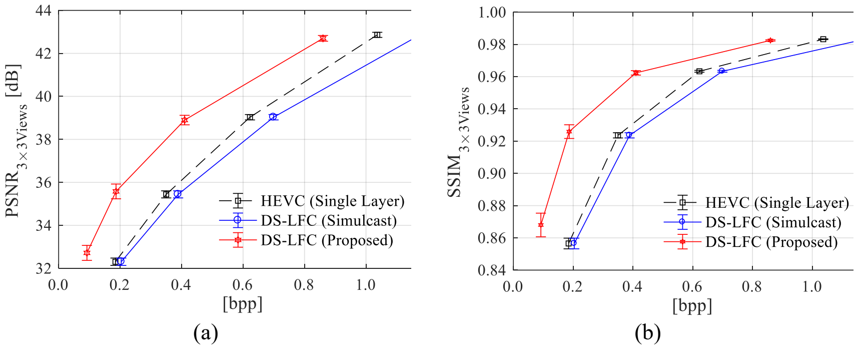
It was observed that there is a consistent relative RD performance gain using the three different quality metrics. In all cases, the proposed DS-LFC outperforms the simulcast cases with significant gains, showing the advantage of using the proposed IL prediction for improving the RD performance. In terms of the $\mathrm{PSNR}_{3\times3\mathrm{Views}}$ metric (Fig. 6.14a), the RD gains (using the BD metric [36]) of the DS-LFC (Proposed) solution go up to 2.79 dB or 14.82 % compared to DS-LFC (Simulcast) and 2.48 dB or 38.62 % with respect to HEVC (Single Layer). In the worst case (Fig. 6.14a), supporting a display scalable bitstream using the DS-LFC (Proposed) solution results in a RD performance penalty (using the BD metric [36]) of 0.37 dB or 4.89 % of bit saving loss.

Regarding the standard deviation values presented in Fig. 6.13 and Fig. 6.14, a more careful analysis of the $\mathrm{PSNR}_{3\times3\mathrm{Views}}$/$\mathrm{SSIM}_{3\times3\mathrm{Views}}$ results for each rendered views showed that views rendered from viewpoint positions near to the border of the MIs presented larger variation in PSNR/SSIM values. These variations are more significant in the case of *Demichelis Spark*, *Robot 3D* (see Fig. 6.13) and *Plane and Toy* mainly due to the increased vignetting that appears in these images, at the border of each MI.



(a) (b)

Fig. Error! Reference source not found.**.13** RD performance for a set of rendered views from image *Robot 3D* (Fig. Error! Reference source not found.**.12**e) in terms of: (a) $\text{PSNR}_{3\times3\text{Views}}$ versus bpp; and (b) $\text{SSIM}_{3\times3\text{Views}}$ versus bpp



(a) (b)

Fig. Error! Reference source not found.**.14** RD performance for a set of rendered views from image *Seagull* (Fig. Error! Reference source not found.**.12**b) in terms of: (a) $\text{PSNR}_{3\times3\text{Views}}$ versus bpp; and (b) $\text{SSIM}_{3\times3\text{Views}}$ versus bpp

7.5 Sparse Set of Micro-Lens Images and Disparities for an Efficient Scalable Coding of Light Field Images

The information in lightfield images has a high degree of correlation, as its elements are projections captured from a single scene out of different angles for many positions. In the previous sections of this chapter, this correlation has been modelled in different ways to enable efficient compression. In the present section, the correlation is described by introducing disparity maps in a similar way as depth maps. It is based on a number of articles that describe and evaluate compression of LF that use disparity maps [38, 39] and multi-hypothesis intra prediction [31, 40] for compression of LF images, both from focused as well as conventional LF cameras [41].

The use of disparity maps to describe the correlation between MIs is particularly suitable when the full LF image is produced by focused LF cameras. The reason is that each MI constitutes a small perspective view of the observed scene with a fair amount of information overlap. The disparities so constitute a shift of pixels between adjacent MIs. This pixel shift is also rather small for data coming from LF cameras, as the distances between the MIs are small and objects at intermediate and long distances from the camera.

The principal reason for using the disparity between MIs can also be found in other compression schemes. For example, in an earlier work [42] the scheme arranges light field images into a grid, images within the grid are recursively predicted from a few intra coded images. It was later improved by using homography transformations to describe the disparities between views in the light field [43].

The compression scheme using a sparse set of micro-lens images and disparities that is presented in this section consists of three parts:

- **Sparse set of MIs** – The MIs of the original LF image is decimated by selecting every *s* MI and so constitute a new LF image of sub-sampled MIs.
- **Disparity maps** – The disparity between adjacent MIs are described by a best value of pixel shift. These disparities so constitute two maps, one describing the horizontal, one the vertical pixel shifts.
- **Refinement by inter and intra prediction** – The two previous parts, sparse set of MIs and the disparity maps, enable the prediction of a LF image of full resolution. The third part contains a refinement to obtain a high quality LF image by predicting from this first LF image of full resolution.

7.5.1 Scalability

The three parts of the compression scheme constitutes a successive refinement of the final LF image reconstruction, and therefore is the basis for a scalability built into the compression scheme. The *first layer* includes a decimated image with a lower angular and spatial resolution. Image reconstruction from this layer can be useful for thumbnails or presentations on smaller displays and devices with lower computational power. The first layer so forms a scalability with respect to resolution. The *second layer* include additional information so that a LF image with full spatial and angular resolution can be reconstructed, although with a reduced image quality than the original. The second layer so forms a scalability with respect to quality. The *third layer* adds further information that enable a full resolution LF image with the highest possible images quality for the selected compression ratio.

7.5.2 Displacement Intra and Inter Prediction Scheme

Intra prediction schemes are efficient compression methods for images that contain correlated information. This is demonstrated in Section 7.4, in which self-similarity prediction for LF images is used. Likewise, the block-copying mode (BC) [44] was introduced into the HEVC codec in order to compress screen contents that contain plenty of correlated information. Both self-similarity and HEVC-BC are single-hypothesis intra predictions, i.e. they find a best block match in the already decoded part of the image and keeps the disparity vector to describe the block to be compressed.

The *displacement intra prediction scheme* was proposed and investigated in [31, 40]. It employs a multi-hypothesis intra prediction by subdividing the already decoded part of the image into two areas, from each of which a best candidate for block prediction is searched. These two blocks are candidates to be used as a reference for the intra prediction. There is also a third candidate. It is the mean of two blocks found by searching near the first two blocks. The best match of the three candidates is used for the intra prediction. The scheme works well both for LF images from focused as well as conventional LF cameras [41].

In the case of light field video coding, i.e. a sequence of LF images, the intra prediction scheme can also incorporate predictions from previous or future frames to search for good candidates. Thereby, the previous or future frames are loaded into the reference picture list, and the rate-distortion optimization of the codec selects the best prediction mode among the inter prediction, the displacement intra and the original HEVC intra. The combination of these modes provides more possible prediction candidates for an efficient compression. See Fig. 6.15.

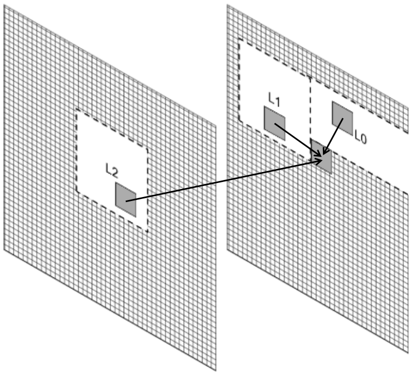


Fig. Error! Reference source not found.**.15** Multi-hypothesis prediction in displacement intra and inter prediction scheme. Each block may be predicted from previously decoded areas in the same frame (L0 and L1), and from previous frames (L2).

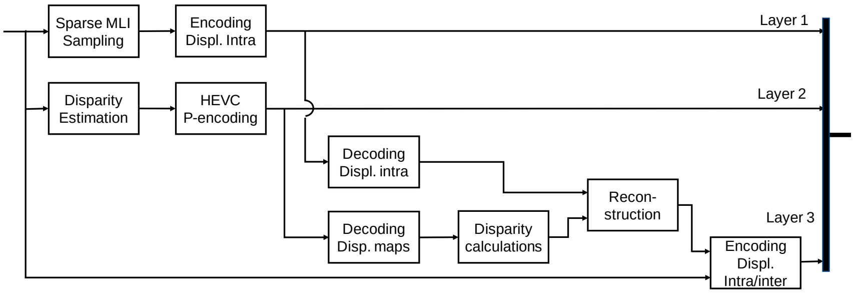


Fig. Error! Reference source not found.**.16** Schematic overview of decoding for sparse set of MIs and disparities. The LF image is encoded in three scalability layers. Layer 1 decimates the LF image to fewer number of MIs and is encoded by the displacement intra prediction scheme. Layer 2 estimates horizontal and vertical disparities between all adjacent MIs. The two disparity maps are encoded using HEVC video coder. Layer 3 uses the reconstruction of layer 1 and 2 as a reference in the displacement intra and inter coder.

7.5.3 Encoding

A schematic description of the encoding of LF images using the sparse set of micro-lens images and disparities is given in Fig. 6.16. The scheme is subdivided into three layers that constitute the basis of the scalability.

7.5.3.1 Sparse set of micro-lens images

The decimation of the full LF image into a sparse set of MIs is done by selecting every s MI. The input LF image $C(x, y, r, t)$ is described by $N \times M$ MIs with coordinates (x, y), each containing $N_t \times M_t$ pixels with coordinates (r, t). So, the sparse set of MIs is described by $C_s(x_s, y_s, r, t) = C(x \cdot s, y \cdot s, r, t)$, such that $x_s \in [1, N/s]$ and $y_s \in [1, M/s]$. See Fig. 6.17a.

The sparse set is itself an LF image with fewer MIs than the original, and therefore has a reduced resolution. The scheme is developed for LF images from focused LF cameras, which has a distribution of angular and spatial information throughout the MIs, which means that the decimation implies a reduction in both angular and spatial resolution.

Another consequence of the sparse set being an LF image is that it can be compressed with any codec developed of this kind of data. HEVC was employed in [39] but Displacement intra has a better compression performance for LF image data and is employed from now on. The compressed sparse set of MIs constitutes the first layer of the scalable codec. The encoder further includes a decoding of the first sparse set to assure that the other layers receives the same data as the decoder on the receiver side.

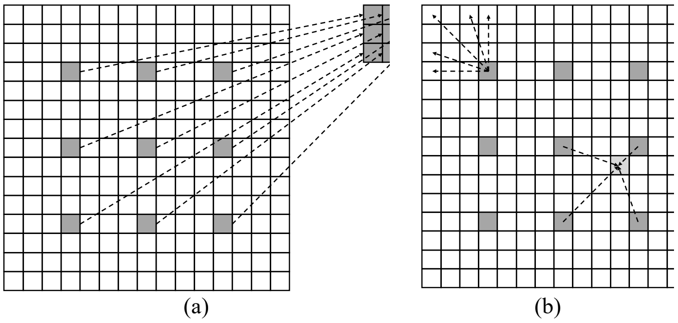


Fig. Error! Reference source not found.**.17** Decimation and reconstruction of sparse set of MIs. (a) The sparse set is constructed by selecting every s MI and combing them into a new LF image of fewer MIs; here $s = 4$. (b) LF image of high resolution is reconstructed by first placing the MIs of the low resolution LF image into their original positions. The remaining MIs are recovered through predictions that use estimated disparities, see Sect. 7.5.3.2. MIs between those MIs part of the sparse set averaged using all surrounding MIs (bottom right arrows)

7.5.3.2 Disparity maps

Disparities between adjacent MIs are estimated on the original LF image. They are later used to calculate the disparities from the sparse set of MIs onto the MIs removed from the original LF image. In fact, these disparities can give an estimated disparity between any two MIs in the LF image. The disparities are gathered in two maps, one for horizontal disparity and one for vertical disparity.

The disparity map is computed by finding a best disparity for the whole MI and its neighbor. The total pixel square error between the MI displaced by the disparity and its neighbor is minimized to obtain the disparity value,

$$\begin{aligned} D_h(x,y) &= \arg\min_{D_h} \left\| C(x,y,r+D_h,t) - C(x+1,y,r,t) \right\|_F \\ D_v(x,y) &= \arg\min_{D_v} \left\| C(x,y,r,t+D_h) - C(x,y+1,r,t) \right\|_F \end{aligned} \tag{6.2}$$

where subscript F denotes the Frobenius norm, in which the summation is done over all r and t. Note that the disparity maps are of sizes $(N-1) \times M$ and $N \times (M-1)$, respectively.

These disparity maps can be compressed in many different ways. It is very important that the decoded values are very accurate in order to retain a good prediction result when reconstructing the full LF image. (See Section 7.5.4 for the reconstruction process.) In [39], HEVC lossless intra codec was used to assure the quality of the disparity maps. However, it turns out that HEVC lossy coding of high quality (low QP-value) give higher compression ratio with sufficient quality. It was chosen to use HEVC to encode the disparity maps as a sequence of two frames, i.e. the first map is encoded by HEVC lossy intra coding and the second is inter-frame predicted. The compressed disparity maps constitute the second layer of the scalable codec.

The encoder further includes a decoding of disparity maps and a reconstruction of the full LF image starting from the sparse set of MIs and calculated disparities. This is done to assure the same data as in the decoder for the process of encoding the third layer of the scalable codec.

7.5.3.3 Refinement by inter and intra prediction

The sparse set of MIs and the disparity maps can be used to predict the other MIs of the full resolution LF image. This LF image reconstruction is as other predictions of lower quality than the original. A straightforward way to have a final LF image of high quality would be to compute the residuals with respect to the original LF image, and compress the residuals by common arithmetic coding. In [39] the LF image reconstruction was instead used as a reference frame in the displacement intra and inter prediction compression as described in Section 7.5.2. and the HEVC RDO chooses the block that gives the least error. The residuals of this prediction are compressed by the HEVC. The compression by inter and intra prediction constitutes the third layer of the scalable codec that produces the full resolution LF image of highest quality.

7.5.4 Decoding and Reconstruction

The encoded LF image data are decoded and reconstructed in three steps that corresponds to the three scalability layers, as schematically depicted in Fig. 6.18.

The first layer is decoded according to the Disparity intra prediction scheme defined in [31]. The result of the decoding is the sparse set of MIs, which itself is a low resolution LF image of the original LF image. It can be used to render images of low spatial resolution and refocusing with limited depth resolution and depth of field as the data contain low angular resolution. This is sufficient for thumbnails

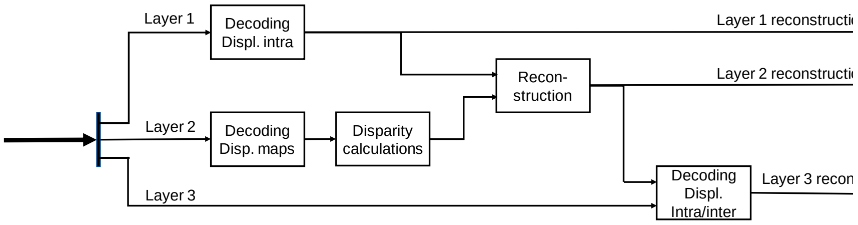


Fig. Error! Reference source not found.**.18** Schematic overview of decoding for sparse set of MIs and disparities. The decoding is divided into three parts, one for each scalability layer. Each layer decoding results in LF images of low resolution (layer 1), high resolution of low quality (layer 2), and high resolution and high quality (layer 3).

and small displays.

The second layer is decoded using HEVC and results in the horizontal and vertical disparity maps, $D_h(x,y)$ and $D_v(x,y)$. These maps are then used to calculate the disparities $D_{hs}(x,y,x_s,y_s)$ and $D_{vs}(x,y,x_s,y_s)$ in (6.3), from the MIs at position (x_s,y_s) in the sparse set to each MI position (x,y) of the original LF image, i.e. to those not being part of the sparse set. See Fig. 6.18.

$$\begin{aligned} D_{hs}(x,y,x_s,y_s) &= \left\{ \begin{array}{ll} \sum_{k=x}^{x_s-1} D_h(k,y) & x < x_s \\ -\sum_{k=x_s}^{x-1} D_h(k,y) & x > x_s \end{array} \right\} \\ D_{vs}(x,y,x_s,y_s) &= \left\{ \begin{array}{ll} \sum_{l=y}^{y_s-1} D_h(x,l) & y < y_s \\ -\sum_{l=y_s}^{y-1} D_h(x,l) & y > y_s \end{array} \right\} \end{aligned} \qquad (6.3)$$

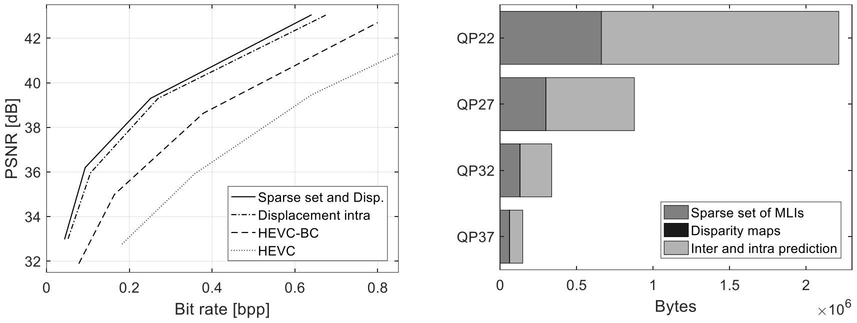
The full LF image of the second layer is reconstructed by shifting the sparse set MIs using the disparities $D_{hs}(x,y,x_s,y_s)$ and $D_{vs}(x,y,x_s,y_s)$, and placing the predicted MI in the corresponding position. If the predicted MI has sparse set MIs on more than one side, the predicted MIs are averaged. See Fig. 6.17b. In case there are still missing areas in a predicted MI, i.e. pixels have not been assigned a value in the disparity-based prediction, these areas need to be filled with plausible information. For this reason, a dynamic inpainting approach [45] was employed in [38] to obtain the final LF image reconstruction of the second layer. This second layer LF image reconstruction has full spatial and angular resolution but has a lower image quality than when also utilizing the third layer data.

The third layer data is fed into the decoder of the disparity intra and inter prediction scheme, along with the output from the second layer. The second layer LF image is put into the reference list and is used for inter prediction along intra prediction of the third layer data. Thereby, the final, third layer, LF image is reconstructed that has full resolution and is of high quality.

7.5.5 Evaluation

The coding scheme using a sparse set of MIs and disparities was evaluated in [38, 39]. The lowest bit rate is obtained for a decimation factor of $s = 2$, which led to a bit rate reduction of 50% - 60% for different input LF images. Larger decimation factors imply a small increase in bit rate. Although it improves the compression efficiency by only a small margin relative to the displacement intra prediction, it provides a scalable structure for coding and rendering. Compared to HEVC BC being a single-hypothesis intra prediction scheme, the sparse set and disparity scheme has a reduction of 20% in bit rate. See Fig. 6.19a.

The third scalability layer contains a fair amount of the bit budget, especially when a very high quality is required for the final LFI. The compression of disparity maps in the second layer results in a very low number of bits when the lossy HEVC coding is employed, whereas the lossless coding use more than 30 times more bits. Yet, the second layer the smallest component among the three. See Fig. 6.19b.



(a) (b)

Fig. Error! Reference source not found.**.19** Efficiency of compression schemes. (a) Rate-distortion graphs for evaluated compression schemes. The scheme using sparse set and disparities performs slightly better than the displacement intra scheme and much better than the single-hypothesis prediction method HEVC-BC and standard HEVC. (b) Distributions of data in the three scalability layers. The third layer contains most data, whereas the second layer (disparities) is constant independent of QP-value, and contains less than 0.5 % of the total for QP37. Data obtained from [38].

The objective quality of the second layer reconstruction is much lower than that of the third and final layer. See Fig. 6.20a. However, the visual quality of the second layer reconstruction is fairly good for the central view rendering, even if improvements can be seen for the full LFI reconstruction. See Fig. 6.20b.

7.5.5.1 Remarks

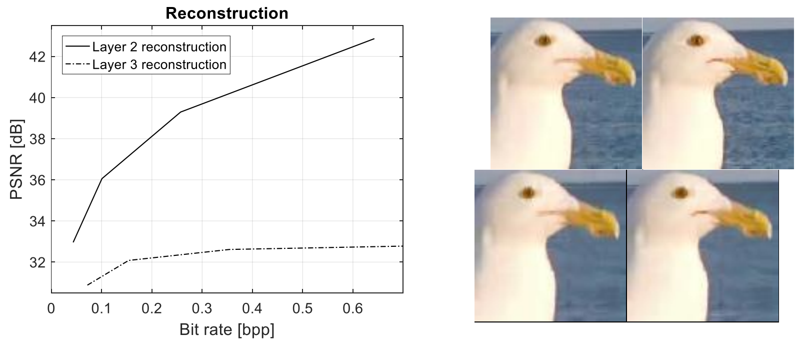
A compression scheme for LF images from focused LF cameras was presented in this section. It uses a sparse set of micro-lens images (MIs) and disparities between these MIs. The scheme exhibits large compression improvements over both HEVC Intra and HEVC BC, and moderate improvements over the multi-hypothesis prediction scheme Displacement intra. The computational complexity is increase. Instead, the scheme introduces scalability in both resolution and quality, and so provides a flexible reconstruction of images.

7.6 Conclusions

This chapter covered recent advances in LF coding, based on different approaches. After a brief description of LF representation formats, the coding efficiency of unmodified standard codecs using various LF image data structures were evaluated and discussed for different coding configurations. In general, it was shown that LFs in pseudo video format provide higher compression efficiency still image formats. Then a scalable LF coding solution, capable of providing compatible substreams to 2D and 3D decoders, is described and evaluated. A display scalable architecture was presented, using a three-layered hierarchical approach, which allows to support a wide range of end-user displays, from conventional 2D to advanced immersive LF applications (e.g., augmented and immersive virtual reality). Furthermore, another recent approach to encode LFs, which exploits spatial correlation based on the disparity maps and multi-hypothesis prediction is also presented and discussed. A sparse set of MIs is used as a first layer, which provides a small resolution representation of the visual content. Then the second and third layers provide higher spatial resolution and the full resolution of the LF, respectively. Such scalable coding schemes also enable seamless interoperability with legacy video systems and smooth transition to emerging applications and services where LFs are increasingly gaining importance and user acceptance.

7.7 References

1. Lippmann G (1908) Épreuves Réversibles Donnant la Sensation du Relief. J Phys Théorique Appliquée 7:821–825.
2. Levoy M (2006) Light fields and computational imaging. Computer (Long Beach Calif) 39:46–55. doi: 10.1109/MC.2006.270



(a)

(b)

Fig. Error! Reference source not found.**.20** Reconstructed images using decimation factor $s = 2$. (a) Objective quality (PSNR) for reconstructed central view using layer 1-2 and layer 1-3, respectively. (b) Central view reconstruction. Upper images were compressed using QP22, lower used QP37. Left images are reconstructed using layer 1-2, and right images using layer 1-3. Images originally published in [38].

3. Levoy M, Hanrahan P (1996) Light Field Rendering. In: Proc. 23rd Annu. Conf. Comput. Graph. Interact. Tech. - SIGGRAPH '96. New Orleans, LA, US, pp 31–42
4. Aggoun A (2006) A 3D DCT Compression Algorithm For Omnidirectional Integral Images. In: 2006 IEEE Int. Conf. Acoust. Speed Signal Process. Proc. Toulouse, France, p II-517-II-520
5. Aggoun A (2011) Compression of 3D Integral Images Using 3D Wavelet Transform. J Disp Technol 7:586–592. doi: 10.1109/JDT.2011.2159359
6. Perra C, Assuncao P (2016) High Efficiency Coding of Light Field Images based on Tiling and Pseudo-Temporal Data Arrangement. In: 2016 IEEE Int. Conf. Multimed. Expo Work. Seattle, WA, US, pp 1–4
7. Perra C (2015) Lossless plenoptic image compression using adaptive block differential prediction. In: 2015 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp 1231–1234
8. Helin P, Astola P, Rao B, Tabus I (2016) Sparse Modelling and Predictive Coding of Subaperture Images for Lossless Plenoptic Image Compression. In: 2016 3DTV-Conference True Vis. - Capture, Transm. Disp. 3D Video. Hamburg, Germany, pp 1–4
9. Santos JM, Assuncao PAA, Cruz LA da S, et al (2017) Performance evaluation of light field pre-processing methods for lossless standard coding. IEEE Commun Soc MMTC Commun - Front 12:44–49.
10. Vieira A, Duarte H, Perra C, et al (2015) Data Formats for High Efficiency Coding of Lytro-Illum Light Fields. In: 2015 Int. Conf. Image Process. Theory, Tools Appl. Orleans, France, pp 494–497
11. Dansereau DGDG, Pizarro O, Williams SBSB (2013) Decoding, Calibration and Rectification for Lenselet-Based Plenoptic Cameras. In: 2013 IEEE Conf. Comput. Vis. Pattern Recognit. Portland, OR, US, pp 1027–1034
12. (2013) Draft Test conditions for HEVC still picture coding performance evaluation. ISO/IEC JTC1/SC29/WG11 MPEG2013/ N13826, Vienna, Austria
13. Conti C, Nunes P, Soares LD (2013) Inter-Layer Prediction Scheme for Scalable 3-D Holoscopic Video Coding. IEEE Signal Process Lett 20:819–822. doi: 10.1109/LSP.2013.2267234
14. Aggoun A, Tsekleves E, Swash MR, et al (2013) Immersive 3D Holoscopic Video System. IEEE Multimed 20:28–37. doi: 10.1109/MMUL.2012.42
15. Arai J, Kawakita M, Yamashita T, et al (2013) Integral Three-Dimensional Television with Video System Using Pixel-Offset Method. Opt Express 21:3474–3485. doi: 10.1364/OE.21.003474
16. Arai J (2015) Integral Three-Dimensional Television. In: 2015 14th Work. Inf. Opt. Kyoto, Japan, pp 1–3
17. (2016) NHK STRL Science & Technology Research Laboratories. https://www.nhk.or.jp/strl/index-e.html. Accessed 10 Jul 2016
18. Wang J, Xiao X, Hua H, Javidi B (2015) Augmented Reality 3D Displays With Micro Integral Imaging. J Disp Technol 11:889–893. doi: 10.1109/JDT.2014.2361147
19. Lanman D, Luebke D (2013) Near-Eye Light Field Displays. ACM SIGGRAPH 2013 Emerg Technol - SIGGRAPH '13 1–1. doi: 10.1145/2503368.2503379
20. Conti C, Nunes P, Soares LD (2013) Using Self-Similarity Compensation for Improving Inter-Layer Prediction in Scalable 3D Holoscopic Video Coding. In: Proc. SPIE 8856 Appl. Digit. Image Process. XXXVI. San Diego, CA, US, p 88561K
21. Sullivan GJ, Ohm J-R, Han W-J, Wiegand T (2012) Overview of the High Efficiency Video Coding (HEVC) Standard. IEEE Trans Circuits Syst Video Technol 22:1649–1668.
22. Vetro A, Wiegand T, Sullivan GJ (2011) Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard. Proc IEEE 99:626–642. doi: 10.1109/JPROC.2010.2098830
23. (2013) White Paper on State of the Art in compression and transmission of 3D Video. ISO/IEC JTC1/SC29/WG11 N13364, Geneva, Switzerland

24. Vetro A, Müller K (2013) Depth-Based 3D Video Formats and Coding Technology. In: Dufaux F, Pesquet-Popescu B, Cagnazzo M (eds) Emerg. Technol. 3D Video. John Wiley & Sons, Ltd, Chichester, UK, pp 139–161
25. Tech G, Chen Y, Muller K, et al (2016) Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. IEEE Trans Circuits Syst Video Technol 26:35–49. doi: 10.1109/TCSVT.2015.2477935
26. Georgiev T, Lumsdaine A (2010) Focused Plenoptic Camera and Rendering. J Electron Imaging 19:021106–021106. doi: 10.1117/1.3442712
27. Conti C, Nunes P, Soares LD (2012) New HEVC Prediction Modes for 3D Holoscopic Video Coding. In: 2012 19th IEEE Int. Conf. Image Process. Orlando, FL, US, pp 1325–1328
28. Conti C, Soares LD, Nunes P (2016) HEVC-Based 3D Holoscopic Video Coding using Self-Similarity Compensated Prediction. Signal Process Image Commun 42:59–78. doi: 10.1016/j.image.2016.01.008
29. Conti C, Nunes P, Soares LD (2016) HEVC-Based Light Field Image Coding with Bi-Predicted Self-Similarity Compensation. In: 2016 IEEE Int. Conf. Multimed. Expo Work. Seattle, WA, US, pp 1–4
30. Conti C, Nunes P, Ducla Soares L (2018) Light field image coding with jointly estimated self-similarity bi-prediction. Signal Process Image Commun 60:144–159. doi: 10.1016/J.IMAGE.2017.10.006
31. Li Y, Sjostrom M, Olsson R, Jennehag U (2016) Coding of Focused Plenoptic Contents by Displacement Intra Prediction. IEEE Trans Circuits Syst Video Technol 26:1308–1319. doi: 10.1109/TCSVT.2015.2450333
32. Sullivan GJ, Wiegand T (1998) Rate-Distortion Optimization for Video Compression. IEEE Signal Process Mag 15:74–90. doi: 10.1109/79.733497
33. Geogiev T Todor Georgiev Gallery of Light Field Data. http://www.tgeorgiev.net/Gallery/. Accessed 17 Sep 2016
34. (2013) 3D Holoscopic Sequences (Download Link). http://3dholoscopicsequences.4shared.com/. Accessed 30 Oct 2016
35. MV-HEVC Reference Software HTM-12.0. https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-12.0/. Accessed 22 Dec 2014
36. Bjøntegaard G (2001) Calculation of Average PSNR Differences between RD Curves. VCEG-M33, Austin, TX, US
37. Bossen F (2013) Common HM Test Conditions and Software Reference Configurations. JCTVC-L1100, Geneva, Switzerland
38. Li Y, Sjöström M, Olsson R, Jennehag U (2016) Scalable Coding of Plenoptic Images by Using a Sparse Set and Disparities. IEEE Trans Image Process 25:80–91. doi: 10.1109/TIP.2015.2498406
39. Li Y, Sjöström M, Olsson R (2015) Coding of Plenoptic Images by using a Sparse Set and Disparities. In: 2015 IEEE Int. Conf. Multimed. Expo. IEEE, pp 1–6
40. Li Y, Sjostrom M, Olsson R, Jennehag U (2014) Efficient intra prediction scheme for light field image compression. In: 2014 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp 539–543
41. Li Y, Olsson R, Sjostrom M (2016) Compression of unfocused plenoptic images using a displacement intra prediction. In: 2016 IEEE Int. Conf. Multimed. Expo Work. IEEE, pp 1–4
42. Magnor M, Girod B (2000) Data compression for light-field rendering. IEEE Trans Circuits Syst Video Technol 10:338–343. doi: 10.1109/76.836278
43. Kundu S (2012) Light field compression using homography and 2D warping. In: 2012 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, pp 1349–1352
44. Rosewarne C, Sharman K, Naccari M, Sullivan G (2014) HEVC Range Extensions Test Model 6 Encoder Description. JCTVC-P1013, San Jose, CA, US

45. Bertalmio M, Bertozzi AL, Sapiro G (2001) Navier-stokes, fluid dynamics, and image and video inpainting. In: Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition. CVPR 2001. Kauai, HI, US, p I-355-I-362